Practical Limits on Achievable Energy Improvements and Useable Delay Tolerance in Correlation Aware Data Gathering in Wireless Sensor Networks

Yujie Zhu, Karthikeyan Sundaresan and Raghupathy Sivakumar
School of Electrical and Computer Engineering
Georgia Institute of Technology
{zhuyujie, sk, siva}@ece.gatech.edu

Abstract*— Correlation of data sent by different sensors in a wireless sensor network can be exploited during the data gathering process to improve energy efficiency. In this paper, we study the energy efficiency of correlation aware data aggregation trees under various sensor network conditions and the tradeoffs involved in using them. The following two related questions are specifically investigated in the study: (i) Is there any practical limit on the achievable improvement in energy efficiency in adopting a correlation aware aggregation structure as opposed to a correlation unaware structure? (ii) Is there a practical maximum useable delay bound that can deliver the maximum achievable improvement? In answering the above questions, we present comprehensive simulation results and draw inferences based on the results. We also conclude two rather surprising results that *the energy improvement in using correlation aware aggregation is not significant under many network scenarios*, and *the maximum useable delay bound is not large compared with the delay along the maximum length shortest-path in the default shortest path tree.

I. Introduction

Wireless Sensor Networks (WSNs) have gained tremendous importance in recent years because of their potential use in various fields. The devices used for sensing and communication in such networks are usually small, cheap and low powered and hence, have limited resources for computation as well as communication. This has spurred a need for energy efficient protocols tailored specifically toward sensor network environments.

One of the key tasks performed by any WSN is the collection of sensor data from the sensors in the field to the sink for processing. This task is also referred to as *data gathering*. In this paper, we consider the problem of data gathering in environments where the data from the different sensors are correlated. Such correlation of the data being collected can be leveraged by appropriately fusing the data inside the network to the best extent possible, thereby reducing the number of transmissions and hence energy consumption, for the gathering process.

On the other hand, a data gathering tree that does not explicitly make use of the correlation between sensor data can be considered to be correlation unaware. The most representative structure for correlation unaware aggregation approaches is a *Shortest Path Tree* (SPT).

Since the primary goal of the structure is to minimize delay, SPT is not considered to be a correlation-aware data gathering structure. Even though opportunistic aggregation may possibly occur when different paths overlap with each other, it does not necessarily maximize the degree of aggregation possible in the network.

The objective of correlation aware data gathering is to reduce the energy cost of an aggregation tree. The energy optimal aggregation structure for a data gathering application depends on the degree of correlation existing between the source data. For statistical queries such as min, max, avg, etc., two pieces of data can be combined and reduced to the same size as that of the original pieces. We call this type of correlation as *perfect* correlation. It is well-known that when sensor data are perfectly correlated, the *Steiner Minimum Tree* (SMT) over all the sources, sink and some of the non-source nodes is optimal. On the other hand, there are other scenarios where the message sizes may not be reduced to the same size as the original data; only a part of each piece of information is redundant. If the correlations between sensor data are not perfect, there is no established optimal structure. Hence, several attempts ([1], [2]) have been made to propose heuristics to approximate the optimal solution.

In this paper, we study the achievable benefits in using correlation aware structures in practical sensor applications. Specifically, we investigate sensor applications with different degrees of delay tolerance, and explore the energy benefits brought about by correlation aware structures in data gathering, as well as the trade-offs for obtaining these energy benefits. Note that the delay tolerance of the application will determine the optimality of the data gathering structure. This is because, to maximize aggregation, some sensor data may have to travel additional hops to combine with other sensor data, thereby increasing the delay of the data gathering process.

This work was funded in part by the National Science Foundation under grants ECS-0225497, ECS-0428329, ANI-0117840, and CCR-0313005.

Hence, to satisfy the delay constraints of the application, some intermediate structure between SPT and SMT may be used to strike a balance between the energy efficiency and the delay requirement.

Also, the benefits of a correlation aware data gathering structure come at the expense of a construction process that typically incurs more overhead than that for a simpler structure such as the SPT, both because of the coordination required for the construction and the fact that unlike the SPT, the correlation aware structure needs to change for each new set of source nodes. Hence, for a correlation aware aggregation tree to be energy efficient, the overhead involved in the tree construction should also be taken into account and the energy savings of the resulting tree should be the net savings after accommodating the cost incurred for the construction.

In this work, we investigate the energy efficiency of the correlation aware aggregation process through comprehensive quantitative analysis. We specifically explore how the improvement in energy efficiency is impacted by network conditions, defined by several parameters including the node density, source density, the physical distribution of sources, the correlation degree, and the delay bound. We present observations from the simulation results, and draw inferences on the trade-offs involved in achieving energy efficiency.

In studying the improvements in energy efficiency with respect to specific network parameters, we also answer two fundamental questions:

1) *Is there a practical limit on the achievable improvement in energy efficiency by adopting a correlation aware aggregation structure as opposed to a correlation unaware structure?* The answer to this question will establish practical bounds on the energy efficiency improvement that can be achieved, and in turn provide a motivation or lack there-of for performing correlation aware aggregation in the first place.
2) *Is there a maximum usable delay bound that can deliver the maximum achievable energy cost improvement?* The answer to this question will establish a practical bound on how delay tolerant a WSN application needs to be in order to get the maximum energy efficiency benefit.

Our contributions can thus be summarized as follows:

- We characterize through quantitative analysis how the energy improvement of a correlation aware aggregation structure is impacted by different network parameters. We show that the energy improvement tends to be bounded by a small constant under many network scenarios. Furthermore, the improvement corresponds to when the additional cost of establishing a correlation aware structure is not taken into account, in the presence of which the improvement will be further reduced.
- We also characterize what the maximum usable delay bound is for achieving the maximum energy efficient structure. We show that the maximum usable delay bound is a small constant times the delay along the maximum length shortest-path in the default shortest path tree.

The rest of the paper is organized as follows: In section II, we describe the evaluation methodology and parameters. The optimization algorithm used for cost evaluation is also briefly introduced. In section III, we present comprehensive simulation results for varying parameters as well as explanations for the inferences drawn from the results. In section IV, we substantiate an important observation made from the simulation study, and derive analytical expressions to corroborate the observation. Finally, we present related work in section VI, followed by a discussion on some practical issues in V, and conclude the paper in section VII.

II. Model

We use a custom-built simulator written in C++ for all our simulations. The simulator takes as input the shape of the network, node density, source density, source distribution and correlation degree, and the outputs are the respective correlation unaware aggregation trees and correlation aware aggregation trees with different delay bounds, along with their respective costs.

A. Evaluation Metrics

Most of the energy consumption in a data gathering process is due to communication. Hence, the amount of communication (number of transmissions) required is directly related to the cost of the aggregation tree. Thus, we consider the aggregation tree cost - the number of edges on a given aggregation tree - as the measure of energy efficiency of the corresponding data gathering process.

The metric we use to measure the energy efficiency improvement provided by correlation aware trees is the *cost ratio*, which is defined as the ratio of the cost of the correlation unaware tree to that of the correlation aware tree over the same set of sources and sink. The shortest path tree is constructed with the purpose of minimizing end to end delay for each source. However, multiple paths from different sources to sink can overlap at some intermediate relay nodes, where opportunistic aggregation is possible. We assume such opportunistic aggregation to take place in all our evaluations. Several synchronization schemes exist to enable such opportunistic aggregation [3]. For a correlation aware tree, the degree of aggregation is higher. Thus, the energy consumption of correlation aware tree tends to be lower. The cost ratio defined in the above fashion measures the relative efficiency of the aggregation aware tree to that of an aggregation unaware tree.

In most sensor applications, delay bound is typically defined to be the maximum delay instead of the average delay required to collect all sensor data. In the aggregation process, messages from sources closer to sink need to be held at some intermediate nodes until other messages from sources farther away arrive at this node in order to achieve maximum aggregation possible. For this reason, the delay incurred in the entire data gathering process is proportional to the maximum delay required to gather data from the source that is farthest from the sink.

In a data gathering process, the delay at each hop of the aggregation tree should include transmission delay, contention delay and aggregation delay. For easy of analysis, we assume a contention-free environment where centralized MAC layer scheduling is used to coordinate transmissions within a contention region. Therefore, the most important factor that contributes to the data gathering latency is the transmission delay and aggregation delay. Aggregation delay comprises not only of the processing time for aggregation at each node, but also the time that an aggregation node takes to wait for data from all downstream nodes in the tree to reach it. Thus, the total delay for a certain data gathering path can be assumed to be proportional to the number of hops on the path. Consequently, we specify delay constraints as the maximum allowable path length in terms of hop count.

B. Evaluation Environment and Parameters

To study the energy efficiency and tradeoffs of correlation aware aggregation trees, we consider a typical sensor network scenario where a total of n sensors are randomly distributed in a disk of radius R. All the sensors communicate using the same transmission range, which is slightly higher than that required for minimum connectivity [4]. Of the n sensors, k are randomly chosen as sources to report data to the sink, which is located at the center of the disk. In this case, data aggregation trees span all sources and are rooted at the sink. This configuration is representative of many sensor network applications and results derived from it are easily extensible to other scenarios such as multiple sink applications.

The following network parameters are used for a comprehensive evaluation:

1) *Delay bound:* deadline imposed by a sensor application to one round of data gathering.
2) *Node density:* total number of nodes distributed in an unit area in the sensor network.
3) *Source density:* ratio of the number of sensors that send data packets to the sink to the total number of sensor nodes in the network.
4) *Source distribution:* geographical distribution of source nodes - uniform or non-uniform.
5) *Correlation degree:* measure of how much information two raw data packets share with each other.

C. Algorithms

We choose Steiner Minimum Tree (SMT) as the correlation aware structure, since it is the optimal aggregation structure [1] when sensor data are perfectly correlated. On the other hand, SPT is selected as the correlation unaware structure since it minimizes the delay required for data aggregation. 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 is an intermediate structure between SMTs and SPTs. Consequently, the energy improvement of SMT over SPT that we study in this paper serves as an upper bound on the energy improvement possible for all other correlation models as well.

The well-known Dijkstra's algorithm is used to compute shortest path tree in the simulations. For SMT, since its computation is a NP-hard problem, we resort to heuristics to generate near-optimal aggregation structures. To evaluate the impact of delay sensitivity of the application on the cost of a near-optimal tree, we need an algorithm that generates near-optimal trees for various delay constraints. Specifically, if the delay bound for a certain data gathering task is D, the delay incurred on the longest path of the near-optimal aggregation tree should be less than or equal to D. From hereon, *we refer to the delay-bounded near-optimal tree as DB-SMT (delay-bounded steiner minimum tree) and the near-optimal tree without delay bound as simply the SMT.*

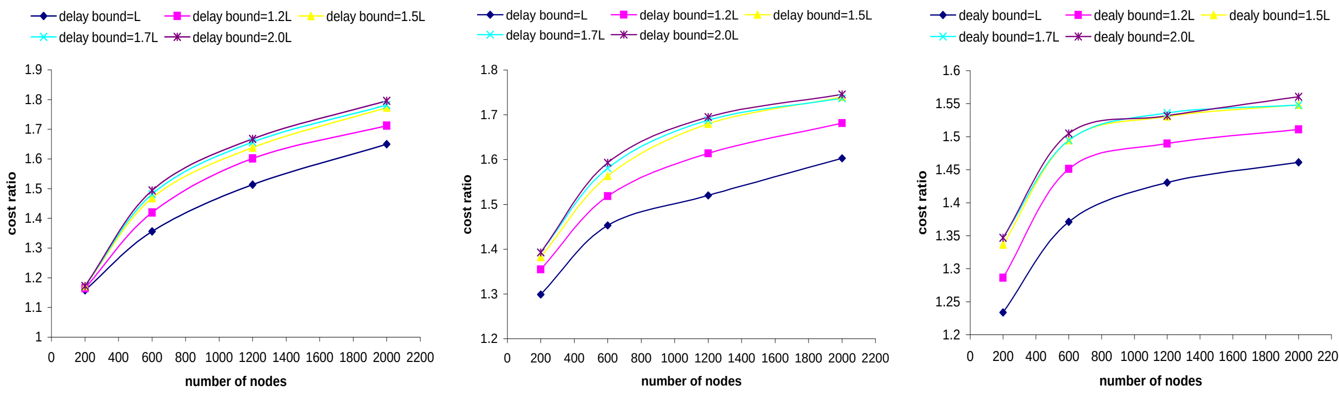
A set of algorithms developed in the context of multicast applications can be used for this purpose. Most multicast routing algorithms are designed to support large number of simultaneous multicast sessions efficiently. A multicast tree that minimizes the total bandwidth utilization of the network links is be established from the source to destinations in these algorithms. Hence, these algorithms can be used for sensor network aggregation, with the only difference being that the data flows in sensor networks are in the reverse direction. Some of these algorithms are specifically tailored to multicast applications that are delay sensitive such as multimedia streaming. Such algorithms, called Constrained Steiner Tree heuristics (CST), generate minimum cost multicast trees within certain delay constraints and can hence be used exactly for our purpose.

We choose a CST algorithm called BSMA (bounded shortest multicast algorithm) to generate the DB-MST. This algorithm has been proven to be able to construct multicast trees with additional costs less than 7% that of the optimal Steiner Minimum Tree, and has been shown to achieve lower costs than other related strategies [5].

D. Methodology

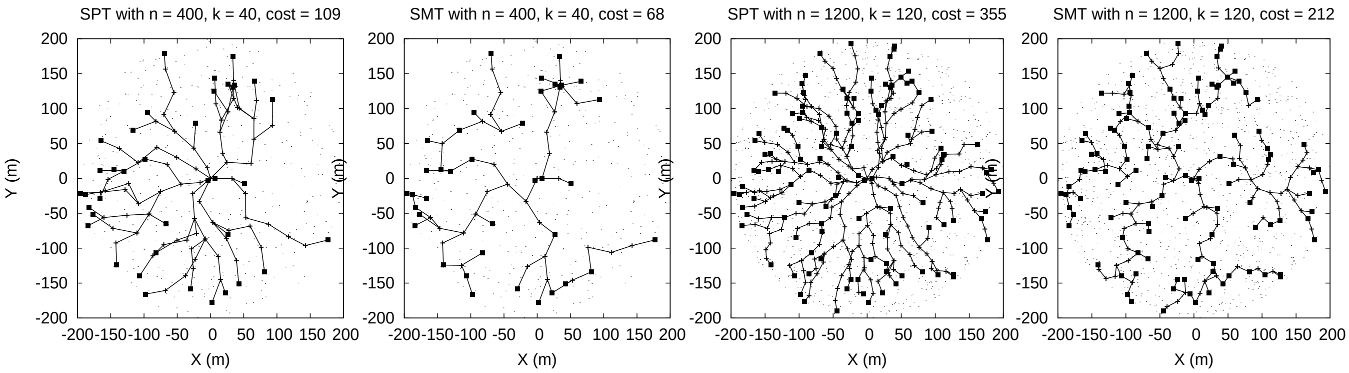
To study the energy efficiency and tradeoffs in WSN data aggregation process, we start from a shortest path tree spanning the sink and all the source nodes in the network, and apply the BSMA algorithm to reduce the cost of the tree. Using this algorithm, different tree structures can be obtained for different delay constraints. Note that, the delay bound has to be higher than the longest shortest path from sources to the sink; otherwise no valid tree can be found. The maximum hop count in the initial shortest path tree is taken as the lowest possible delay bound and the DB-SMT tree is initially generated with this delay bound. The delay bound is then relaxed to obtain DB-SMT trees with further reduced costs.

For each network configuration, we vary the network parameters specified in the previous subsection, generate SPT and DB-SMT trees for each configuration, and take the cost ratio between SPT and DB-SMT trees. Each network configuration is run for several random seeds, and the average of the cost ratio across the seeds serves as a data point in the graphs



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

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



(a) $SPT, n = 400$ (b) $SMT, n = 400$ (c) $SPT, n = 1200$ (d) $SMT, n = 1200$

Fig. 2. Topology of SPTs and SMTs

presented subsequently. Each graph contains several curves displaying the relationship between cost ratio and one of the varying network parameters, with each curve corresponding to a specific delay constraint. The varying trends, thresholds and bounds for each graph are identified and discussed in the following section.

III. Performance Analysis

In this section we present simulation results to show the energy-delay tradeoffs of aggregation trees under various network conditions.

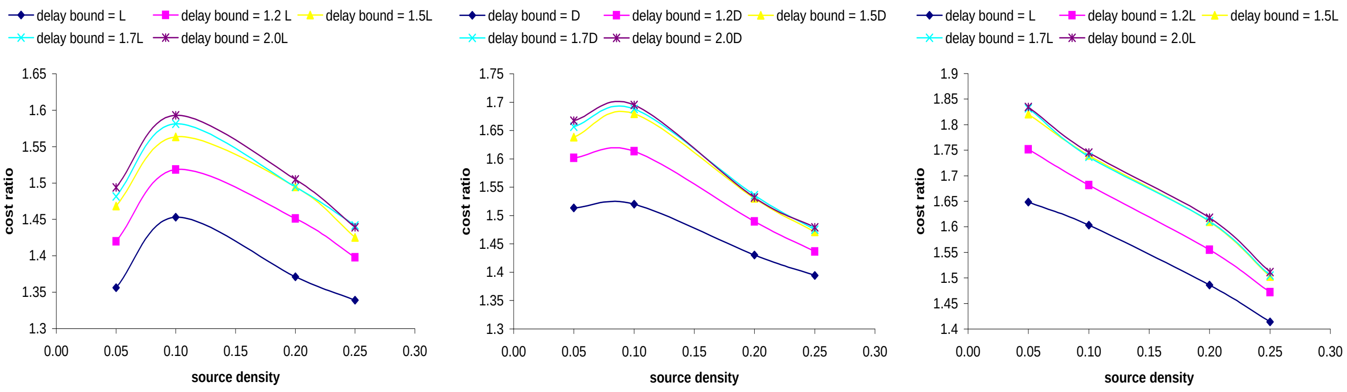
A. Varying Node Density

To study the impact of node density on the energy efficiency of aggregation trees, the number of sensor nodes distributed in the field (n) is increased from 200 to 2000. Figure 1 shows the cost ratio of SPT vs. DB-SMT when the source densities are $1/20$, $1/10$ and $1/5$.

It can be observed from the results that the cost ratio between SPT and DB-SMT increases with node density. This implies that correlation aware data gathering is more efficient when the density of sensor nodes is large. This can be intuitively explained as follows: with high node density, the probability of shortest paths over-lapping with each other is low; hence, SPT has very low aggregation efficiency and there is greater potential for energy improvement using a DB-SMT. Consequently, the cost ratio improves as node density increases. To further illustrate this observation, we plot the structure of SPT and DB-SMT trees constructed when n is 400 and 1200 in Figure 2. In both configurations, the number of source nodes is $1/10$ that of total sensor nodes. For the case of $n = 1200$, it can be seen that many parallel shortest paths exist in the SPT structure. However, after optimization for aggregation, most of these separated paths are combined, thereby enabling great cost savings in DB-SMT. However, for the $n = 400$ case, SPT is already an efficient structure in terms of path sharing. Thus, the improvement after optimization is not significant.

We also observe that for different source densities, the increasing trend of cost ratio remains to be the same. however, the absolute value of cost ratio reduces as source density increases. Further, for the same source density, when the delay constraint is increased, the cost ratio between SPT and DB-SMT increases, but the increase in ratio tends to saturate when the delay bound is more than 1.5 times the longest shortest path length. The specific reasons for these observations will be presented when the impact of source density and delay constraints are investigated subsequently.

From the three results, we can observe that the cost ratio



(a) $n = 600$

(b) $n = 1200$

(c) $n = 2000$

Fig. 3. Performance Improvement over SPT for Different Source Densities

tends to saturate when node density is high (corresponding to $n = 1200$). In other words, the rate of increase of cost ratio tends to slow down with increasing node density. This implies that, in contrast to common belief, energy improvement of DB-SMT over SPT does not scale significantly with node density. In fact, we theoretically analyze and provide a tight bound on the rate at which the cost ratio improves with node density in Section IV.

The results also indicate that at low node density, correlation aware data gathering does not bring significant cost improvement. If we require a correlation aware aggregation tree to provide a factor of at least 1.5 times in cost improvement (50% improvement), then from the results it is clear that this is possible only when the node density is sufficiently high with $n > 600$. This in turn implies that correlation aware aggregation does not provide desired energy efficiency for low node densities.

Thus, from the study of cost ratio variation with node density, we have the following insights:

Cost ratio of SPT over DB-SMT increases with node density in sensor networks, but tends to saturate with increasing node density. For correlation aware aggregation trees to achieve a desirable energy improvement, the node density of the sensor network should be relatively high.

B. Varying Source Density

To investigate how the density of source sensor nodes affects the efficiency of aggregation tree, we compare the cost of SPT and DB-SMT across a range of source densities and node densities. Figure 3 shows how cost ratio varies with source densities when $n = 600$, $n = 1200$ and $n = 2000$. Each curve consists of four data points with respect to source densities of $1/20$, $1/10$, $1/5$ and $1/4$.

It can be observed that when node density is low, the cost ratio increases with source density, reaches a maximum, and then starts to decrease again. However, for high node density, the cost ratio decreases monotonically with source density. When there are fewer sensor nodes in the network, due to the relatively small SPT cost at low source densities, the possible cost reduction achievable from optimization in DB-SMT is limited. This explains the low cost ratio at source density of $1/20$ for $n = 600$ and $n = 1200$ cases. On the other hand, when source density is higher than $1/5$, a considerable fraction of nodes on SPT are sources, implying that SPT is already an efficient structure. Consequently, the possible cost reduction from optimization in DB-SMT is once again less. An important factor that determines the degree of cost improvement is the inefficiency of the SPT structure. Thus, at very low and very high source densities, the higher efficiency of SPT structures reduces the cost ratio improvement, resulting in a peak value at an intermediate value of source density.

The inference with respect to node density is the same as before, where as the node density increases, the path diversity in SPT also increases. Thus, the shortest paths in SPT diverge from each other even at low source densities, leaving considerable margins for cost improvement in DB-SMT. This results in the monotonically reducing cost ratio at $n = 2000$.

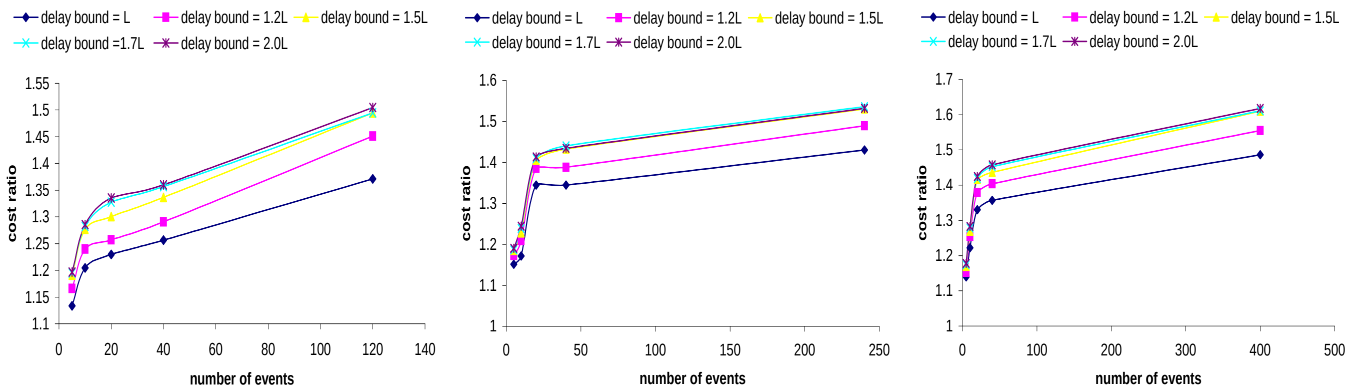
Simulations with source densities larger than $1/4$ were also conducted, and cost ratios were observed to be less than 1.2. This can also be extrapolated from the trend of the curves in Figure 3. From the study of cost ratio variation with source density, we obtain the following insights:

Cost ratio of SPT over DB-SMT decreases with increasing source density when node density is high. However, with low node density, medium source density ensures the best possible cost improvement.

C. Varying Source Distribution

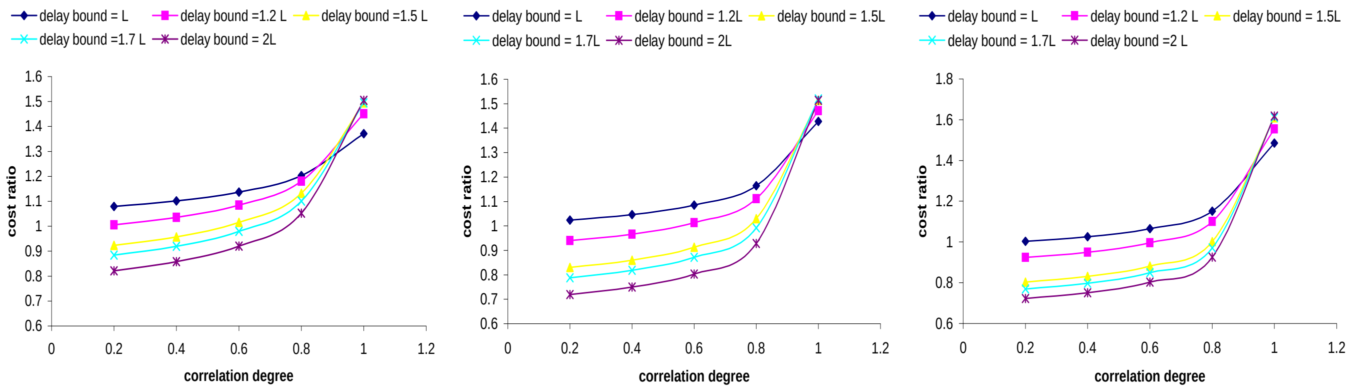
In the previous discussions, we had assumed that sources are uniformly distributed in the network. However, this may not always be the case in sensor network applications. There are situations where only certain specific locations (where scattered events occur) in the network need to be monitored, in which case the sink gathers data from sensor nodes around these events. Under these circumstances, source nodes can no longer be considered to be uniformly distributed. In this subsection, we study how the distribution of sources affects the effectiveness of correlation aware aggregation.

In this set of simulations, n increases from 200 to 2000, and each configuration has a total of $s = n/5$ sources distributed in



(a) $n = 600$ (b) $n = 1200$ (c) $n = 2000$

Fig. 4. Performance Improvement over SPT for Different Source Distributions



(a) $n = 800$ (b) $n = 1200$ (c) $n = 2000$

Fig. 5. Performance Improvement over SPT for Different Correlation Degrees

the network. The number of events e (locations) in the network for each scenario increases from 5, 10, 20, 40 to s. Sources are equally distributed in the different event locations. When the number of events is 5 and 10, the sources are highly cluttered around the event locations, and as event number increases, the source distribution becomes closer to uniform distribution. This model is similar to the event-radius model used in [6].

From the results presented in Figure 4, it can be seen that the cost ratio increases with the number of events. The overall trend of the cost ratio improvement can be explained as follows. The sources tend to be densely distributed around event locations when there are few events in the network. Hence, the shortest paths from the same event location to the sink can combine with each other at an early stage, thereby making SPTs inherently efficient in terms of path sharing. This can be observed from Figure 4, where the cost ratio is less than 1.3 when the number of events is 5 and 10. However, the path diversity of SPTs tends to increase as the number of event locations increases with the source distribution tending towards uniform distribution. Consequently, the cost reduction for correlation aware data aggregation becomes greater.

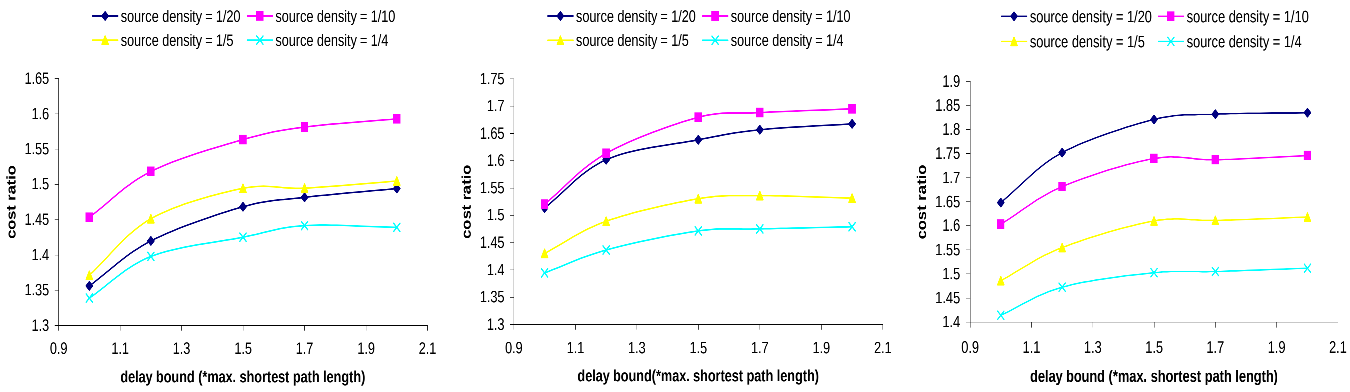
From the study of cost ratio variation with source distribution, we gain the following insight:

Cost ratio of SPT over DB-SMT increases as the distribution of source nodes tends towards uniform distribution

D. Varying Correlation Degree

It is possible that the data gathered in certain sensor network applications are not perfectly correlated, in which case the correlation degree will be less than one. The total message size after aggregation would no longer be the same size as the original message, but instead would be larger. Several works [1], [2] have studied this problem before. However, none of them have identified the effectiveness of SPT versus SMT with respect to varying correlation degree.

Characterizing the correlation existing between data collected in sensor networks is a fairly complicated task, since the nature of correlation differs with the type applications considered. Even for a simple correlation model, the mathematical representation becomes difficult when multiple distributed sources are involved. [2] presents a correlation model that uses joint entropy to define correlation between two sources. A constructive technique is also proposed to characterize correlation when multiple sources are involved, but the calculation becomes intractable when there are large number of sources, uniformly distributed in a 2-dimensional field. For simplicity, we adopt the same correlation model used in [1], where each data packet is assumed to bring a fixed amount of new



(a) $n = 600$

(b) $n = 2000$

(c) $n = 2000$

Fig. 6. Performance Improvement over SPT for Different Delay Bounds

information into the aggregated data packet. Specifically, if ρ is defined to be the correlation degree, and the sizes of raw data packets generated by sensor nodes to be m, then after aggregation of two data packets, the message size becomes $m + (1 - \rho)m$. Similarly, for n sources, the aggregated data packet has a size of $m + (n - 1)(1 - \rho)m$.

In this set of simulations we choose the number of nodes to be 600, 1200 and 2000; source density to be $1/5$; and delay constraints of L, $1.2L$, $1.5L$, $1.7L$ and $2.0L$. Figure 5 illustrates how cost ratio varies with the correlation degree and delay constraints.

It can be seen that the cost ratio increases with correlation degree. This trend remains the same for all node densities. However, for higher node densities, the cost improvement of DB-SMT over SPT is lower. We explain the overall increasing trend of cost ratio with increasing correlation degree as follows: when $\rho \longrightarrow 0$ (raw data packets are un-correlated with each other), SPT is the optimal structure since aggregation does not help reducing the transmission and hence the energy cost. Thus, the best approach is to deliver each message along the shortest possible route to the sink. On the other hand, when $\rho \longrightarrow 1$, SMT is the optimal structure with respect to energy efficiency, as established earlier. Therefore, we expect the optimal structure to be close to SPT for small correlation degrees, wherein progress towards the sink is more important than en-route aggregation. Due to this reason, the cost ratio of SPT over DB-SMT increases with increasing correlation degree.

In each of the results, it can be seen that for most of the correlation degrees, DB-SMT with a lower delay bound results in a higher cost ratio than DB-SMT with a higher delay bound. However, this trend is completely reversed when $\rho = 1$. Also notice that when delay bound is higher than $1.5L$, the cost ratio between SPT and SMT is less than one for some of the lower correlation degrees. This in turn implies that SPT is a more efficient structure for aggregation than DB-SMT under those circumstances. These trends are counter-intuitive, because it is expected that higher delay bounds assist in path sharing and hence energy cost reduction in DB-SMT by traversing as many nodes as possible at an early stage of the aggregation path. However, results indicate that increasing the delay constraint and hence extending the path for more aggregation does not bring in improvements in energy efficiency for most of the partially-correlated ($\rho < 1$) cases. The reasoning for this observation is as follows.

When the simulation results were further analyzed, it turned out that as the maximum path length (delay constraint) increases, the average path length for a DB-SMT also increases. For example, when delay bound was 10 hops, the corresponding average path length was 7.8 hops. However, when the delay bound was 20 hops, the average path length increased to 10.2 hops. The average path length of an aggregation tree has two conflicting impacts on its energy efficiency. On one hand, the smaller the average path length, the lesser the number of hops (transmissions) towards the sink and hence lower energy cost. On the other hand, a shorter average path length also implies lesser room for aggregation, leading to a lower energy efficiency. The relative impact of the two components and the net resulting impact on energy efficiency is in turn dependent on the correlation degree of the sensor data. For lower correlation degrees, the room for aggregation is inherently low. Hence, a shorter average hop length would help reduce the energy cost. Due to this reason, DB-SMTs with lower delay constraints that facilitate explicit aggregation while at the same time maintaining smaller average path length perform the best. However, at high correlation degrees, the larger room for aggregation and hence cost reduction overcomes the additional cost due to increased average hop length, resulting in DB-SMTs with larger delay constraints performing the best. These observations and results clearly indicate that SMT serves to be the optimal aggregation structure only when data from different sources are highly correlated. For scenarios where correlation between the sensor data is low, SPT or DB-SMT with lower delay bound is a better structure for energy efficiency.

From the study of cost ratio variation with correlation degree, we obtain the following insights:

Energy efficiency of DB-SMT increases with correlation degree. DB-SMT with the lowest delay bound proves to be the most energy efficient for low to moderate correlation degrees. Higher delay bounds helps improve aggregation efficiency only when the correlation degree is relatively high. The high correlation degress also ensure the optimality of the SMT structure.

E. Varying Delay Bounds

One of the objectives of this work is also to understand the limit of data gathering delay bounds on the energy efficiency of correlation aware aggregation trees. In all the results discussed thus far, we present curves corresponding to delay bounds (D) from L to $1.2L$, $1.5L$, $1.7L$, and $2.0L$. To study the variation of cost ratio with respect to delay bounds in depth, results from some of the simulations ($\rho = 1$) are re-plotted in Figure 6.

It can be clearly seen that the cost ratio increases with increasing delay bounds, which indicates that less restrictive delay tolerance helps improve the aggregation and hence the cost efficiency.

Higher delay bounds imply that the aggregation path can be longer in order to maximize en-route aggregation. Both $D = 1.2L$ and $D = 1.5L$ result in significant cost improvement over $D = L$ scenario. However, the growth of cost ratio slows down and tends to saturate after $D = 1.7L$. This is a very interesting observation. Generally speaking, the intuition is that the longer a path is, the more data packets can be aggregated en-route. Thus, higher delay bounds allow the creation of aggregation trees with lower cost. But simulation results show otherwise: aggregation path longer than twice the longest shortest path do not help significantly in reducing the cost.

To understand this phenomenon better, let's revisit the structures of SMT and SPT for $n = 1200$ in figure 2 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. The difference is that there are lesser number of shortest paths in the backbone of SMT. Sources not on the ”backbone” are connected by ”branches” to the backbone structure. 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.

However, notice that all the above discussions are pertaining to $\rho = 1$ correlation model. If the correlation degree of sensor data is low, then a lower delay bound would yield a better performance. Hence, the cost ratio trend would reverse in that case for low correlation degrees.

From the study of cost ratio variation with respect delay bounds, we have the following insight:

The cost ratio of SPT over DB-SMT increases as delay bound increases for high correlation degrees and tends to saturate. Further, delay bounds beyond twice the maximum shortest path length do not help reduce the DB-SMT cost further in this case. However, the cost ratio tends to decrease as delay bound increases when correlation degrees are low.

F. Summary

In this subsection, we summarize all the observations and insights derived from simulation studies.

- We have shown that the cost ratio of SPT over DB-SMT increases with node density in the sensor network, but tends to saturate with increasing node density.
- Further, when node density is high, the cost ratio of SPT over DB-SMT decreases with increasing source density. However, at low node density, a moderate source density delivers the best cost improvement.
- With respect to the impact of source distribution on aggregation efficiency, we observe that cost ratio of SPT over DB-SMT increases as the distribution of source nodes tends closer to uniform distribution.
- For different correlation models, we find that the energy efficiency of DB-SMT increases with correlation degree, and DB-SMT with the lowest delay bound is the most energy efficient for low to moderate correlation degrees. Higher delay bounds help improve aggregation efficiency only when correlation degree ρ is sufficiently high. The high correlation degree also ensures the optimality of SMT.
- Most importantly, the energy delay tradeoff of correlation aware and unaware tree can be summarized as follows: The cost ratio of SPT over DB-SMT increases as delay bound increases for high correlation degrees. Delay bounds beyond twice the maximum shortest path length do not help reduce DB-SMT cost further. Furthermore, the cost ratio tends to decrease as delay bound increases for low correlation degrees.

Finally, we highlight two major observations we inferred from simulation study:

1. The cost ratios of SPT over DB-SMT scales very slowly (tends to saturate) with respect to node density.
2. Increasing delay bound beyond a (small) constant order of the longest shortest path length does not help reduce aggregation tree cost further.

IV. Analytical Reasoning

In this section, we theoretically substantiate the slow rate of growth of the cost ratio of SPT over SMT with respect to node density. Specifically, we show that the expected (energy) cost improvement obtained by a SMT over SPT scales very slowly (as $\sqrt{\log n}$) with node density.

Before going into the details of the proofs, we present the details of the SPT and SMT structures considered in estimating the costs.

A. Expected SPT Cost

We consider a network graph where nodes are uniformly distributed in a unit area disk and the root of the SPT tree is at the center of the disk. For the convenience of analysis, we divide the network into layers of concentric rings, each ring consisting of all the nodes that are at the same distance (in terms of hops) away from the sink, i.e. nodes in between the i^{th} and $(i-1)^{th}$ rings are assumed to be i hops away from the sink. The distribution of nodes and sources are assumed to be uniform in the network. The uniform distribution of nodes assumed in the network corresponds to a poisson point process with a certain rate λ. A property of such a poisson point process is that the expected number of nodes in a certain subregion with area A is equal to $A * \lambda$. Hence, the expected number of nodes that are i hops away from sink increases with i^2.

In a SPT structure, each source is connected to the sink located at the center of the unit disk. For sources further away from the sink, the shortest paths can be considered to be independent of each other with a high probability. However, at a certain distance away from the sink, all shortest paths tend to converge, and the nodes within this distance belong to at least one of the shortest paths with a high probability. Hence, we assume that there exists a threshold distance and hence ring i^* exists, such that for all $i \leq i^*$, all nodes on ith ring are part of the SPT structure. However, for rings beyond ring i^*, only some of the nodes on each ring will be part of the SPT structure.

Based on the SPT structure defined above, we define a relaxed SPT structure SPT_r, such that the cost of SPT_r is higher than that of SPT. This relaxed SPT structure also consists of two components. The first component (SPT_0) is a SPT spanning all the nodes within i^* hops from the sink. And the second component (SPT_1) is a set of independent shortest paths such that each path connects exactly one source to a leaf on SPT_0. In other words, paths on SPT_1 never overlap with each other. Thus, the cost of SPT_r is always higher than the cost of SPT.

Now, let m be the hop number of longest shortest path in the network, n and s be the total number of nodes and sources in the network respectively.

The expected cost of SPT_r (C_{spt}) is given by

$$E[C_{sptr}] = E[C_1] + E[C_2] \tag{1}$$

Since all nodes in SPT_0 component are part of the SPT structure, the net cost of C_1 is contributed by the number of nodes in the SPT_0 component, which in turn is given by the number of nodes with the i^{*th} ring. According to the property of poisson point process for uniform node distribution, we have, and we have

$$E[C_1] = \sum_{1}^{i^*} E[n_j] = \frac{i^{*2}}{m^2} n \tag{2}$$

To obtain $E[C_2]$, we condition the product of the number of sources present in a ring j ($i^* < j \leq m$) along with their shortest distance to a leaf in SPT_0. This results in,

$$\begin{aligned} E[C_1] &= \sum_{j=i^*+1}^{m} s_j d_j = \frac{m^2 - i^{*2}}{m^2} s \times \frac{2}{3}(m - i^*) & (3) \\ &= \frac{2}{3}\frac{s}{m^2}(m - i^*)(m^2 - i^{*2}) & (4) \end{aligned}$$

where, $\frac{m^2-i^{*2}}{m^2}s$ is the total number of sources on SPT_1, and $\frac{2}{3} * (m - i^*)$ is the expected length of shortest paths on SPT_1.

Thus, the expected cost of the relaxed SPT structure from the costs of the two components (C_1 and C_2) is now given by,

$$E[C_{sptr}] = \frac{i^{*2}}{m^2} n + \frac{2}{3}\frac{s}{m^2}(m - i^*)(m^2 - i^{*2}) \tag{5}$$

The radius of transmission and hence the hop length is the minimum connectivity range defined in [4],

$$r = R\sqrt{\frac{\log 10n}{n}} \tag{6}$$

where R is the radius of the entire network.

Let n_{i^*} represents the total number of nodes on i^{*th} ring (with i^* hops), and s_j denotes the number of sources on j^{th} ring. Since the shortest paths on SPT_1 are independent, each node on i^{*th} ring is connected to at least one shortest path on SPT_1. Therefore, we have:

$$\begin{aligned} n_{i^*} &< \sum_{j=i+1}^{m} s_j & (7) \\ \Rightarrow (2i^* - 1)\frac{n}{m^2} &< \frac{m^2 - i^{*2}}{m^2} s & (8) \\ \Rightarrow si^{*2} + 2n_{i^*} - n - m^2 s &< 0 & (9) \end{aligned}$$

Solving the above inequality, we get where i^* is given by

$$\begin{aligned} i^* &= \frac{-2n \pm \sqrt{4n^2 + 4s(n + m^2 s)}}{2s} & (10) \\ &\simeq \frac{-2n \pm 2ms}{2s} & (11) \\ &\simeq m - \frac{n}{s} & (12) \end{aligned}$$

When the fraction of sources is large such that, $\frac{s}{n} > \frac{1}{m}$.

And m can be approximated as

$$\begin{aligned} m &= \frac{R}{r}\beta & (13) \\ &= 1.32\sqrt{\frac{n}{\log 10n}} & (14) \end{aligned}$$

where R is the radius of the network, r is the minimum transmission range for connectivity defined in [4] and β is a constant. This constant is introduced to account for a path connecting a furthermost node to sink not being a straight line. Plugging in the transmission range defined with 6, we get,

$$m = 1.32\sqrt{\frac{n}{\log 10n}} \tag{15}$$

B. Expected SMT Cost

Determining the cost of SMT in a network graph directly is rather difficult. However, for the sensor network environment considered, we can translate the cost of SMT in Euclidean space (ESMT) (whose cost is known directly) into the cost of SMT in network graphs (NSMT).

Lemma 1: The expected cost ESMT is $\Theta(R\sqrt{s})$.

Proof: From [7], the cost of a minimum spanning tree in Euclidean space (EMST) has the following upper bound:

$$E[C_{EMST}] \leq 0.707\sqrt{s}R + o(\sqrt{s}) \tag{16}$$

and has the following lower bound:

$$E[C_{EMST}] \geq \frac{1}{2}R\frac{s-1}{\sqrt{s}} \tag{17}$$

Combining the two bounds, we have:

$$E[C_{EMST}] = \Theta(R\sqrt{s}) \tag{18}$$

On the other hand, it is shown in [8] that the cost ratio of EMST over ESMT for the same set of source nodes is bounded by a small constant:

$$\frac{E[C_{EMST}]}{E[C_{ESMT}]} < \frac{2}{\sqrt{3}} \tag{19}$$

Therefore, we have

$$E[C_{ESMT}] = \Theta(R\sqrt{s}) \tag{20}$$

Lemma 2: The expected cost of NSMT is $\Theta(m\sqrt{s})$ for the sensor network considered on network graphs.

Proof:

Note that, the distance between two sources on ESMT can be translated into hop count directly via the following relationship:

$$H = \lceil \frac{L}{r} \rceil \tag{21}$$

where H is the hop count of path between two nodes, and L is the Euclidean distance. Because such a translation maintains the order of the cost, and m is the equivalent of R in network space, we have

$$E[C_{NSMT}] = \Theta(m\sqrt{s}) \tag{22}$$

Accordingly the expected cost of a SMT in network graph can be approximated as:

$$E[C_{smt}] = c\sqrt{s}m, \tag{23}$$

where c is a constant.

C. Cost Ratio

Proposition 1: The expected cost improvement of SMT over SPT in sensor network graph increases at $\Theta(\sqrt{\log n})$, where n is the total number of node in the sensor network, and s is $\Theta(n)$.

Proof: Combining equations 5 and 23, and observing the fact that the SPT_r structure considered for the analysis is a relaxed variant of the actual SPT structure, we obtain the ratio of the expected costs of the SPT and SMT structures as,

$$\begin{aligned} Cost\ Ratio &\leq \frac{E[C_{sptr}]}{E[C_{smt}]} \qquad (24) \\ &= \frac{\frac{i^{*2}}{m^2}n + \frac{2s}{3m^2}(m-i^*)(m^2-i^{*2})}{cm\sqrt{s}} \qquad (25) \end{aligned}$$

where m and i^* are computed using equation (15) and (10). Plugging in m and i^*, we get:

$$\begin{aligned} E[C_{sptr}] &= \Theta(\sqrt{n\log n}) + \Theta(n) \qquad (26) \\ &= \Theta(n) \qquad (27) \end{aligned}$$

and

$$E[C_{smt}] = \Theta(\frac{n}{\sqrt{\log n}}) \tag{28}$$

Combining the above two equations, we get:

$$Cost\ Ratio = \Theta(\sqrt{\log n}) \tag{29}$$

From the above analysis, we arrive at the conclusion that the cost ratio of SPT over SMT increases only with $\Theta(\sqrt{\log n})$ for large n. This increase rate is responsible for making the cost ratio improvement saturate at high node densities in the simulations. Hence, we can expect that when n is sufficiently large, the energy improvement of SPT over SMT tends to saturate. Although theoretically speaking, the cost ratio still increases as a function of n, practically the improvement in energy efficiency provided by such a slow increasing rate is negligible beyond a certain node density. Consequently, for large scale sensor networks, the energy improvement of correlation aware aggregation trees is not as significant as normally expected. We discuss the practical implications of this observation in the next section.

V. Practical Implications

A. Practical Implications of Limited Energy Improvement

As inferred in Section III, the energy improvement of correlation aware tree structures over correlation unaware tree structures is bounded by a small factor of two for all the scenarios we simulated. We also shown through analysis that cost ratio increases as $\Theta(\sqrt{\log n})$. This indicates that even for a large scale sensor network with node densities greater than those simulated in this work, the perceivable energy improvement will still be limited due to the slow rate of energy improvement.

This observation implies that correlation aware aggregation data gathering may not always be a good choice for sensor data gathering. As discussed in section I, explicit communication is required for setting up correlation aware aggregation trees. Furthermore, for highly dynamic sensor applications where sources change rapidly with time, the overhead of tree construction may offset the energy benefits resulting from correlation aware aggregation. It is also possible that the source nodes that are going to report data packets to sink are not known a priori, in which case correlation aware aggregation trees cannot be computed before the data gathering process.

On the contrary, correlation unaware trees such as shortest path trees can usually be established in a distributed fashion or pro-actively before the data gathering process. Furthermore, different shortest path trees for various sets of sources can be derived from the same shortest path tree over the sink and all sensor nodes by trimming branches that are entirely over non-source sensor nodes. The cost of explicit tree construction is eliminated for correlation unaware aggregation trees. Thus, given the cost and feasibility issues involved in the construction of correlation aware trees and the moderate energy improvement possible, correlation unaware approach may be a more desirable choice under several circumstances.

B. Practical Implications of Limited Delay Tolerance

We also observed that increasing delay tolerance does not always help reduce the aggregation tree cost. With increasing delay constraint, the cost of the SMT structure reduces. But this is not always true: beyond a certain delay tolerance, which is comparable to the longest shortest path length, the cost ratio improvement tends to saturate. This is because, it becomes possible to construct the optimal aggregation tree for the given network and delay tolerance. Practically, this means that an application does not have to be designed with large delay tolerances to ensure energy efficiency that is close to the maximum possible.

VI. Related Works

In this section, we discuss related works that have done similar studies as ours presented in this paper. For each related work, we explain its scope of study and introduce observations and results made by the authors. The similarities and differences between their results and ours are compared, and reasons for those differences are identified.

A. Related Works on Correlation Aware Aggregation Trees

[1] provided in-depth discussions related to efficient data gathering structure, and proved that the generation of an optimal aggregation structure is a NP-complete problem. Two heuristics: leaves deletion algorithm and SPT/TSP balanced tree algorithm are proposed to approximate the optimal tree and the cost ratios of approximation trees over SPTs are presented in this paper. The results presented in this paper are similar to ours in that the cost ratio of optimized tree over SPT is bounded by 0.5. This paper made a similar observation to ours that SMT resembles a combination of a SPT core and TSP paths in the outskirts. A SPT/TSP balanced tree is proposed as an approximation of optimal tree according to this observation. For an aggregation tree, SPT is built for nodes within a radius $q(\rho)$ from the root, and for the rest of the nodes, TSP paths are used to connect sources in a certain subregion to the existing shortest path tree. The advantage of this algorithm is that correlation degree is taken into account during tree construction process. But this structure is not adaptive to number of nodes in the network. As illustrated in figure 2, when the number of nodes is high, SPT is rather inefficient even at area close to sink due to path diversity. Therefore, we speculate that performance of this approximation algorithm degrades as node number increases.

[9] studied the energy efficiency of aggregation tree to some extent. But the main focus of this paper is to propose an approximation of SMT called Greedy Incremental Tree(GIT) and study its performance, therefore the scope of this paper is different from this paper. In its simulation study, [9] compared the energy dissipation of GITs and SPTs. Since the energy model they use is different with ours, their results is not directly comparable to ours. Nonetheless, this paper pointed out that SPT and GIT are similar in low density networks but achieve significant energy savings at higher node densities (each node has more neighbors). This observation is similar to the observation we made in this paper.

B. Related Works on Data Aggregation Tree Efficiency

[6] first systematically studied data-centric routing approaches in wireless sensor networks. However, the focus of [6] is on comparing data-centric routing with traditional end-to-end routing scheme(address-centric routing). In this paper, address centric routing scheme is defined as shortest path tree without aggregation (overlapped paths are counted separately). Therefore, the emphasis of [6] is to compare the performance differences between “aggregate” and “do not aggregate”, while our work investigates energy cost differences between aggregation aware and unaware schemes (SPT with aggregation and DB-SMT tree). So the focus of the two works are different.

[2] compares two major classes of data aggregation scheme: routing-driven compression (RDC) and compression-driven routing (CDR) across a broad range of spatial correlations. In this paper, RDC routes data through shortest paths toward sink, and performs opportunistic aggregations when routes overlap with each other. For CDR, routes are selected in order to compress data from all sources sequentially. This work mainly investigate the impact of correlation degrees on optimal aggregation structure. While for our study, we consider not only correlation degrees, but delay bounds and other parameters when comparing efficiency of correlation aware and unaware data aggregation trees. [2] uses grid topologies to compare RDC and CDR performance for different correlations, and we use more general uniform distribution topologies. Therefore, in this paper, the cost ratio of RDC over CDR is higher than the bound we observed in our study. We suspect that this difference is caused by the grid topology used in their paper. Nonetheless, results from this paper indicate that

CDR outperforms RDC for high correlation scenarios, whilst CDR performs better for low correlation scenarios, which is comparable to our conclusion in more general network settings.

VII. Conclusions

In this paper we study the energy efficiency of correlation aware aggregation trees in wireless sensor networks. Sensor applications with and without delay tolerance are considered, and how delay tolerance and other network conditions affect the efficiency of an correlation aware aggregation tree is explored. Through quantitative study and analysis, we conclude two rather surprising results: the energy improvement in using correlation aware aggregation is not significant under many network scenarios compared to the cost and complexity incurred in the tree construction process; and the maximum useable delay bound required to achieve the best possible energy efficiency is not high compared with the delay along the maximum length shortest-path in the default shortest path tree. Practical implications of these results have also been identified.

References

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

[2] Pattem S., Krishnamachari B., and Govindan R., "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.

[3] W. Yuan, S. Krishnamurthy, and S. K. Tripathi, "Synchronization of multiple levels of data fusion in wireless sensor networks," in *IEEE Global Communications Conference (GLOBECOM'03)*, December 2003.

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

[5] Salama H.F., Reeves D.S., and Viniotis Y., "Evaluation of multicast routing algorithms for real-time communication on high-speed networks," in *IEEE Journal on Selected Areas in Communications*, April 1997, vol. 15, pp. 332 – 345.

[6] Bhaskar Krishnamachari, Deborah Estrin, and Stephen B. Wicker, "The impact of data aggregation in wireless sensor networks," in *Proceedings of the 22nd International Conference on Distributed Computing Systems*, 2002, pp. 575–578.

[7] Ning-Yang B. Wang and Reui-Chuan Chang, "An Upper Bound for the Average Length of the Euclidean Minimum Spanning Tree," in *J. Computer Math*, 1989, vol. 30, pp. 1–12.

[8] Ding-Zhu Du and Frank K. Hwang, "An approach for proving lower bounds: Solution of gilbert-pollak's conjecture on steiner ratio," in *FOCS 1990*, pp. 76–85.

[9] 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.