Data Collection with Bounded-Sized Messages in Wireless Sensor Networks

Min Kyung An

Abstract—In this paper, we study the data collection problem in Wireless Sensor Networks (WSNs) adopting the two interference models: The graph model and the more realistic physical interference model known as Signal-to-Interference-Noise-Ratio (SINR). The main issue of the problem is to compute schedules with the minimum number of timeslots, that is, to compute the minimum latency schedules, such that data from every node can be collected without any collision or interference to a sink node. While existing works studied the problem with unit-sized and unbounded-sized message models, we investigate the problem with the bounded-sized message model, and introduce a constant factor approximation algorithm. To the best of our knowledge, our result is the first result of the data collection problem with bounded-sized model in both interference models.

Keywords—Data collection, collision-free, interference-free, physical interference model, SINR, approximation, bounded-sized message model, wireless sensor networks, WSN.

I. INTRODUCTION

WSNs consist of a number of tiny wireless sensor devices (nodes). These nodes are scheduled to turn on their power to emit signals (i.e., to send data), or turn it off to conserve their limited power for specific time duration. When emitting signals, a collision or interference can occur at a node if the data transmission is interfered by signals concurrently sent by other nodes. In this case, the data should be re-transmitted. Because the tiny nodes have limited energy resources, it is crucial to reduce such unnecessary retransmissions in order to prolong the network’s lifetime.

One important task of a WSN is to collect data periodically and send (forward) the data to a sink node in the network. This type of application is commonly known as data collection. An interesting approach for the data collection is to assign timeslots to nodes to obtain a good schedule through which data from every node is collected to the sink node. Here, if nodes are assigned the same timeslot in a schedule, then they can send data concurrently without causing any collision or interference. The objective of the problem is to compute schedules with the minimum number of timeslots, that is, to compute the minimum latency schedules, such that data from every node can be collected without any collision or interference.

For the data collection problem, there are three models in the literature: Unit-sized, bounded-sized, or unbounded-sized messages. In the unit-sized message model, a node can send a single unit-sized message at a timeslot, and therefore merging (combining) messages is not allowed. In the bounded-sized message model, a node can merge messages up to some limit before it sends, whereas in the unbounded-sized message model, there is no limit on the length of the merged message.

The data collection problem has been widely investigated by researchers in two interference models: The graph model and the physical interference model. In the graph model, given a transmission range \( r(u) \) for every node \( u \) (i.e., the radius of the broadcasting disk covered by the signal sent by \( u \) using its transmission power \( p(u) \)), the interference range of \( u \) is defined as \( \rho \cdot r(u) \), where \( \rho \geq 1 \) is the interference factor [2]. When \( \rho = 1 \), it is called a collision-free graph model that concerns collision only, and when \( \rho \geq 1 \), it is called a collision-interference-free graph model that concerns both collision and interference. Although the traditional graph model has been widely used in many studies, it is not an adequate model since cumulative interference caused by all the other concurrently transmitting nodes is ignored [2]. Thus, the more realistic physical interference model which is known as SINR has been used by many researchers for investigating problems in WSNs since its introduction by Gupta et al. in [3].

In the graph model, Bermond et al. [9] and Coleri et al. [23] proved the NP-hardness of the data collection problem when \( \rho \geq 1 \) and \( \rho = 1 \), respectively, with the unit-sized message model. With the unbounded-sized message model, the data collection is also known as data aggregation, and Chen et al. [24] and An et al. [22] proved the NP-hardness of the problem with \( \rho = 1 \) and \( \rho \geq 1 \), respectively. Because of the NP-hardness of the problem, many researchers have focused on proposing approximation algorithms, and the existing approximation algorithms with the unit-sized and unbounded-sized message models are summarized in Table I. Note that results in [4]–[6], [8]–[10] apply to special topologies or general graphs only. Lastly, with the bounded-sized message model, there currently exist no studies which investigated the data collection problem, to the best of our knowledge. There exist few studies [25], [26] which investigated a related application called gossiping assuming that messages can be merged into a single message whose size is bounded by \( \log n \), where \( n \) is the number of nodes in a network.

In the SINR model, few researchers have investigated the data collection problem with the unit-sized message model, whereas there exists several studies which proposed approximation algorithms with the unbounded-sized message model, and Lam et al. [16], [17] showed the first result of the NP-hardness of the problem with the model. Like the
In this paper, we consider a WSN that consists of a set of sensor nodes deployed in a plane. Each node $u \in V$ is assigned a transmission power level $p(u)$, and its transmission range $r(u)$ is defined as the radius of the broadcasting disk covered by the signal sent by $u$ using its power $p(u)$. Accordingly, a directed edge $(u, v)$ exists from node $u$ to node $v$, if $v$ resides in $u$’s broadcasting disk (i.e., $d(u, v) \leq r(u)$, where $d(u, v)$ denotes the Euclidean distance between $u$ and $v$).

1) *Graph Model*: Let $C_u = \{ v \mid v \in V, d(u, v) \leq r(u) \}$ denote the set of nodes that reside in $u$’s transmission range. Then, two nodes $u$ and $v$ can communicate each other if $u \in C_v$ and $v \in C_u$. Next, let $I_u$ denotes the set of nodes that reside in $u$’s interference range $\rho \cdot r(u)$, where $\rho \geq 1$ is the interference factor. Then, the collision is said to occur at node $w$ if there exist other concurrently sending nodes $u$ and $v$ such that $w \in C_u \cap I_v$, where $\rho = 1$ (i.e., $C_u = I_u$). Also, the interference is said to occur at node $w$ if there exist other concurrently sending nodes $u$ and $v$ such that $w \in C_u \cap I_v$, where $\rho > 1$ (i.e., $C_u \subset I_v$).

In the graph model, we model a communication graph as a directed graph $G = (V, E)$ where $E = \{(u, v) \mid u, v \in V, d(u, v) \leq r(u) \}$ and $d(v, u) \leq r(v)$.

2) *SINR Model*: In the SINR model, when a node $u$ sends data using its power level $p(u)$, the signal sent to a receiver $v$ may not be strong enough to be received and hence the transmitted data is lost. It is because the signal sent by $u$ fades and $v$ is interfered by the cumulative interference caused by all the other concurrently transmitting nodes. In this model, the received power at the receiver $v$ is defined as $p(u) \cdot d(u, v)^{-\alpha}$, where $\alpha > 2$ is the path loss exponent, and $v$ can receive the data transmitted by the sender $u$ without any interference only if the ratio of the received power at $v$ to the total interference caused by all the other concurrently transmitting nodes and background noise is beyond an SINR threshold $\beta \geq 1$.

Formally, node $v$ can successfully receive data via the communication edge $(u, v)$ only if

$$SINR_{u,v} = \frac{p(u)}{N + \sum_{w \in X \setminus \{u, v\}} \frac{p(w)}{d(u, w)^{\beta}}} \geq \beta \geq 1 \tag{1}$$

where $N > 0$ is the background noise, and $X$ is the set of other concurrently transmitting nodes.

As $u$ can send its data to the nodes within the distance $(\frac{r(u)}{\beta})^\frac{1}{\beta}$ (i.e., $r(u) = (\frac{N^\beta}{p(u)})^\frac{1}{\beta}$) only, we model the communication graph as a directional disk graph $G = (V, E)$, where $E = \{(u, v) \mid u, v \in V, d(u, v) \leq (\frac{r(u)}{\beta})^\frac{1}{\beta} \}$ and $d(v, u) \leq (\frac{r(u)}{\beta})^\frac{1}{\beta}$, as in [2].

B. Problem Definition

We define the Minimum Latency Collection Scheduling (MLCS) problem as follows. Given a set of nodes for a network in a plane, we assign every node a timeslot such that nodes assigned the same timeslot, say $t$, can send data to their receivers simultaneously, satisfying the following conditions:

- (Graph Model) Neither collision nor interference occurs at any receiver.
- (SINR Model) The SINR inequality (1) is satisfied for every receiver.

A *schedule* is defined as a sequence of such timeslots, $(t_1, t_2, \cdots, t_L)$, where $L$ denotes the latency of the schedule. A schedule is *successful* if all data of every node $v \in V - \{s\}$ is collected to a sink node $s \in V$. See Table II for notations.

III. CONSTANT FACTOR APPROXIMATION ALGORITHM

In this section, we introduce our constant factor approximation algorithm for the MLCS problem with the bounded-sized message model where each node can merge messages into a single message up to size of $K$ before it sends. We further assume that every node $u$ has its buffer storage $B(u)$ whose size is unlimited, and is assigned the transmission power level $P$, i.e., for every $u \in V, p(u) = P$.
A. Interference Models
We consider both graph and physical interference (SINR) models with the following assumptions as in [2]:

1) Graph Model: We set the maximum link length (i.e., the maximum transmission range $r$) to be the given $P$, and make the assumption that the undirected unit disk graph $G$, where $E = \{(u,v) \mid d(u,v) \leq r\}$, is connected and its interference factor $\rho \geq 1$.

2) SINR Model: From the SINR inequality (1) (Section II), we can compute the possible maximum link length as $r_{\text{max}} = \left(\frac{P}{\sqrt{N}}\right)^{\frac{2}{\alpha}}$. We do not consider the links whose length is $r_{\text{max}}$ because only node $u$ can be a sender to send its data to some receiver $v$, where $d(u,v) = r_{\text{max}}$ (i.e., other nodes cannot transmit concurrently). Thus, we consider the links $(u,v)$, where $d(u,v) \leq \delta\left(\frac{P}{\sqrt{N}}\right)^{\frac{2}{\alpha}}$, for some constant $\delta \in (0,1)$ as in [15] thereby setting $r = \delta\left(\frac{P}{\sqrt{N}}\right)^{\frac{2}{\alpha}}$. We also make the assumption that the undirected graph $G$, where $E = \{(u,v) \mid d(u,v) \leq r\}$, is connected and $\alpha > 2$ [3].

B. Algorithm
MLCS algorithm starts by constructing a collection tree $T$ which is a breadth-first-search (BFS) tree (cf. [27]) on $G$ rooted at the sink node $s$. Then, a number of iterations are performed to find a schedule based on $T$. Assigning timeslots for data collection is based on a constant value $H$. The value $H$ guarantees that for any two sender nodes $u$ and $u$’s descendant node $v$ on $T$, if $\ell(u) - \ell(v) \geq H$, then they can send data simultaneously without interference, where $\ell(u)$ denotes the level of $u$ on $T$. The constant value $H$ is set as follows in the two interference models (See Lemmas 1 and 2):

- Graph Model: $H = \lfloor \rho + 2 \rfloor$
- SINR Model: $H = \left\lfloor \frac{P \cdot 2\pi}{N (\delta)^{2}} \right\rfloor + 1$

The details of data collection scheduling are contained in Algorithms 1 and 2.
say $u$, whose $|B(u)|$ is largest, is chosen. If there are less than or equal to $K$ messages in its buffer, then extract all the $|B(u)|$ messages from the buffer; otherwise, if there are more than $K$ messages in its buffer, then extract only $|B(u)|$ messages from the buffer. Then the extracted messages are merged into a single message $M$, and the node $u$ is scheduled to forward $M$ to its parent node $parent(u)$ on $T$ at the timeslot $t$. At the timeslot $t$, $M$ is unmerged and the unmerged messages are buffered at $parent(u)$’s buffer.

IV. ANALYSIS

In this section, we analyze the Minimum Latency Collection Scheduling (MLCS) algorithm (Algorithm 1) and bound the latency of schedules produced by it.

First, we set the constant value $H$ for the graph and SINR models.

Lemma 1 (Graph Model): For an interference factor $\rho \geq 1$, let $H = \lceil \rho + 2 \rceil$. In MLCS algorithm, for any two nodes $u$ and $v$’s descendant $v$ on $T$, if $|r(u) - r(v)| \geq H$, then they can send data simultaneously without interference.

Proof: Consider a pair of sender and receiver, denoted by $s_1$ and $r_1$, and let $s_2$ be the closest sender to $r_1$ that does not interfere with $r_1$. Without loss of generality, let us assume $r_1$ is a descendant of $s_1$ and $s_2$ is a descendant of $r_1$ on $T$. Then, $d(r_1, s_2) > \rho \cdot r$. In order to bound the shortest number of hops between $r_1$ and $s_2$, assume a straight line between $r_1$ and $s_2$, and relay nodes with the power level $P$ on the line. As we are assuming that $r_1$ and $s_2$ are connected with the shortest number of hops, we need at least $\lceil \frac{\pi}{\alpha} \rceil = \lceil \rho \rceil$ relay nodes for the connection. This implies that there are at least $\lceil \rho + 1 \rceil$ hops between $r_1$ and $s_2$. Thus, in the MLCS algorithm, we can set $H = \lfloor \rho + 2 \rfloor$.

Lemma 2 (SINR Model): For SINR threshold $\beta \geq 1$, path loss exponent $\alpha > 2$, background noise $N > 0$, and some constant $\delta \in (0, 1)$, let $H = \lceil \tau + r - 1 + 1 \rceil$, where $\tau = \lceil \frac{P}{N(\delta - 1)(\alpha - 0.5)} \rceil$ and $r = \lceil \frac{P}{N} \rceil$. In the MLCS algorithm, for any two nodes $u$ and $v$’s descendant $v$ on $T$, if $|r(u) - r(v)| \geq H$, then they can send data simultaneously without interference.

Proof: Consider a sender $s_1$ trying to send its data to its farthest possible receiver $r_1$, and let $s_2$ be the closest sender to $r_1$ that does not interfere with $r_1$. Without loss of generality, let us assume $r_1$ is a descendant of $s_1$ and $s_2$ is a descendant of $r_1$ on $T$. Then $\tau = \lceil \frac{P}{N(\delta - 1)(\alpha - 0.5)} \rceil$ is a lower bound for the shortest distance between $r_1$ and $s_2$ [28], and therefore $d(r_1, s_2) \geq \tau$.

Next, let us bound the shortest number of hops between $r_1$ and $s_2$ as follows. Assume a straight line between $r_1$ and $s_2$, and relay nodes with the power level $P$ on the line. As we are assuming that $r_1$ and $s_2$ are connected with the shortest number of hops, we need at least $\lceil \tau \cdot r - 1 \rceil$ relay nodes for the connection. This implies that there are $\lceil \tau \cdot r - 1 \rceil$ hops between $r_1$ and $s_2$. Thus, in the MLCS algorithm, we can set $H = \lceil \tau \cdot r - 1 \rceil$.

Next, we bound the latency of the data collection schedules produced by the algorithm.

Lemma 3 (Lower Bound): If $n$ is the number of nodes in a network, then every data collection schedule with bounded-sized model where several messages can be merged into a single message whose size is bounded by $K$ takes at least $\lceil \frac{n}{K} \rceil$ timeslots.

Proof: Consider a node $u$, and $n - 1$ messages that the node $u$ has to receive. As a node can merge up to $K$ messages, the node must receive at least $\lceil \frac{n - 1}{K} \rceil$ distinct messages. Therefore, any data collection schedule allowed to merge messages up to size of $K$ needs at least $\lceil \frac{n - 1}{K} \rceil$ timeslots.

Theorem 4: The MLCS algorithm collects data from all the other nodes successfully to sink node $s$ with at most $H \cdot \lceil \frac{n - 1}{K} \rceil$ timeslots, and it is a constant-factor approximation with the factor of $2H$.

Proof: First note that there are $n - 1$ messages that the sink node $s$ must receive. In the MLCS algorithm, $s$ receives single merged message every $H$ timeslots, and as the subroutine, Collection-Scheduling algorithm (Algorithm 2), merges messages up to size of $K$, $s$ receives at most $\lceil \frac{n - 1}{K} \rceil$ messages to collect data without collision or interference (Lemmas 1 and 2). Therefore, it takes at most $H \cdot \lceil \frac{n - 1}{K} \rceil$ timeslots.

Next, letting $SOL$ denote the upper bound of the latency of the algorithm, and $OPT$ be the lower bound (Lemma 3), we get $SOL \leq \lceil H \cdot \frac{n - 1}{K} \rceil \leq 2H$. Thus, it is an approximation algorithm with the constant-factor of $2H$.

V. CONCLUSION

In this paper, we focused on the Minimum Latency Collection Scheduling (MLCS) problem of WSNs in the graph model as well as the more realistic physical interference model known as SINR. We proposed a $O(1)$-approximation algorithm that works in both the interference models with bounded-sized message model. To the best known of our knowledge, our result is the first result of the problem with bounded-sized model in both interference models. For future work, we plan to study another related problem, gossiping, adopting both the interference models with bounded-sized message model.

REFERENCES


