On the Construction of Data Aggregation Tree with Minimum Energy Cost in Wireless Sensor Networks: NP-Completeness and Approximation Algorithms

Tung-Wei Kuo and Ming-Jer Tsai
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan, ROC
E-mail: mjtsai@cs.nthu.edu.tw

Abstract—In many applications, it is a basic operation for the sink to periodically collect reports from all sensors. Since the data gathering process usually proceeds for many rounds, it is important to collect these data efficiently, that is, to reduce the energy cost of data transmission. Under such applications, a tree is usually adopted as the routing structure to save the computation costs for maintaining the routing tables of sensors. In this paper, we work on the problem of constructing a data aggregation tree that minimizes the total energy cost of data transmission in a wireless sensor network. In addition, we also address such a problem in the wireless sensor network where relay nodes exist. We show these two problems are NP-complete, and propose $O(1)$-approximation algorithms for each of them. Simulations show that the proposed algorithms each have good performance in terms of the energy cost.

I. INTRODUCTION

In many applications, sensors are required to send reports to a specific target (e.g. base station) periodically [1]. In habitat monitoring [2] and civil structure maintenance [3], it is a basic operation for the sink to periodically collect reports from sensors. Since the data gathering process usually proceeds for many rounds, it is necessary to reduce the number of the packets, which carries the reports, transmitted in each round for energy saving. In this paper, we undertake the development of data gathering in wireless sensor networks.

Data aggregation is a well-known method for data gathering, which can be performed in various ways. In [1], a fixed number of reports received or generated by a sensor are aggregated into one packet. In other applications, a sensor can aggregate the reports received or generated into one report using a divisible function (e.g. SUM, MAX, MIN, AVERAGE, top-k, etc.) [4].

The effectiveness of data aggregation is mainly determined by the routing structure. In many data aggregation algorithms, a tree is used as the routing structure [5], [6], [7], especially for the applications that have to monitor events continuously. The reason is that sensors, which usually have limited resources, can save relatively high computational costs for maintaining routing tables if sensors route packets based on a tree. While several papers target at the maximization of the network lifetime [5], [6], it is sometimes desirable to minimize the energy cost. For example, in rechargeable sensor networks [8], [9], as it is hard to predict the energy replenishment profile, minimizing the energy cost is a simple way to prolong the network lifetime. For some indoor applications, sensors may have AC power plugs. Under such circumstance, energy saving then becomes the major issue. In this paper, the problem of constructing a data aggregation tree with minimum energy cost will be studied.

The remainder of this paper is organized as follows. Section II describes the network model and shows the MECAT problem is NP-complete. Section III provides a 2-approximation algorithm for the MECAT problem. In Section IV, we show the MECAT_RN problem is NP-complete and give a 7-approximation algorithm. Using simulations, we evaluate the performances of the proposed algorithms in Section V. Related works are studied in Section VI. Finally, we conclude the paper in Section VII.

II. NETWORK MODEL AND PROBLEM DEFINITION

A. The Network Model

We model a network as a connected graph $G = (V, E)$ with weights $s(v) \in \mathbb{Z}^+$ and 0 associated with each node $v \in V \setminus \{r\}$ and $r$, respectively, where $V$ is the set of nodes, $E$ is the set of edges, and $r \in V$ is the sink. Each node $v$ has to send a report of size $s(v)$ to sink $r$ periodically in a multi-hop fashion based on a routing tree. A routing tree constructed for a network $G = (V, E)$ with sink $r$ is a directed tree $T = (V_T, E_T)$ with root $r$, where $V_T = V$ and a directed edge $(u, v) \in E_T$ only if an undirected edge $(u, v) \in E$. A node $u$ can send a packet to a node $v$ only if $(u, v) \in E_T$, in which case $u$ is a child of $v$, and $v$ is the parent of $u$. For the energy consumption, we only consider the energy cost of the radio [6]. Let $Tx$ and $Rx$ be the energy needed to send and receive a packet, respectively. While routing, a hop-by-hop aggregation is performed according to the aggregation ratio, $q$, which is the size of reports that can be aggregated into
one packet. The aggregation ratio is assumed to be an integer through this paper.

Example 1. Fig. 1(b) is a routing tree constructed for the wireless sensor network shown in Fig. 1(a). Assume the aggregation ratio is 3, and both $Tx$ and $Rx$ are equal to 1. Using the routing tree, node 6 first sends a packet containing its report to node 7. After node 7 receives the packet from node 6, node 7 aggregates the reports of nodes 6 and 7 into one packet and then sends the packet to node 3. The process proceeds until node $r$ receives the reports of all nodes. Clearly, 3, 2, 2, 1, 1, 1, and 1 packets are sent by nodes 1, 2, 3, 4, 5, 6, and 7, respectively; therefore, a total of 11 packets are sent (and received) by the nodes.

B. The Problem and Its Hardness

We first describe our problem in the following.

Problem 1. Given a network $G = (V, E)$ with weights $s(v) \in Z^+$ and 0 associated with each node $v \in V \setminus \{r\}$ and $r$, respectively, a sink $r \in V$, an aggregation ratio $q \in Z^+$, energy costs $Tx \in R^+$ and $Rx \in R^+$ for transmitting and receiving a packet, respectively, and $C \in R^+$, the Minimum Energy-Cost Aggregation Tree (MECAT) problem asks for a routing tree $T = (V_T, E_T)$ with root $r$ and $V_T = V$, such that the total transmission and reception energy consumed by all sensors is not greater than $C$. In addition, $MECAT(G, r, q, Tx, Rx, C)$ denotes an instance of the MECAT problem, and $COST(T)$ denotes the energy cost of a routing tree $T$.

By a polynomial-time reduction from the load-balanced semi-matching problem [10], we can prove the MECAT problem is NP-complete.

III. APPROXIMATION ALGORITHM

As the MECAT problem is NP-complete, we provide an approximation algorithm. Observe that while sending a packet to the sink, the longer the routing path is, the greater the energy cost is. Naturally, we would route each packet via a shortest path to the sink. The resulting routing structure is then a shortest path tree. There are at least two benefits to route packets using a shortest path tree. First, a shortest path tree is easy to construct in a distributed manner, as described in the following two steps. The sink node first broadcasts a message such that each node can evaluate the hop distance from the sink [11], Then, each node sets its parent to the node with a smaller hop distance from the sink. Second, in many time-critical applications, it is necessary to route packets using a shortest path tree to achieve the minimum packet transmission delay. Although a shortest path tree may not have minimum energy cost, Theorem 1 shows a shortest path tree algorithm has an approximation ratio of 2.

Theorem 1. Every shortest path tree algorithm is a 2-approximation algorithm.

Proof: Let $T$ be a routing tree. Since the number of packets sent by nodes equals that received by nodes in $T$, we have

$$COST(T) = (Tx + Rx) \sum_{v \not= r} \left\lceil \frac{des_T(v) + s(v)}{q} \right\rceil.$$ 

Let $T_{OPT}$ be a routing tree with minimum energy cost, we then obtain

$$COST(T_{OPT}) = (Tx + Rx) \sum_{v \not= r} \left\lceil \frac{des_{T_{OPT}}(v) + s(v)}{q} \right\rceil < (Tx + Rx) \sum_{v \not= r} \left( \frac{des_{T_{OPT}}(v) + s(v)}{q} + 1 \right) \leq 2COST(T_{OPT}).$$

On the other side, it is easy to show that

$$T_{SPT} = \arg \min_T (Tx + Rx) \sum_{v \not= r} \frac{des_T(v) + s(v)}{q}.$$ 

Thus, we have

$$2COST(T_{OPT}) \geq (Tx + Rx) \sum_{v \not= r} \left( \frac{des_{T_{SPT}}(v) + s(v)}{q} + 1 \right) > COST(T_{SPT}).$$

IV. DATA AGGREGATION WITH RELAY NODES

To improve the network connectivity or survivability, the relay node placement problem in a wireless sensor network has been extensively investigated in recent years. These relay nodes, which do not produce reports, are used to forward the packets received from other nodes. In this section, we study the problem of constructing a data aggregation tree with minimum energy cost in the presence of relay nodes.

A. The Problem and Its Hardness

Here, a routing tree only needs to span all non-relay nodes. For the convenience of description, we assume every relay node has a zero-sized report. In the following, the problem is described and shown to be NP-complete.

Problem 2. Given a network $G = (V, E)$ with weights $s(u) \in Z^+$ and 0 associated with each source $u \in U \subseteq V \setminus \{r\}$ and $v \in V \setminus U$, respectively, a set of sources $U$, a sink $r \in V$, an aggregation ratio $q \in Z^+$, energy costs $Tx \in R^+$ and $Rx \in R^+$ for transmitting and receiving a
Given a graph \( G = (V, E) \) with weight \( w(e) \in \mathbb{R}^+ \) associated with each edge \( e \in E \) indicating the length and weight \( s(u) \in \mathbb{Z}^+ \) associated with each source \( u \in U \subseteq V \) indicating the demand size to route to sink \( r \in V \), a set of sources \( U \), a sink \( r \), and a transmission facility capacity \( q \in \mathbb{Z}^+ \), the Capacitated Network Design (CND) problem is to find a path from \( u \) to sink \( r \) for each source \( u \in U \), such that the total cost of installing all facilities is minimized, where the cost of installing \( k \) facilities on an edge with length \( l \) is \( k \cdot l \). Note that a node might have multiple outgoing edges in a feasible solution of the CND problem. That is, a feasible solution of the CND problem might not be a tree. Moreover, \( \text{CND}(G, U, r, q, C) \) denotes an instance of the CND problem, and \( \text{COST}_{\text{CND}}(R) \) denotes the cost of installing facilities of a route \( R \).

**Definition 1.** [14] Given a graph \( G = (V, E) \) with weight \( w(e) \in \mathbb{R}^+ \) associated with each edge \( e \in E \), a spanning tree \( T \) rooted at \( r \) is called an \((\alpha, \beta)\)-LAST, where \( \alpha \geq 1 \) and \( \beta \geq 1 \), if the following two conditions are satisfied:

1. For every node \( v \), the distance from \( v \) to \( r \) in \( T \) is at most \( \alpha \) times the minimum distance from \( v \) to \( r \) in \( G \).
2. The weight of \( T \) is at most \( \beta \) times that of the minimum spanning tree of \( G \).

**Algorithm 1:** Salman’s Algorithm for the CND Problem

**Input:** \( G, U, r, q, C \)

1. Construct a complete graph \( G' \) with node set \( U \cup \{r\} \).
2. Set the weight of each edge \((u, v)\) in \( G' \) to the length of the shortest path from \( u \) to \( v \) in \( G \).
3. Compute a \((3, 2)\)-LAST \( T_L \) in \( G' \).
4. Let \((u, u_1, \cdots, u_n, r)\) be the shortest path from \( u \) to \( r \) in \( T_L \). Then, the concatenation of paths \( P_{u, u_1}, P_{u_1, u_2}, \cdots, P_{u_n, r} \) is the output path from \( u \) to \( r \), where \( P_{x,y} \) denotes the shortest path from \( x \) to \( y \) in \( G \).
5. Return the output path from \( u \) to \( r \) for each \( u \in U \).

**Algorithm 2:** Our Algorithm for the MECAT_RN Problem

**Input:** \( G, U, r, q, T_x, R_x, C \)

1. Construct a complete graph \( G' \) with node set \( U \cup \{r\} \).
2. Set the weight of each edge \((u, v)\) in \( G' \) to the length of the shortest path from \( u \) to \( v \) in \( G \).
3. Compute a \((3, 2)\)-LAST \( T_L \) in \( G' \).
4. Compute \( G'' = (V'', E'') \), where \( V'' = \{w|w \in P_{u,v} \text{ for some } (u, v) \in T_L\} \), \( E'' = \{\{x, y\}|\{x, y\} \in P_{u,v} \text{ for some } (u, v) \in T_L\} \), and \( P_{u,v} \) is the shortest path from \( u \) to \( v \) in \( G \).
5. Construct a shortest path tree \( T_{SPT} \) rooted at \( r \) and spanning \( U \) in \( G'' \).
6. Return \( T_{SPT} \).

Theorem 4 shows Algorithm 2 is a 7-approximation algorithm of the MECAT_RN problem. Lemma 1, derived from the proof of Lemma 2.1 in [12], is used in the proof of Theorem 4. We omit the proof of Lemma 1 due to the page limit.

**Lemma 1.** Let \( R = \bigcup_{u \in U} P_{u,r} \) be a route of the CND problem, where \( P_{u,r} \) is the routing path from source \( u \) to sink \( r \), and let \( R_{OPT} \) be the route with minimum cost of the CND problem. Then, \( \text{COST}_{\text{CND}}(R) \leq (\alpha' + \beta')\text{COST}_{\text{CND}}(R_{OPT}) \), if the following two conditions are satisfied:

1. For every source \( u \), the length of \( P_{u,r} \) is at most \( \alpha' \) times the minimum distance from \( u \) to \( r \) in \( G \).
2. The total lengths of the edges of \( R \) is at most \( \beta' \) times that of the Steiner tree of \( G \) spanning \( U \).

**Theorem 4.** Algorithm 2 is a 7-approximation algorithm of the MECAT_RN problem.

**Proof:** Let Algorithm \( A \) be obtained from Algorithm 2 by replacing Line 2 with Line 2 of Algorithm 1 and modifying Line 6 to return the path from \( u \) to \( r \) in \( T_{SPT} \) for each \( u \in U \) instead of \( T_{SPT} \). We first claim that Algorithm \( A \) is a 7-approximation of the CND problem. Let \( R_1 \) and \( R_A \) be
the solutions generated by Algorithms 1 and A, respectively. Clearly, the following two facts hold:

1) For every source $u$, the length of $P_{u,r}$ in $R_A$ is less than that in $R_1$.
2) The total lengths of the edges of $R_A$ is less than that of $R_1$.

[12] tells us that the length of $P_{u,r}$ in $R_1$ is at most 3 times the minimum distance from $u$ to $r$ in $G$ for every source $u$ and the total lengths of the edges of $R_1$ is at most 4 times that of the Steiner tree of $G$ spanning $U$. Thus, Algorithm $A$ is a 7-approximation of the CND problem by Lemma 2.

Next, given MECAT_RN$(G_1, U, r, q, Tx, Rx, C)$, we construct CND$(G_2, U, r, q, C)$, where $G_2$ is obtained from $G_1$ by setting the weight of each edge to $T_x + Rx$. Let $T_2$ and $T_{OPT}$ be a routing tree generated by Algorithm 2 and the routing tree with minimum energy cost for MECAT_RN$(G_1, U, r, q, Tx, Rx, C)$, respectively. Let $R_A$ and $R_{OPT}$ be a route generated by Algorithm A and the route with minimum cost of installing facilities for CND$(G_2, U, r, q, C)$, respectively. Note that for each $u \in U$, the sequence of the nodes in the path from the node $u$ to $r$ in $T_2$ is equal to that in the path from $u$ to $r$ in $R_A$. Thus, $COST(T_2) = COST_{CND}(R_A)$. It is also noted that a collection of the path from $u$ to $r$ in $T_{OPT}$ for each $u \in U$ can be a route $R$ for CND$(G_2, U, r, q, C)$, in which case $COST(T_{OPT}) = COST_{CND}(R)$. It implies $COST_{CND}(R_{OPT}) \leq COST(T_{OPT})$. Finally, due to the fact that $COST_{CND}(R_A) \leq 7COST_{CND}(R_{OPT})$, we obtain $COST(T_2) \leq 7COST(T_{OPT})$.

V. NUMERICAL RESULTS

Two simulations were conducted here. In the first and second simulations, algorithms of the MECAT problem (data aggregation without relay nodes) and the MECAT_RN problem (data aggregation with relay nodes) were compared respectively. We also compared our algorithms with the lower bound of the minimum energy cost $LB$ evaluated by $LB = (Tx + Rx) \cdot \max \{\sum_{u \in U}^{s(u)} l(u,r), |E(T_{Steiner})|\}$, where $U$ is the set of sources, $q$ is the aggregation ratio, $l(u,r)$ is the hop distance from $u$ to $r$ in a shortest path tree and $|E(T_{Steiner})|$ is the number of edges in a Steiner Tree. This is due to the fact that the corresponding minimum energy cost of the MECAT problem and the MECAT_RN problem is impossible to be smaller than each of $(Tx + Rx) \cdot \sum_{u \in U}^{s(u)} l(u,r)$ and $(Tx + Rx) \cdot |E(T_{Steiner})|$. Since a Steiner tree cannot be obtained in polynomial time, we use a 2-approximation algorithm to construct a Steiner tree $T(V_{ST}, E_{ST})$ [15], and evaluate $|E(T_{Steiner})|$ as $\max \{\frac{|E_{ST}|}{2}, |U|\}$.

In a wireless sensor network, 100 sensor nodes were uniformly deployed in a $100 \times 100$ field. A link exists between two sensor nodes with distance less than or equal to the transmission range $R = 20$. Since the transmission power is about two times the reception power [16], $Tx$ and $Rx$ are set to 2 and 1 respectively. If all reports have the same size (uniform report size), the size is set to 1; otherwise (non-uniform report size), the sizes are randomly set to range from 1 to 5. The energy cost of each algorithm approaches $LB$ as the aggregation ratio is great. Second, although a shortest path tree algorithm and a Steiner tree algorithm have bad performances in the worst cases (see Theorems 2 and 3), they have good performances in average cases. Third, a shortest path tree performs better and worse than a Steiner tree algorithm when the aggregation ratio is small and great, respectively. This is because as the aggregation ratio is great, a packet can carry a large number of reports, and thus, the energy cost highly depends on the number of edges in the data aggregation tree. On the contrary, the energy cost highly depends on the lengths of the paths from sources to the sink as the aggregation ratio is small.

VI. RELATED WORKS

In [5], an algorithm is demonstrated to find the best shortest path tree that maximizes the network lifetime. In [6], the authors prove the problem of finding an optimal aggregation...
tree that maximizes the network lifetime is NP-complete and propose an approximation algorithm.

In [1], the problem of finding a routing structure minimizing the number of transmitted packets is studied, in which neither the proof of NP-completeness nor an approximation algorithm is given. They show that routing packets on any two shortest path trees does not significantly affect the effectiveness of data aggregation. In addition, all reports are assumed to have the same size and the existence of relay nodes are not taken into consideration, which is different from this paper.

A problem similar to ours is studied in [7], but the aggregation model is different. It is assumed that any \( j \) reports can be aggregated into \( f(j) \) packets, where \( f \) is concave. However, in our data aggregation model, \( f(j) = \lceil \frac{j}{2} \rceil \) is not concave. Thus, the analysis in [7] cannot be applied here.

VII. CONCLUSION

In this paper, we study the problem of constructing energy-efficient data aggregation trees. Two types of this problem are investigated: the one without relay nodes and the one with relay nodes. Both of them are shown to be NP-complete. For the problem without relay nodes, we find that a shortest path tree algorithm turns out to be a 2-approximation algorithm and can be easily implemented in a distributed manner. For the problem with relay nodes, we first show that a shortest path tree algorithm and a Steiner tree algorithm each have bad performance in the worst cases. We then obtain an \( O(1) \)-approximation algorithm based on Salman’s algorithm for the Capacitated Network Design problem.

REFERENCES