Capacity of Data Collection in Arbitrary Wireless Sensor Networks

Siyuan Chen* Minsu Huang* Shaojie Tang† Yu Wang*

***Abstract*— Data collection is a fundamental function provided by wireless sensor networks. How to efficiently collect sensing data from all sensor nodes is critical to the performance of sensor networks. In this paper, we aim to understand the theoretical limits of data collection in a TDMA-based sensor network in terms of possible and achievable maximum capacity. Previously, the study of data collection capacity [1]–[6] has concentrated on large-scale random networks. However, in most of the practical sensor applications, the sensor network is not uniformly deployed and the number of sensors may not be as huge as in theory. Therefore, it is necessary to study the capacity of data collection in an arbitrary network. In this paper, we first derive the upper and lower bounds for data collection capacity in arbitrary networks under protocol interference model and disk graph model. We show that a simple BFS tree based method can lead to order-optimal performance for any arbitrary sensor networks. We then study the capacity bounds of data collection under a general graph model, where two nearby nodes may be unable to communicate due to barriers or path fading, and discuss performance implications. Finally, we provide discussions on the design of data collection under physical interference model or Gaussian channel model.**

***Index Terms*— capacity, data collection, arbitrary networks, wireless sensor networks.**

I. Introduction

Due to their wide-range potential applications in various scenarios such as battlefield, emergency relief and environment monitoring, wireless sensor networks have recently emerged as a premier research topic. The ultimate goal of a sensor network is often to deliver the sensing data from all sensors to a sink node and then conduct further analysis at the sink node. Thus, data collection is one of the most common services used in sensor network applications. In this paper, we study some fundamental capacity problems arising from data collection in wireless sensor networks.

We consider a wireless sensor network where n sensors are *arbitrarily* deployed in a finite geographical region. Each sensor measures independent field values at regular time intervals and sends these values to a sink node. The union of all sensing values from n sensors at a particular time is called a *snapshot*. The task of data collection is to deliver these snapshots to a single sink. Due to spatial separation, several sensors can successfully transmit at the same time if these transmissions do not cause any destructive wireless interference. As in the literature, we first adopt the *protocol interference model* in our analysis and assume that a successful transmission over a link has a fixed data-rate W bit/second. Later, we relax these assumptions to more realistic models: *physical interference model* and *Gaussian channel model*.

The performance of data collection in sensor networks can be characterized by the rate at which sensing data can be collected and transmitted to the sink node. In particular, the theoretical measure that captures the limits of collection processing in sensor networks is the capacity of many-to-one data collection, *i.e.*, the maximum data rate at the sink to continuously receive the snapshot of data from sensors. *Data collection capacity* reflects how fast the sink can collect sensing data from all sensors with interference constrain. It is critical to understand the limit of many-to-one information flows and devise efficient data collection algorithms to improve the performance of wireless sensor networks.

Capacity limits of data collection in random wireless sensor networks have been studied in the literature [1]–[6]. In [1], [2], Duarte-Melo *et al.* first introduced the many-to-one transport capacity in dense and random sensor networks under protocol interference model. Both El Gamal [3] and Barton and Zheng [4] investigated the capacity of data collection with complex physical layer techniques, such as antenna sharing, channel coding and cooperative beam-forming. Liu *et al.* [5] recently studied the capacity of a general some-to-some communication paradigm under protocol interference model in random networks with multiple randomly selected sources and destinations. Chen *et al.* [6] studied the capacity of data collection under protocol interference model with multiple sinks. However, all the research above shares the standard assumption that a large number of sensor nodes are either located on a grid structure or randomly and uniformly distributed in a plane. Such an assumption is useful to simplify the analysis and derive nice theoretical limits, but may be invalid in many practical sensor applications.

In this paper, we focus on *deriving capacity bounds of data collection for arbitrary networks*, where sensor nodes can be deployed in any distribution and can form any network topology. We summarize our contributions as follows:

- For arbitrary sensor networks under protocol interference model and disk graph model (if two sensors are within the transmission ranges of each other then they can communicate), we propose a simple data collection method which performs data collection on branches of the Breadth First Search (BFS) tree. We prove that this method can achieve collection capacity of $\Theta(W)$ which matches the theoretical upper bound.
- Since the disk graph model is idealistic, we also consider a more practical network model: *general graph model*.

S. Chen, M. Huang and Y. Wang are with Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA. S. Tang is with Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois, USA.

Digital Object Indentifier 10.1109/TPDS.2011.96 1045-9219/11/$26.00 © 2011 IEEE

In the general graph model, two nearby nodes may be unable to communicate due to various reasons such as barriers and path fading. We first show that $\Theta(W)$ may not be achievable for a general graph. Then we prove that a greedy scheduling algorithm on BFS tree can achieve capacity of $\Theta(\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*})$ while the capacity is bounded by $\Theta(\frac{W}{\Delta^*})$ from above. Here, Δ^*, λ^*, and λ are three new interference related parameters defined in Section V.

- Finally, we discuss the data collection capacity under more general communication models, physical interference model and Gaussian channel model. For physical interference model, we prove that the capacity of data collection is in the same order as the one under protocol interference model. For Gaussian channel model, we derive an upper bound of data collection capacity.

The results above not only help us to understand the theoretical limits of data collection in sensor networks, but also provide practical and efficient data collection methods (including how to construct data collection structure and how to schedule data collection) to achieve near-optimal capacity. Even though we are focusing on arbitrary networks, all of our solutions can be applied to random networks since any random network is just a special case of arbitrary networks.

The rest of this paper is organized as follows. We first review related work in Section II, and then describe our network model in Section III. We study the data collection capacity under disk graph model and protocol interference model in Section IV. In Section V, we relax the disk graph model in our analysis and derive the bounds of data collection capacity in a general graph model. We discuss the collection capacity under physical interference model and Gaussian channel model in Section VI, and conclude the paper in Section VII. A preliminary conference version of this paper appeared in [7]. Due to space limit, some detailed proofs and simulation results are ignored here, and provided as *Supplemental Material*.

II. Related Work

Gupta and Kumar initiated the research on capacity of random wireless networks by studying the unicast capacity in the seminal paper [9]. A number of following papers studied capacity under different communication scenarios in random networks: unicast [10]–[12], multicast [13]–[15], broadcast [16], [17]. In this paper, we focus on the capacity of data collection in a many-to-one communication scenario.

Capacity of data collection in random wireless sensor networks has been investigated in [1]–[6]. Duarte-Melo *et al.* [1], [2] first studied the many-to-one transport capacity in random sensor networks under protocol interference model. They showed that the overall capacity of data collection is $\Theta(W)$. El Gamal [3] studied data collection capacity subject to a total average transmitting power constraint. They relaxed the assumption that every node can only receive from one source node at a time. It was shown that the capacity of random networks scales as $\Theta(\log nW)$ when n goes to infinity and the total average power remains fixed. Their method uses antenna sharing and channel coding. Barton and Zheng [4] also investigated data collection capacity under more complex physical layer models (non-cooperative SINR model and cooperative time reversal communication (CTR) model). They first demonstrated that $\Theta(\log nW)$ is optimal and achievable using CTR for a regular grid network in [18], then showed that the capacities of $\Theta(\log nW)$ and $\Theta(W)$ are optimal and achievable by CTR when operating in fading environments with power path-loss exponents that satisfy $2 < \beta < 4$ and $\beta \geq 4$ for random networks [4]. Recently, Chen *et al.* [6] have studied data collection capacity with multiple sinks. They showed that with k sinks the capacity increases to $\Theta(kW)$ when $k = O(\frac{n}{\log n})$ or $\Theta(\frac{nW}{\log n})$ when $k = \Omega(\frac{n}{\log n})$. Liu *et al.* [5] lately introduced the capacity of a more general some-to-some communication paradigm in random networks where there are $s(n)$ randomly selected sources and $d(n)$ randomly selected destinations. They derived the upper and lower bounds for such a problem. However, all research above shares the standard assumption that a large number of sensor nodes are either located on a grid structure or randomly and uniformly distributed in a plane. Such an assumption is useful to simplify the analysis and derive nice theoretical limits, but may be invalid in many practical sensor applications. To our best knowledge, our paper is *the first* to study data collection capacity for arbitrary networks.

III. Network Models and Collection Capacity

A. Basic Network Models

In this paper, we focus on the capacity bound of data collection in arbitrary wireless sensor networks. For simplicity, we start with a set of simple and yet general enough models. Later, we will relax them to more realistic models.

We consider an arbitrary wireless network with n sensor nodes $v_1, v_2, \cdots, v_n$ and a single sink v_0. These n sensors are arbitrarily distributed in a field. At regular time intervals, each sensor measures the field value at its position and transmits the value to the sink. We first adopt a *fixed data-rate channel model* where each wireless node can transmit at W bits/second over a common wireless channel. We also assume that all packets have unit size b bits. The time is divided into time slots with $t = b/W$ seconds. Thus, only one packet can be transmitted in a time slot between two neighboring nodes. TDMA scheduling is used at MAC layer.

Under the fixed data-rate channel model, we assume that every node has a fixed transmission power P. Thus, a fixed transmission range r can be defined such that a node v_j can successfully receive the signal sent by node v_i only if $||v_i - v_j|| \leq r$. Here, $||v_i - v_j||$ is the Euclidean distance between v_i and v_j. We call this model *disk graph model*. We further define a communication graph $G = (V, E)$ where V is the set of all nodes (including the sink) and E is the set of all possible communication links. We assume graph G is connected.

Due to spatial separation, several sensors can successfully transmit at the same time if these transmissions do not cause any destructive wireless interferences. As in the literature, we first model the interference using *protocol interference model*. All nodes have a uniform interference range R. When node v_i transmits to node v_j, node v_j can receive the signal successfully if no node within a distance R from v_j is transmitting simultaneously. Here, for simplicity, we assume that $\frac{R}{r}$ is a constant α which is larger than 1. Let $\delta(v_i)$ be the number of nodes in v_i's interference range (including v_i itself) and Δ be the maximum value of $\delta(v_i)$ for all nodes v_i, $i = 0, \cdots, n$. We summarize all notations used in this paper in a table given in Section VI of *Supplemental Material*.

B. Capacity of Data Collection

We now formally define delay and capacity of data collection in wireless sensor networks. Recall that each sensor at regular time intervals generates a field value with b bits and wants to transport it to sinks. We call the union of all values from all n sensors at particular sampling time a *snapshot* of the sensing data. The goal of data collection is to collect these snapshots from all sensors to the sinks. It is clear that the sink prefers to get each snapshot as quickly as possible. In this paper, we assume that there is no correlation among all sensing values and no network coding or aggregation technique is used during the data collection.

Definition 1: The **delay of data collection** D is the time used by the sink to successfully receive a snapshot, i.e., the time needed between completely receiving one snapshot and completely receiving the next snapshot at the sink.

Definition 2: The **capacity of data collection** C is the ratio between the size of data in one snapshot and the time to receive such a snapshot (i.e., $\frac{nb}{D}$) at the sink.

Thus, the capacity C is the maximum data rate at the sink to continuously receive the snapshot data from sensors. Here, we require the sink to receive the complete snapshot from all sensors (*i.e.*, data from all sensors need to be delivered). Notice that data transport can be pipelined in the sense that further snapshots may begin to transport before the sinks receiving prior snapshots. In this paper, we focus on *capacity analysis of data collection in an arbitrary sensor network*.

IV. Collection Capacity for Disk Graph Model

Upper Bound of Collection Capacity: It has been proved that the upper bound of capacity of data collection for random networks is W [1], [2]. It is obviously that this upper bound also holds for any arbitrary network. The sink v_0 cannot receive at rate faster than W since W is the fixed transmission rate of individual link. Therefore, we are interested in design of data collection algorithm to achieve capacity in the same order of the upper bound, *i.e.* $\Theta(W)$.

In this section, we propose a simple BFS-based data collection method and demonstrate that it can achieve the capacity of $\Theta(W)$ under our network model: disk graph model. Our data collection method includes two steps: data collection tree formation and data collection scheduling.

A. Data Collection Tree - BFS Tree

The data collection tree used by our method is a classical Breadth First Search (BFS) tree rooted at the sink v_0. The time complexity to construct such a BFS tree is $O(|V| + |E|)$. Let T be the BFS tree and $v_1^l, \cdots, v_c^l$ be all leaves in T. For each leaf v_i^l, there is a path P_i from itself to the root v_0. Let $\delta^{P_i}(v_j)$ be the number of nodes on path P_i which are inside the interference range of v_j (including v_j itself). Assume the maximum interference number Δ_i on each path P_i is $\max\{\delta^{P_i}(v_j)\}$ for all $v_j \in P_i$. Hereafter, we call Δ_i *path interference* of path P_i. Then we can prove that T has a nice property that the path interference of each branch is bounded by a constant.

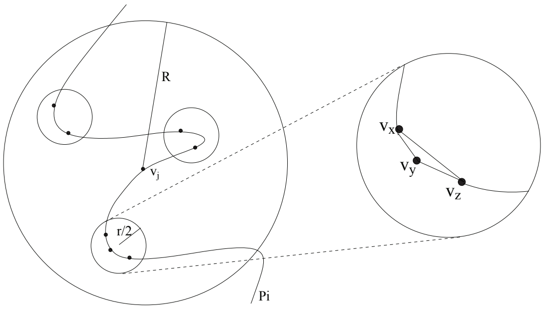


Fig. 1. Proof of Lemma 1: on a path P_i in BFS T, the interference nodes for a node v_j is bounded by a constant.

Lemma 1: Given a BFS tree T under the protocol interference model, the maximum interference number Δ_i on each path P_i is bounded by a constant $8\alpha^2$, i.e., $\Delta_i \leq 8\alpha^2$.

Proof: We prove by contradiction with a simple area argument. Assume that there is a v_j on P_i whose $\delta^{P_i}(v_j) > 8\alpha^2$. In other words, more than $8\alpha^2$ nodes on P_i are located in the interference region of v_j. Since the area of interference region is πR^2, we consider the number of interference nodes inside a small disk with radius $\frac{r}{2}$. See Figure 1 for illustration. The number of such small disks is at most $\frac{\pi R^2}{\pi(\frac{r}{2})^2} = 4\alpha^2$ inside πR^2. By the Pigeonhole principle, there must be more than $\frac{8\alpha^2}{4\alpha^2} = 2$ nodes inside a single small disk with radius $\frac{r}{2}$. In other words, three nodes v_x, v_y and v_z on the path P_i are connected to each other as shown in Figure 1. This is a contradiction with the construction of BFS tree. As shown in Figure 1, if v_x and v_z are connected in G, then v_z should be visited by v_x not v_y during the construction of BFS tree. This finishes our proof. ■

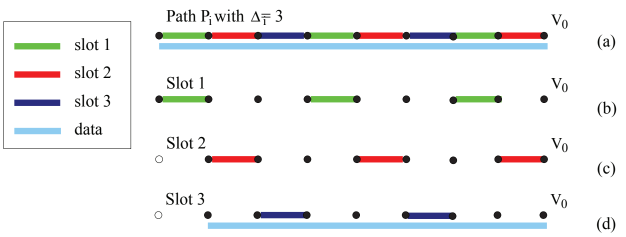
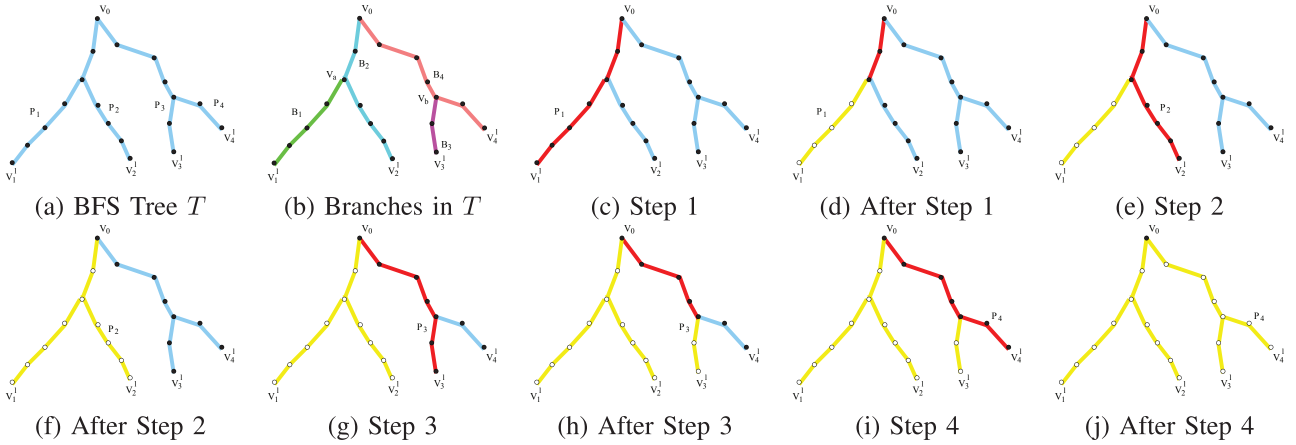


Fig. 2. Scheduling on a path: after Δ_i slots the sink gets one data.

B. Branch Scheduling Algorithm

We now illustrate how to collect one snapshot from all sensors. Given the collection tree T, our scheduling algorithm basically collects data from each path P_i in T one by one.

First, we explain how to schedule collection on a single path. For a given path P_i, we can use Δ_i slots to collect one data in the snapshot at the sink. See Figure 2 for illustration.



(a) BFS Tree T (b) Branches in T (c) Step 1 (d) After Step 1 (e) Step 2 (f) After Step 2 (g) Step 3 (h) After Step 3 (i) Step 4 (j) After Step 4

Fig. 3. Illustrations of our scheduling on the data collection tree T.

In this figure, we assume that $R = r$, i.e., only adjacent nodes interfere with each other. Thus $\Delta_i = 3$. Then we color the path using three colors as in Figure 2(a). Notice that each node on the path has unit data to transfer. Links with the same color are active in the same slot. After three slots (Figure 2(d)), the leaf node has no data in this snapshot and the sink got one data from its child. Therefore, to receive all data on the path, at most $\Delta_i \times |P_i|$ time slots are needed. We call this scheduling method *Path Scheduling*.

Now we describe our scheduling algorithm on the collection tree T. Remember T has c leaves which define c paths from P_1 to P_c. Our algorithm collects data from path P_1 to P_c in order. We define that i-th branch B_i is the part of P_i from v_i^l to the intersection node with P_{i+1} for $i = [1, c-1]$ and c-th branch $B_c = P_c$. For example, in Figure 3(b), there are four branches in T: B_1 is from v_1^l to v_a, B_2 is from v_2^l to v_0, B_3 is from v_3^l to v_b, and B_4 is from v_4^l to v_0. Notice that the union of all branches is the whole tree T. Algorithm 1 shows the detailed branch scheduling algorithm. Figure 3(c)-(j) give an example of scheduling on T. In the first step (Figure 3(c)), all nodes on P_1 participate in the collection using the scheduling method for a single path (every Δ_1 slots, sink v_0 receives one data). Such collection stops until there is no data in this snapshot on branch B_1, as shown in Figure 3(d). Then Step 2 collects data on path P_2. This procedure repeats until all data in this snapshot reaches v_0 (Figure 3(j)).

Algorithm 1 Branch Scheduling on BFS Tree

Input: BFS tree T.

```
1: for each snapshot do
2:     for $t = 1$ to $c$ do
3:         Collect data on path $P_i$. All nodes on $P_i$ transmit
           data towards the sink $v_0$ using Path Scheduling.
4:         The collection terminates when nodes on branch $B_i$
           do not have data for this snapshot. The total slots
           used are at most $\Delta_i \cdot |B_i|$, where $|B_i|$ is the hop
           length of $B_i$.
5:     end for
6: end for
```

C. Capacity Analysis

We now analyze the achievable capacity of our data collection method by counting how many time slots the sink needs to receive all data of one snapshot.

Theorem 2: The data collection method based on path-scheduling in BFS tree can achieve data collection capacity of $\Theta(W)$ at the sink.

Proof: In Algorithm 1, the sink collects data from all c paths in T. In each step (Lines 3-4), data are transferred on path P_i and it takes at most $\Delta_i \cdot |B_i|$ time slots. Recall that *Path Scheduling* needs at most $\Delta_i \cdot k$ time slots to collect k packets from path P_i. Therefore, the total number of time slots needed for Algorithm 1, denoted by τ, is at most $\sum_{i=1}^{c} \Delta_i \cdot |B_i|$. Since the union of all branches is the whole tree T, *i.e.*, $\sum_{i=1}^{c} |B_i| = n$. Thus, $\tau \leq \sum_{i=1}^{c} \Delta_i |B_i| \leq \sum_{i=1}^{c} \tilde{\Delta} |B_i| \leq \tilde{\Delta} n$. Here $\tilde{\Delta} = \max\{\Delta_1, \cdots, \Delta_c\}$. Then, the delay of data collection $D = \tau t \leq \tilde{\Delta} nt$. The capacity $C = \frac{nb}{D} \geq \frac{nb}{\tilde{\Delta} nt} = \frac{W}{\tilde{\Delta}}$. From Lemma 1, we know that $\tilde{\Delta}$ is bounded by a constant. Therefore, the data collection capacity is $\Theta(W)$. ■

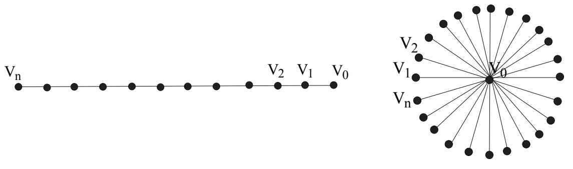
Remember that the upper bound of data collection capacity is W, thus our data collection algorithm is order-optimal. Consequently, we have the following theorem.

Theorem 3: Under protocol interference model and disk graph model, data collection capacity for arbitrary wireless sensor networks is $\Theta(W)$.

V. Collection Capacity for General Graph Model

So far, we assume that the communication graph is a disk graph where two nodes can communicate if and only if their distance is less than or equal to transmission range r. However, a disk graph model is idealistic since in practice two nearby nodes may be unable to communicate due to various reasons such as barriers and path fading. Therefore, in this section, we consider a more general graph model $G = (V, E)$ where V is the set of sensors and E is the set of possible communication links. Every sensor still has a fixed transmission range r such that the necessary condition for v_j to receive correctly the signal from v_i is $||v_i - v_j|| \leq r$. However, $||v_i - v_j|| \leq r$ is not the sufficient condition for an edge $v_i v_j \in E$. Some links do not belong to G because of physical barriers or the selection of routing protocols. Thus, G is a subgraph of a disk graph. Under this model, the network topology G can be any general graph (for example, setting $r = \infty$ and putting a barrier between any two nodes v_i and v_j if $v_iv_j \notin G$). Notice that even though we still consider the protocol interference model, our analysis still holds for arbitrary interference graph.

In general graph model, the capacity of data collection could be $\frac{W}{n}$ in the worst-case. We consider a simple straight-line network topology with n sensors as shown in Figure 4(a). Assume that the sink v_0 is located at the end of the network and the interference range is large enough to cover every node in the network. Since the transmission on one link will interfere with all the other nodes, the only possible scheduling is transferring data along the straight-line via all links. The total time slots needed are $n(n+1)/2$, thus the capacity is at most $\frac{nb}{n(n+1)t/2} = \Theta(\frac{W}{n})$. Notice that in this example, the maximum interference number Δ of graph G is n. It seems the upper bound of data collection capacity could be $\frac{W}{\Delta}$. We now show an example whose capacity can be much larger than $\frac{W}{\Delta}$. Again we assume all n nodes with the sink interfering with each other. The network topology is a star with the sink v_0 in center, as shown in Figure 4(b). Clearly, a scheduling that lets every node transfer data in order can lead to a capacity W which is much larger than $\frac{W}{\Delta} = \frac{W}{n}$.



(a) Straight-line Topology (b) Star Topology

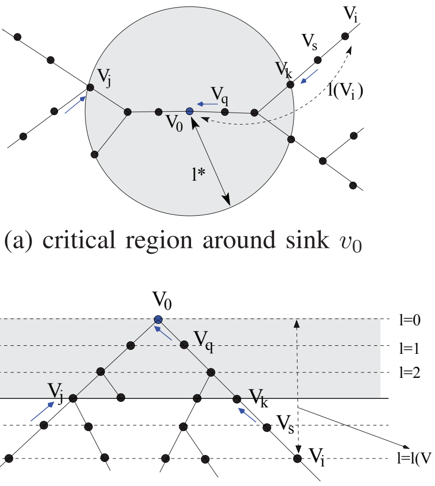
Fig. 4. The optimum of BFS-based method under two extreme cases.

A. Upper Bound of Collection Capacity

We first present a tighter upper bound of data collection capacity for general graph model than the natural one W. Consider all packets from one snapshot, we use p_i to represent the packet generated by sensor v_i. For any v_i, let $l(v_i)$ be its level in the BFS tree rooted at the sink v_0 (which is the minimum number of hops required for packet p_i or a packet at v_i to reach v_0). We use $D(v_0, l)$ to represent a virtual disk centered at the sink node v_0 with radius of hop distance l. The *critical level* (or called the *critical radius*) l^* is the greatest level l such that no two nodes within l level from v_0 can receive a message in the same time slot, *i.e.*, $l^* = \max\{l | \forall v_i, v_j \in D(v_0, l)$ cannot receive packets at the same time$\}$. The region defined by $D(v_0, l)$ is called *critical region*. See Figure 5 for illustration. For any packet p_i originated at node v_i, we define

$$\lambda_i^* = \begin{cases} l(v_i) & \text{if } v_i \in D(v_0, l^*) \\ l^* + 1 & \text{otherwise.} \end{cases}$$

Here, λ_i^* gives the minimum number of hops needed to reach the sink v_0 after packet p_i reaches the critical region around v_0. Let $\lambda^* = \max_i\{\lambda_i^*\}$. Then we can prove the following lemma on the lower bound of delay for data collection.



(a) critical region around sink v_0

(b) a tree view of critical region

Fig. 5. Illustration of the definition of critical region, *i.e.* l^*. The grey area is the critical region, where no any two nodes can receive a message in the same time slot due to interference around v_0.

Lemma 4: For all packets from one snapshot, the delay to collect them at sink v_0

$$D \geq t \sum_i \lambda_i^*.$$

Proof: It is clear the critical region around the sink v_0 is a bottleneck for the delay. Any packet inside the critical region can only move one step at each time slot. First, the total delay must be larger than the delay which is needed for the case where all packets originated outside critical region are just one hop away from the critical region. In other words, assume that we can move all packets originated outside critical region to the surrounding area without spending any time. Then each packet p_i needs λ_i^* time slots to reach the sink. By the definition of the critical region, no simultaneous transmissions around the critical region (1-hop from it) can be scheduled in the same slot. Therefore, the delay is at least the summation of λ_i^*. ■

Let $\Delta^* = \frac{\sum_i \lambda_i^*}{n}$, we have a new upper bound of data collection capacity, $C \leq \frac{W}{\Delta^*} \leq W$. Notice that $\Delta^* \geq 1$ and it represents the limit of scheduling due to interference around the sink (and its critical region).

B. Lower Bound of Collection Capacity

The data collection algorithm based on branch-scheduling in BFS tree can still achieve the capacity of $\frac{W}{\tilde{\Delta}}$. However, in general graph model we can not bound $\tilde{\Delta}$ by a constant any more, and it could be $O(1)$ or $O(n)$. Though this simple method can match the tight upper bounds $\Theta(\frac{W}{n})$ and W of examples shown in Figure 4, it is still not a tight bound. We show such an example and discuss a tighter lower bound based on this method in Section I of *Supplemental Material*.

Now we introduce a new greedy-based scheduling algorithm which is inspired by [19]. The scheduling algorithm still uses the BFS tree as the collection tree. All messages will be sent along the branch towards the sink v_0. For n messages from one snapshot, it works as follows. In every time slot, it sends each message along the BFS tree from the current node to its parent, without creating interference with any higher-priority message. The priority ρ_i of each packet p_i is defined as $\frac{1}{l(v_i)}$. It is clear that packets originated from the children of the sink have the highest priority $\rho_i = 1$ while packets originated from other nodes have lower priority $\rho_i < 1$. For two packets with the same priority (on the same level in the BFS tree), ties can be broken arbitrarily. Given a schedule, let v_j^τ be the node of packet p_j in the end of time slot τ. The detailed greedy algorithm is given in Algorithm 2.

Algorithm 2 Greedy Scheduling on BFS Tree

```
Input: BFS tree $T$.
Compute the priority $\rho_i = 1/l(v_i)$ of each message $p_i$.
for each snapshot do
    while $\exists p_j$ such that $v_j^\tau \neq v_0$ do
        for all such $p_i$ in decreasing order of priority $\rho_i$ do
            if sending $p_i$ from node $v_i^\tau$ will not create interference with any higher-priority messages that are already scheduled for this time slot then
                node $v_i^\tau$ sends $p_i$ to its parent $par(v_i^\tau)$ in $T$.
            end if
        end for
        $\tau = \tau + 1$.
    end while
end for
```

Now we analyze the capacity achieved by this greedy data collection method. Before presenting the analysis, we first introduce some new notations. For two nodes v_i and v_j, $h(v_i, v_j)$ denotes the shortest hop number from v_i to v_j in graph G. The delay of packet p_j is defined as the time until it reach the sink v_0, *i.e.*, $D_j = t \cdot \min\{\tau : v_j^\tau = v_0\}$.

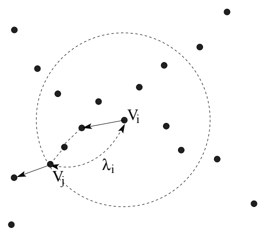


Fig. 6. Illustration of the definitions of λ_i.

Let λ_i be the minimal hops that a packet needs to be forwarded from node v_i before a new packet at v_i can be safely forwarded along the BFS tree. So $\lambda_i = \max\{l | \exists v_j, h(v_i, v_j) = l$ and transmission from v_i to $par(v_i)$ interferes with transmission from v_j to $par(v_j)\} + 1$. Here $par(v_i)$ is the parent of v_i in T. See Figure 6 for illustration. Here $\lambda_i = 4$ for v_i. We define that $\lambda = \max_i\{\lambda_i\}$. Both λ and λ_i are integers (hop counts). In addition, we can prove that $\lambda \geq \lambda^*$. A detailed proof is provided in Section II of *Supplemental Material*.

Packet p_j is said to be blocked in time slot τ if, in time slot τ, p_j is not sent out. We define the following blocking relation on our greedy algorithm schedule: $p_k \prec p_j$ if in the last time slot in which p_j is blocked by the transmission of higher priority packets in that time slot, p_k is the one closest to p_j in term of hops among these packets (ties broken arbitrarily). The blocking relation induces a directed blocking tree T_D where nodes are all message p_i and edge (p_k, p_j) representing $p_k \prec p_j$. The root p_r of the tree T_D is a message with highest priority (originated in a child of v_0) which is never blocked. Let $P(j)$ the path in T_D from p_r to p_j and $h(j)$ be the hop count of $P(j)$. We then derive an upper bound on the delay D_j of packet p_j in the greedy algorithm.

Lemma 5: For each packet p_j in the snapshot, its delay $D_j \leq t \cdot \sum_{p_i \in P(j)} \min\{l(v_i), \lambda\}$.

Proof: We prove this lemma by induction on $h(j)$. For any packet p_j, if $h(j) = 0$, which means p_j is the root p_r of T_D, it will not be blocked. So $D_j = t \cdot l(v_j)$. Then consider the right side of the inequation $t \cdot \sum_{p_i \in P(j)} \min\{l(v_i), \lambda\} = t \cdot \min\{l(v_j), \lambda\}$. Since p_j is packet with highest priority, $l(v_j) = 1$ and $l(v_j) \leq \lambda$. Thus, $t \cdot \sum_{p_i \in P(j)} \min\{l(v_i), \lambda\} = t \cdot l(v_j)$ and the claim in this lemma holds for the case where $h(j) = 0$.

If $h(j) > 0$, i.e., $p_j \neq p_r$, let τ be the last time slot in which p_j is blocked by packet p_k, *i.e.*, $p_k \prec p_j$. Notice that $t \cdot h(v_k^\tau, v_0) \leq D_k - t \cdot \tau$, otherwise p_k would not reach v_0 by time D_k. Also $h(v_j^t, v_k^t) \leq \lambda - 1$ since after p_k moves one hop p_j is safe to move. From time slot $\tau + 1$, p_j may be forwarded towards v_0 over one hop in each time slot, and reach v_0 at the earliest time slot,

$$\begin{aligned} D_j &\leq t \cdot (\tau + 1 + h(v_j^t, v_0)) \\ &\leq t \cdot (\tau + 1 + h(v_k^t, v_0) + h(v_j^t, v_k^t)) \\ &\leq t \cdot (\tau + 1) + D_k - t \cdot \tau + t \cdot \lambda - 1 \\ &= D_k + t \cdot \lambda. \end{aligned}$$

On the other hand, $D_j \leq D_k + t \cdot l(v_j)$ because after p_k reaches the sink v_0, p_j needs at most $l(v_j)$ to reach the sink. Consequently, $D_j \leq D_k + t \cdot \min\{l(v_j), \lambda\}$. This completes our proof. ■

Lemma 6: The data collection capacity of our greedy algorithm is at least $\frac{\lambda^*}{\lambda} \frac{W}{\Delta^*}$.

Proof: Let p_j be the packet having maximum D_j . By Lemma 5 and $\lambda \geq \lambda^*$,

$$\begin{aligned} D_j &\leq t \sum_{p_i \in P(j)} \min\{l(v_i), \lambda\} \leq \frac{\lambda}{\lambda^*} t \sum_{p_i \in T_D} \min\{l(v_i), \lambda^*\} \\ &\leq \frac{\lambda}{\lambda^*} t (\sum_{v_i \in D(v_0, l^*)} l(v_i) + \sum_{v_i \notin D(v_0, l^*)} (l^* + 1)) \\ &= \frac{\lambda}{\lambda^*} t \sum_i \lambda_i^* = \frac{\lambda}{\lambda^*} n t \Delta^*. \end{aligned}$$

Thus, the capacity achieved by our greedy algorithm is at least $\frac{nb}{D_j} = \frac{\lambda^*}{\lambda} \frac{W}{\Delta^*}$. ■

Remark: In summary, we show that under protocol interference model and general graph model data collection capacity for arbitrary sensor networks has the following bounds:

Theorem 7: Under protocol interference model and general graph model, data collection capacity for arbitrary sensor networks is at least $\frac{\lambda^*}{\lambda} \frac{W}{\Delta^*}$ and at most $\frac{W}{\Delta^*}$.

Here λ^* describes the interference around the sink v_0, while λ describes the interference around a node v_i. Since $\lambda \geq \lambda^*$, $\frac{\lambda^*}{\lambda} \leq 1$. For disk graph model, $\frac{\lambda^*}{\lambda}$ is a constant. However, for general graph model it may not, thus, there is still a gap between the lower and upper bounds (such an example is given in Section I of *Supplemental Material*). We leave finding tighter bounds to close the gap as one of our future works. For two examples in Figure 4, the greedy method matches the optimal solutions in order. For the straight-line topology in Figure 4(a), $\lambda^* = \lambda = n$ and $\Delta^* = \Theta(n)$. Thus, the capacity $\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*} = \Theta(\frac{W}{n})$ matches the upper bound. For the star topology in Figure 4(b), $\lambda^* = \lambda = 1$ and $\Delta^* = 1$. In this case, $\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*} = \Theta(W)$ also matches the upper bound. Compared with the branch scheduling method, greedy method can achieve much better capacity in practice, since greedy algorithm allows packet transmissions among multiple branches of the BFS tree in the same time slot. This is confirmed by our simulation results on random networks (Section V of *Supplemental Material*).

VI. Discussions on Other Models

A. Physical Interference Model

So far, we only consider the protocol interference model, which is an ideal and simple model. We can extend our analysis to the physical interference model by applying a technique introduced by Li *et al.* [8] when they studied the broadcast capacity of wireless networks. In *physical interference model*, node v_j can correctly receive signal from a sender v_i if and only if, given a constant $\eta > 0$, the SINR (Signal to Interference plus Noise Ratio)

$$\frac{P \cdot ||v_i - v_j||^{-\beta}}{B \cdot N_0 + \sum_{k \in I} P \cdot ||v_k - v_j||^{-\beta}} \geq \eta,$$

where B is the channel bandwidth, N_0 is the background Gaussian noise, I is the set of actively transmitting nodes when node v_i is transmitting, $\beta > 2$ is the pass loss exponent, and P is the fixed transmission power. We can prove the following theorem which indicates that data collection capacity under physical interference model is still $\Theta(W)$.

Theorem 8: Under physical interference model and disk graph model, data collection capacity for arbitrary wireless sensor networks is $\Theta(W)$.

Due to space limit, the detailed proof of this theorem is given in Section III of *Supplemental Material*.

B. Gaussian Channel Model

For both protocol interference model and physical interference model, as long as the value of a given conditional expression (such as transmission distance or SINR value) beyond some threshold, the transmitter can send data successfully to a receiver at a specific constant rate W due to the fixed rate channel model. While widely studied, fixed rate channel model may not capture well the feature of wireless communication. We now discuss the capacity bounds under a more realistic channel model: *Gaussian channel model*. In such model, it determines the rate under which the sender can send its data to the receiver reliably, based on a continuous function of the receiver's SINR. Again, we assume every node transmits at a constant power P. Any two nodes v_i and v_j can establish a direct communication link $v_i v_j$, over a channel of bandwidth W, of rate

$$W_{ij} = W \log_2 \left(1 + \frac{P \cdot ||v_i - v_j||^{-\beta}}{N_0 + \sum_{k \in I} P \cdot ||v_k - v_j||^{-\beta}}\right).$$

This model assigns a more realistic transmission rate at large distance than the fixed rate channel model with protocol or physical interference model.

In order to derive an upper bound for the capacity of data collection under Gaussian channel model, we consider the congestion at the sink node. In particular, we prove that whatever scheduling scheme is implemented, the total transmission rate of all the incoming links at the sink node is upper bounded by some value. As a bottleneck, the capacity of the whole network is always bounded by that value. Our proof basically follows the same idea proposed in [12] [13], which is firstly used to study the capacity bound for multicast session under Gaussian channel model. Due to space limit, the detailed proof is given in Section IV of *Supplemental Material*.

Theorem 9: An upper bound for data collection capacity under Gaussian channel model is at most

$$\max_i(W_{i0}) + W \cdot \log_2(n).$$

The first part of this upper bound depends on the rate of the shortest incoming link at sink, while the second part depends on the total number of nodes. Notice that $\max_i(W_{i0}) \leq W \log_2(1 + \frac{P}{N_0})$. Thus, which part in the bound playing an important role depends on the relationship between n and $1 + \frac{P}{N_0}$. When the network is a regular grid or a random homogeneous topology, it is satisfied that $l_{i0} \geq n^{\gamma}$ for some constant $\gamma < 0$. Then we have $\max_i(W_{i0}) = O(W \log n)$. Therefore, the total rate of all incoming links at sink node v_0 is at most $O(\log n \cdot W)$. A lower bound of data collection capacity in this model is still open.

VII. Conclusion

In this paper, we study the theoretical limits of data collection in terms of capacity for arbitrary wireless sensor networks. We first propose a simple data collection method based on BFS tree to achieve capacity of $\Theta(W)$, which is order-optimal under protocol interference model and disk graph model. However, when the underlying network is a general graph, we show that $\Theta(W)$ may not be achievable. We prove that a new BFS-based method using greedy scheduling can still achieve capacity of $\Theta(\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*})$ and also give a tighter upper bound $\Theta(\frac{W}{\Delta^*})$. At last, we discuss the collection capacity under more general models, physical interference model or Gaussian channel model. Table I summarizes our results. All of our methods can achieve these results for random networks too. We also provide some simulation results on random networks in Section V of *Supplemental Material*.

There are still several open problems left as our future work. First, we would like to close the gap of upper and lower bounds of data collection capacity for general graph; Second, the lower bound of data collection capacity under Gaussian channel model is still open. We plan to design new data collection schemes to approximate the upper bound better. Third, even though the capacity of data aggregation for arbitrary networks has been studied in [20], the author only considered the worst case capacity. It is interesting to study aggregation capacity for any arbitrary network. Fourth, different collection methods may cost different amount of energy. It is desired to study the trade-off between the achievable capacity and the energy consumption for data collection in sensor networks. Recent study [21] provides a nice start on this direction. Last, we also plan to study the collection capacity under more practical models (considering data correlation, fading effects, and time varying channels).

TABLE I

SUMMARY OF DATA COLLECTION CAPACITY

Network Model	Interference Model	Capacity C
Disk Graph	Protocol Interference	$C = \Theta(W)$
Disk Graph	Physical Interference	$C = \Theta(W)$
General Graph	Protocol Interference	$\Theta(\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*}) \leq C \leq \Theta(\frac{W}{\Delta^*})$
General Graph	Gaussian Channel	$C \leq \max_i(W_{i0}) + W \cdot \log_2(n)$

VIII. ACKNOWLEDGMENTS

The work of S. Chen, M. Huang and Y. Wang is supported in part by the US National Science Foundation (NSF) under Grant No. CNS-0721666, CNS-0915331, and CNS-1050398, and by Tsinghua National Laboratory for Information Science and Technology (TNList).

REFERENCES

[1] E.J. Duarte-Melo and M. Liu, "Data-gathering wireless sensor networks: Organization and capacity," *Computer Networks*, 43, 519–537, 2003.

[2] D. Marco, E.J. Duarte-Melo, M. Liu, and D.L. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proc. Int'l Workshop on Information Processing in Sensor Networks*, 2003.

[3] H.E. Gamal, "On the scaling laws of dense wireless sensor networks: the data gathering channel," *IEEE Trans. on Information Theory*, vol. 51, no. 3, pp. 1229–1234, 2005.

[4] R. Zheng and R.J. Barton, "Toward optimal data aggregation in random wireless sensor networks," in *Proc. of IEEE Infocom*, 2007.

[5] B. Liu, D. Towsley, and A. Swami, "Data gathering capacity of large scale multihop wireless networks," in *Proc. of IEEE MASS*, 2008.

[6] S. Chen, Y. Wang, X.-Y. Li, and X. Shi, "Capacity of data collection in randomly-deployed wireless sensor networks," *ACM Springer Wireless Networks (WINET)*, to appear, 2010. Short version in *Proc. of IEEE SECON*, 2009.

[7] S. Chen, S. Tang, M. Huang, and Y. Wang, "Capacity of data collection in arbitrary wireless sensor networks," in *Proc. of IEEE Infocom*, 2010.

[8] X.-Y. Li, J. Zhao, Y.W. Wu, S.J. Tang, X.H. Xu, and X.F. Mao, "Broadcast capacity for wireless ad hoc networks," in *Proc. of IEEE MASS*, 2008.

[9] P. Gupta and P.R. Kumar, "The capacity of wireless networks," *IEEE Trans. on Information Theory*, 46(2), 388-404, 2000.

[10] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Proc. of IEEE Infocom*, 2001.

[11] B. Liu, P. Thiran, and D. Towsley, "Capacity of a wireless ad hoc network with infrastructure," in *Proc. of ACM MobiHoc*, 2007.

[12] Franceschetti, M. and Dousse, O. and Tse, D.N.C. and Thiran, P. "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. on Information Theory*, 53(3), 1009-1018, 2007.

[13] Keshavarz-Haddad, A. and Riedi, R.H. "Bounds for the capacity of wireless multihop networks imposed by topology and demand," in *Proc. of ACM MobiHoc*, 2007.

[14] X.-Y. Li, S.-J. Tang, and O. Frieder, "Multicast capacity for large scale wireless ad hoc networks," in *Proc. of ACM MobiCom*, 2007.

[15] S. Shakkottai, X. Liu, and R. Srikant, "The multicast capacity of large multihop wireless networks," in *Proc. of ACM MobiHoc*, 2007

[16] A. Keshavarz-Haddad, V. Ribeiro, and R. Riedi, "Broadcast capacity in multihop wireless networks," in *Proc. of MobiCom*, 2006.

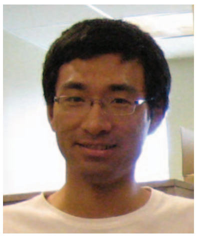
[17] B Tavli, "Broadcast capacity of wireless networks," *IEEE Communications Letters*, 10, 68-69, 2006.

[18] R.J. Barton and R. Zheng, "Order-optimal data aggregation in wireless sensor networks using cooperative time-reversal communication," in *Proc. of Annual Conf. on Information Sciences and Systems*, 2006.

[19] V. Bonifaci, P. Korteweg, A. Marchetti-Spaccamela, and L. Stougie, "An approximation algorithm for the wireless gathering problem," *Operations Research Letters*, 36, 605-608, 2008.

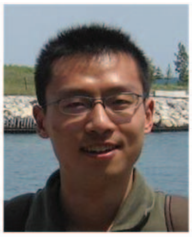
[20] T. Moscibroda, "The worst-case capacity of wireless sensor networks," in *Proc. of ACM IPSN*, 2007.

[21] X.-Y. Li, Y. Wang, and Y. Wang, "Complexity of data collection, aggregation, and selection for wireless sensor networks," *IEEE Trans. on Computers*, to appear, 2010.



Siyuan Chen received his B.S. degree from Peking University, China in 2006. He is currently a PhD student in the University of North Carolina at Charlotte, majoring in computer science. His current research focuses on wireless networks, ad hoc and sensor networks, and algorithm design.

Minsu Huang received his BS degree in computer science from Central South University in 2003 and his MS degree in computer science from Tsinghua University in 2006. He is currently a PhD student in the University of North Carolina at Charlotte, majoring in computer science. His current research focuses on wireless networks, ad hoc and sensor networks, and algorithm design.



Shaojie Tang has been a PhD student of Computer Science Department at the Illinois Institute of Technology since 2006. He received BS degree in Radio Engineering from Southeast University, China, in 2006. His current research interests include algorithm design and analysis for wireless ad hoc network and online social network.

Yu Wang is an Associate Professor of Computer Science at the University of North Carolina at Charlotte. He received his PhD degree (2004) in computer science from Illinois Institute of Technology, his BEng degree (1998) and MEng degree (2000) in computer science from Tsinghua University, China. His current research interests include wireless networks, ad hoc and sensor networks, and algorithm design. He is a recipient of Ralph E. Powe Junior Faculty Enhancement Awards from ORAU. He is a member of ACM and senior member of IEEE.