THROUGHPUT OPTIMIZATION FOR DATA COLLECTION IN WIRELESS SENSOR NETWORKS

by

Siyuan Chen

Approved by:

_____

Dr. Yu Wang

_____

Dr. Teresa Dahlberg

_____

Dr. Weichao Wang

_____

Dr. Jiang (Linda) Xie

_____

Dr. Wei Zhao

ABSTRACT

SIYUAN CHEN. Throughput optimization for data collection in wireless sensor networks.
(Under the direction of DR. YU WANG)

Wireless sensor networks are widely used in many application domains in recent years. 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 dissertation, we aim to study the theoretical limits of data collection in a TDMA-based sensor network in terms of possible and achievable maximum capacity. Various communication scenarios are considered in our analysis, such as with a single sink or multiple sinks, randomly-deployed or arbitrarily-deployed sensors, and different communication models. For both randomly-deployed and arbitrarily-deployed sensor networks, an efficient collection algorithm has been proposed under protocol interference model and physical interference model respectively. We can prove that its performance is within a constant factor of the optimal for both single sink and regularly-deployed multiple sinks cases. We also 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. In addition, we further discuss the problem of data collection capacity under Gaussian channel model.

ACKNOWLEDGMENTS

It is a pleasure to thank the many people who made this dissertation possible

First and foremost, I want to express my deeply-felt thanks to my advisor, Dr. Yu Wang. He patiently provide me with warm encouragement and thoughtful guidance to overcome tough situations and complete my dissertation. I appreciate all his contributions of time and ideas to make my Ph.D. experience productive and stimulating.

Special thanks to my committee, Dr. Teresa Dahlberg, Dr. Weichao Wang, Dr. Jiang (Linda) Xie and Dr. Wei Zhao for their support, guidance and helpful suggestions. I am grateful for their time and assistance over the years.

I sincerely acknowledge Prof. Qiansheng Cheng for offering me opportunities to experience research in Peking University and providing advice on my study. Without his help, I would not go to pursue my Ph.D. degree. I wish his soul roots in peace and solace in the heaven.

Lastly, I wish to thank my family for all their love and encouragement. My parents, Jianjun Chen and Qing Fang, deserve special mention for their inseparable support. They raised me with a love of science and supported me in all my pursuits. My grandfather, Xiyin Fang, receive my deepest gratitude for his love and persistent confidence in me. Words fail me to express my appreciation to my wife Fang Yan for being so patient and self-sacrificing, and for always giving me strength and hope. Without her love and faithful support, I would not have complete this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TDMA   Time Division Multiple Access

WSN   Wireless Sensor Network

ProIM   Protocol Interference Model

PhyIM   Physical Interference Model

ALERT   Automated Local Evaluation in Real-Time

MAC   Media Access Control

SINR   Signal to Interference plus Noise Ratio

CTR   Cooperative Time Reversal Communication

BFS   Breadth-First Search

CHAPTER 1: INTRODUCTION

A wireless sensor network (WSN)[2, 3, 4, 5] is a network consisting of a set of sensor nodes and is usually distributed randomly or strategically over a geographical area where traditional wired or wireless networks are difficult to be established or even unavailable. A sensor is a kind of computing device that is made up of sensing component, computing component, communication component, memory and power component and is able to perform data sensing, first stages of processing, and routing.

The main difference between sensor networks and traditional wired and wireless network is that a sensor is a very tiny and cheap device which has limited battery resource, limited memory storage and limited sensing range. Thus, a single sensor can only sense and collect some special events in its sensing range, which is significantly smaller than the target region area that needs to be monitored. To solve these weakness, a sensor must work in cooperation with other sensors to provide services by communicating with each other. As a result, a wireless sensor network often includes a large number of sensor nodes.

Due to its wide-range potential applications [6, 7, 8, 9, 10, 11, 12, 13] , such as wildlife tracking [10], volcanic activity monitoring [11, 12], sensor-based structure health monitor [13] and so on, wireless sensor network has recently emerged as a premier research topic. In these application scenarios, a sensing system usually requires that all of the wireless sensor nodes send their sensing data into a single sink node or multiple sink nodes in order to perform further analysis. For example, Automated Local Evaluation in Real-Time (ALERT) is a very famous system using sensor networks to transmit environmental data to a central computer in real time which was

developed by the National Weather Service [14, 5]. Sensors in the ALERT continuously sense local water level and temperature, and transmits data from the their location to the sink. In United States, ALERT has been widely used for flood alarming. Another well-known system of wireless sensor networks is the network deployed at Reventador volcano to monitor the volcano's activity [11], as show in Figure 1.1 [1]. Every sensor will relay the data to the sink (the red equipment)in the network when it detects an event. The sink will forward the data it has collected to base station for analyzing. From these examples above, data collection is one of the most common services used in wireless sensor network applications. Most of the sensor applications require the data's timeliness which means the data generated by sensors need to arrive at sinks as quickly as possible. And if the data can not be sent to sinks before its deadline, the data will probably be meaningless and be dropped. However, a wireless sensor network could cover a vast expanse and the scale of wireless sensor network may achieves tens of thousands of sensors. It is possible that the distance between the location of a sensor and the sink is very long. Information generated from such sensors may need numerous sensors' relays to reach the sink, and the delay of such information will be extremely large. Therefore, the largest delay of a given sensor network is a critical problem. Furthermore, most applications need sensors sense and generate data from their surrounding and forward it to sinks ceaselessly. Here, we call the data rate at the sink to continuously receive multiple snapshot from sensors in a long period as *capacity of data collection*. We are also interested in the capacity limit of a wireless sensor network under different interference models.

More formally, the *delay* of data collection is the time to transmit one single *snapshot* to the sinks from the snapshot generated at sensors. Considering the size of data in the snapshot, we can define *delay rate* as the ratio between the data size and the delay. Clearly, large delay rate is desired. When multiple snapshots from sensors are generated continuously, data transport can be pipelined in the sense that further
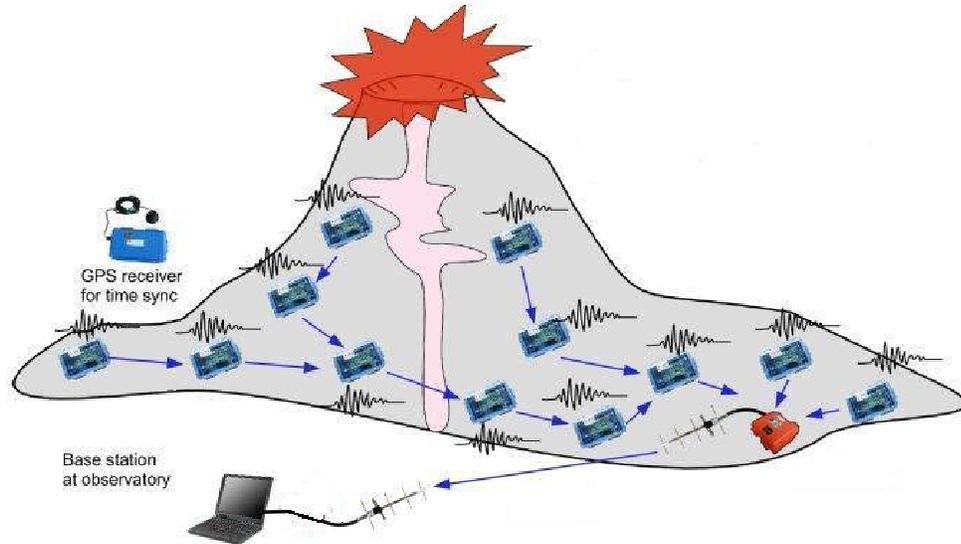
Figure 1.1: Data collection process in a WSN deployed at a volcano.[1]

snapshot may begin to be transported before the sink receiving the prior snapshot. The maximum data rate at the sinks to continuously receive the snapshot data from sensors is defined as the *capacity* of data collection. Notice that the capacity is always larger than or at least equal to the delay rate. Both *delay rate* and *capacity* reflect that how fast the sinks can collect data from all sensors. It is critical to understand the limitations of many-to-one information flows and devise efficient data collection algorithms to maximize performance of wireless sensor networks.

There already exists several literature about capacity problems in wireless networks [15, 16, 17, 18, 19, 20, 21, 22]. But they mainly focus on the capacity problem for unicast, multicast or broadcast in wireless ad hoc networks. Capacity limits of data collection in random wireless sensor networks have been studied in the literature [23, 24, 25, 26, 27, 28]. In [23, 24], Duarte-Melo *et al.* first studied the many-to-one transport capacity in dense and random sensor networks. But they only considered the simplest case with a single sink under the protocol interference model(ProIM). El Gamal [25] studied the capacity of data collection subject to a total average transmitting power constraint where a node can receive data from multiple source nodes at a

time. Recently, Barton and Zheng [26] also investigated the capacity of data collection under general physical layer models (e.g. cooperative time reversal communication model) where the data rate of an individual link is not fixed as a constant W but dependent on the transmitting powers and transmitting distances of all simultaneous transmissions. Both [25] and [26] assumed complex physical layer techniques, such as antenna sharing, channel coding and cooperative beam-forming. More related work could be found in Section 2.

This dissertation focuses on the fundamental capacity problems of data collection. We are interested in the theoretical upper bounds of the data collection capacity for a given sensor network, which answers the question that how fast the data generated by sensors can be sent to the sink theoretically. Such network capacity results are not only important in theory, but also provide practical guidelines for protocol design in real wireless sensor networks. We will also design efficient collection algorithms and scheduling mechanisms to achieve or approach to such upper bounds.

1.1   Network Model

We focus on the theoretical capacity boundary of data collection in wireless sensor networks. For simplicity, a simple and general model that is widely used in the community is introduced. We consider a static sensor network which includes n wireless sensor nodes $V = \{v_1, v_1, \cdots, v_n\}$ and a sink set S. Here, we assume that both sensor nodes and sink nodes are deployed in a plain area. At regular time intervals, each sensor node measures the field value at its position and transmits the value to one of the sinks. We first adopt a fixed data-rate channel model where each wireless node can transmit at W bits/second over a common wireless channel. Then we discuss the data collection capacity problem under a varying data-rate channel model-Gaussian Channel Model. We also assume that all packets have a unit size of b bits and each sensor has a fixed transmission power P. Time is partitioned into slots with $t = b/W$ seconds. Accordingly, only one packet can be transmitted in each
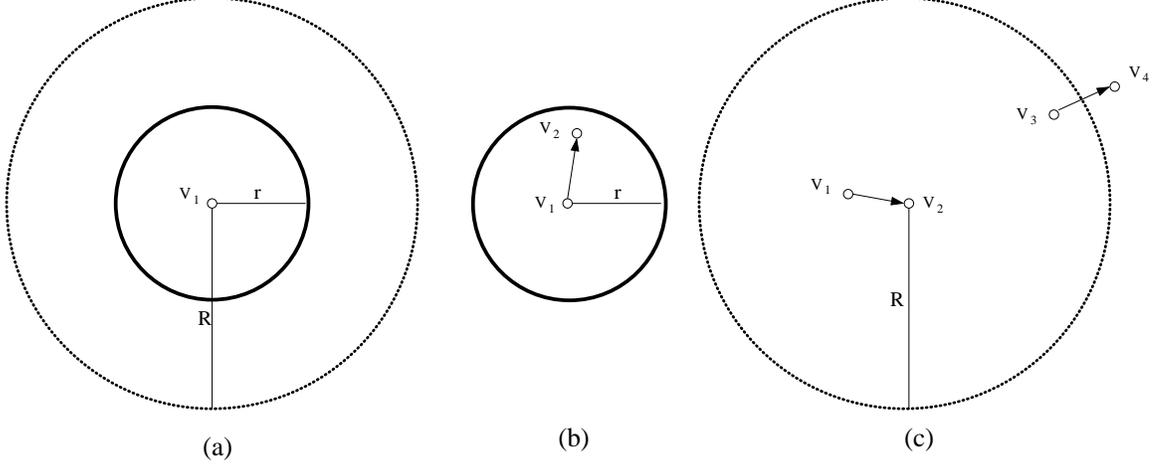
Figure 1.2: An example of protocol interference model.

time slot between two neighboring nodes. TDMA scheduling is used at MAC layer.

1.2 Communication Models

There are three different communication models that are widely used in wireless sensor networks: protocol interference model (ProIM) [15, 23, 24, 27], physical interference model (PhyIM) [15] and Gaussian channel model [29, 30].

Under *protocol interference model*, a fixed transmission range r and interference range R are defined such that a node v_j can successfully receive the signal sent by node v_i only if $||v_i - v_j|| \leq r$ and no node expect v_i within a distance R from v_j is transmitting simultaneously. Here, $||v_i - v_j||$ is the Euclidean distance between v_i and v_j. We show an example in Figure 1.2, in Figure 1.2(a), the circle of real line is the transmission range of node v_1 while the circle of dashed line means the interference range of v_1. Only the receiver v_2 is in the transmission range of its sender v_1, v_2 can get the data successfully as shown in Figure 1.2(b). And the Figure 1.2(c) illustrates that since the distance between v_2 and v_3 is less than interference R, v_2 can not receive the data from v_1 if v_3 is transmitting at the same time. In this dissertation, for simplicity, we assume that $\frac{R}{r}$ is a constant α which is larger than 1.

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, $N_0 > 0$ 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 assume that each node uses the same transmission power and the background noise N_0 is a fixed constant.

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) reaches some threshold, the sender can send data successfully to a receiver at a specific constant rate W due to the fixed rate channel model. However, fixed rate channel model may not capture well the feature of wireless communication. As a result, a more realistic channel model: *Gaussian channel model* is introduced. 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. 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

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

Where N_0 is the background Gaussian noise, I is the set of actively transmitting nodes when node v_i is transmitting, $l(v_i, v_j) = \min\{1, ||v_i - v_j||^{-\beta}\}$ is the pass loss between node v_i and v_j, and $\beta > 2$, and P is the fixed transmission power. This model assigns a more realistic transmission rate at large distance than the fixed rate channel model(protocol interference model and physical interference model).

1.3 Capacity and Delay

We formally define the delay and capacity of data collection in sensor networks. Each sensor measures independent field values at regular time intervals and sends these values to one sink node. The union of all sensing values from n sensors at

a particular time is called *snapshot*. The task of data collection is to deliver these snapshots to sinks.

Definition 1. The *delay* of data collection Δ is the time transpired between the time a snapshot is taken by the sensors and the time the sinks have all data of this snapshot.

Definition 2. The *delay rate* of data collection Γ is the ratio between the data size of one snapshot $n \cdot b$ and the delay Δ.

It is clear that we prefer smaller delay and larger delay rate so that the sink can get each snapshot more quickly.

On the other hand, the data transport can be pipelined in the sense that further snapshots may begin to transport before the sinks receive prior snapshots completely. Therefore, we need to define a new data rate of continuously data collection.

Definition 3. The *usage rate* of data collection U is the number of time slots needed at sinks between completely receiving one snapshot and completely receiving next snapshot at the sinks.

Thus, the time used by sinks to successfully receive a snapshot is $T = U \times t$. Notice that due to pipelining, T is always smaller than or equal to Δ. Clearly, small usage rate and T are desired.

Definition 4. The *capacity* of data collection C is the ratio between the size of data in one snapshot and the time to receive such a snap shot (i.e., $\frac{nb}{T}$) at the sinks.

Thus, the capacity C is the maximum data rate at the sinks to continuously receive the snapshot data from sensors. Clearly, C is at least as large as the delay rate Γ, and is usually substantially larger.

In this dissertation, we analyze the delay rate and capacity for data collection in both random and arbitrary wireless sensor networks under various communication models. Notice that in our definitions we require data from every sensor to reach the sink in the same rate, thus, the fairness among all sensors is guaranteed.

1.4   Dissertation Organization

The rest of this dissertation is organized as follows. We first review related work on capacity analysis of wireless networks in Section 2. We study the data collection and data aggregation capacity of random sensor networks under various communication models in Section 3. In Section 4, we then discuss the collection capacity in arbitrary sensor networks under different interference models. We conclude the dissertation in Section 5 by summarizing the completed work and some possible future work.

CHAPTER 2: RELATED WORK

In this chapter, we briefly review existing results on capacity analysis in wireless networks or wireless sensor networks.

2.1 Capacity Analysis in Wireless Networks

Gupta and Kumar initiated the research on capacity of wireless ad hoc networks by studying the fundamental capacity limits in the seminal paper [15] under both protocol interference model and physical interference model. They mainly focused on the capacity of unicast that every node in the network chose a destination and sent its packet. In the case of random networks, where n nodes were independently distributed over a unit area, they showed that both the upper bound and lower bound of capacity are $\Theta(\frac{W}{\sqrt{n \log n}})$, while in the case of physical interference model, the upper bound is $O(\sqrt{n})$ and lower bound is $\Omega(\frac{W}{\sqrt{n \log n}})$.

The authors also studied the unicast capacity in arbitrary networks in [15], where n nodes were arbitrarily located in a unit area and each node had an arbitrarily chosen destination. They showed the capacity could achieve $\Theta(\sqrt{n})$ under protocol interference model. And under physical interference model, the capacity of upper bound and low bound are $O(n^{1-\frac{1}{\alpha}})$ and $\Omega(\sqrt{n})$ respectively.

A number of following papers studied capacity under different communication scenarios in random wireless networks: unicast [16, 17] and multicast [18, 19, 20] Grossglauser *et al.* [16] examined the asymptotic unicast capacity in large wireless ad hoc networks and showed that the average long-term unicast throughput was kept constant even when the density of nodes increased by introducing mobility into the network model. In their algorithm, the source node didn't transmit packet to the

destination node until their distance was within $O(\frac{1}{\sqrt{n}})$, when the network was distributed in a unit area. Liu *et al.* [17] studied the unicast capacity of a wireless ad hoc network with infrastructure which had n ordinary nodes and b base stations. For the one dimensional network model, they showed that the gain in capacity is increasing significantly with the number of base stations as long as $b \log b \leq n$. However, in the two-dimensional model, a hybrid wireless network requires a large number of base stations $b = \Omega(\sqrt{n})$ to obtain a linear capacity increase.

The multicast means that in a wireless network with n nodes, each node randomly picks $k-1$ nodes as destination from the other $n-1$ nodes and sends its packet. Li *et al.* [18] studied asymptotic multicast capacity for the large-scale random wireless networks. They showed the total multicast capacity is $\Theta(\sqrt{\frac{n}{\log n}} \cdot \frac{W}{\sqrt{k}})$, when $k = O(\frac{n}{\log n})$ and $\Theta(W)$ when $k = \Omega(\frac{n}{\log n})$. Then [19] studied the multicast capacity for the hybrid random wireless networks which consisted n ordinary wireless nodes and m base stations, which derived analytical upper bounds and lower bounds on multicast capacity in the case of base stations were distributed regularly in a grid. Shakkottai *et al.* [20] studied the multicast capacity of random networks with the number of source nodes n^ϵ for some $\epsilon > 0$, and the number of receivers per source $n^{1-\epsilon}$. And they proposed a comb-based architecture for multicast routing which achieves the upper bound for capacity in an order sense.

Broadcast capacity in an arbitrary network was studied in [21, 22]. Both of the two papers showed that the broadcast capacity of a given network is $\Theta(W)$. And if every node in the network served as a source and broadcasted, the capacity per flow is $\Theta(W/n)$. To achieve such capacity, [21] presented a broadcast scheme based on Minimum Connected Dominating Set. For random networks, the same capacity bounds can be obtained.

2.2   Capacity of Data Collection in Wireless Sensor Networks

In this dissertation, we focus on the capacity of data collection which is an all-to-one communication scenario. Capacity of data collection in random wireless sensor networks was studied in [23, 24, 25, 26, 28, 27, 31, 32, 33, 34]. In [23, 24], Duarte-Melo *et al.* first studied the many-to-one transport capacity in random sensor networks under protocol model and gave the result of overall capacity of data collection as $\Theta(W)$. They also showed that compressing data are inefficient to improve the capacity when the density of the sensor network increases to infinity in [24]. El Gamal [25] studied data collection capacity subject to a total average transmitting power constraint. They relaxed the assumption that every node can only receive a packet 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 [26] also investigated data collection capacity under more complex physical models (non-cooperative SINR model and cooperative time reversal communication (CTR) model). They first demonstrated that $\Theta(\log nW)$ is optimal and achievable by using CTR for a regular grid network in [28], 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 [26]. Liu *et al.* [27] recently 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. Note that data collection is a special case for their problem when $s(n) = n$ and $d(n) = 1$. Most recently, Ji *et al.* [31, 32, 33, 34] also studied data collection methods in random wireless sensor networks under different communication models. In [31], an order-optimal continuous data collection method was proposed for single-radio multi-channel wireless sensor networks under proto-

col model and a pipeline scheduling algorithm for data collection was proposed for dual-radio multi-channel networks under protocol model. In [32], a cell-based path scheduling algorithm was proposed for data collection under physical model which can achieve $\Theta(W)$ of capacity, then a segment-based pipeline scheduling algorithm using compressive data gathering technique to further improve the collection capacity was presented. In [33], Ji and Cai investigated the achievable data collection capacity for asynchronous wireless sensor networks under generalized physical model by giving a scalable distributed data collection algorithm with $\Theta(W)$ of achievable capacity. In [34], Ji *et al.* considered data collection capacity under a probabilistic network model where the successful transmission over a link is a random variable related to the SNIR. They proposed two scheduling algorithms (a cell-based multipath algorithm and a zone-based pipeline algorithm) for such a model. All research above (including this paper itself) shares the standard assumption where large number of sensor nodes are randomly and uniformly distributed in a plane. Such assumption is useful for simplifying the analysis and deriving nice theoretical limitations, but may be invalid in some practical sensor applications.

For capacity of data collection with aggregation, there are also several studies. Giridhar and Kumar [35] investigated a general aggregation problem in random sensor network where a symmetric function of the sensor measurements is used for data aggregation. It was shown that for random planar network, the maximum rate for computing divisible functions (a subset of symmetric functions) is $\Theta(\frac{W}{\log n})$. In addition, using a technique called block-coding, they further showed that type-threshold functions can be computed at a rate of $\Theta(\frac{W}{\log \log n})$ in the physical model. Mosci-broda [36] further studied the aggregation capacity for arbitrarily deployed networks (named as *worst-case capacity*) under both protocol and physical models. He showed that the worst-case capacities of data aggregation are $\Theta(\frac{W}{n})$ under protocol model and $\Omega(\frac{W}{\log^2 n})$ under physical model respectively. Notice that the worst-case capacity

definition in his paper was in terms of the union of all arbitrary networks instead of any fixed arbitrary network.

Finally, there are also some results [37, 38, 39] on how to schedule data aggregation in sensor network such that the delay or latency is minimized. In [37], Huang *et al.* developed a algorithm which had the latency bound $23R + \Delta - 18$, where Δ is the maximum node degree and R is the network radius. In [38], the minimum data aggregation time problem was proved NP-hard and aggregation schedules of latency at most $(\Delta - 1)R$ was proposed. In [39], the authors propose a new tree based approximation algorithm guaranteed performance ratio $\frac{7\Delta}{log_2|S|} + c$, where S is the set of sensors containing data, Δ is the maximal node degree, and c is a constant.

CHAPTER 3: DATA COLLECTION FOR RANDOM WSNS

We first consider the simplest situation: data collection under protocol interference model in a random wireless sensor network. We will consider different communication models, and in most cases, we will give order-optimal data collection schemes.

3.1 Preliminaries

In this section, we first introduce the network model, the communication model and a partition method which will be used for collection methods and theoretical analysis.

3.1.1 Network Model and Communication Model

We focus on the capacity bound of data collection in random wireless sensor networks that n wireless sensor nodes $V = \{v_1, v_1, \cdots, v_n\}$ are randomly and uniformly deployed in a square of unit area, which is called *random dense network*[42]. While the k sink nodes $S = \{s_1, s_2, \cdots, s_k\}$ can be distributed either regularly or randomly. We adopt the *protocol interference model* in our analysis first, and study the data collection problem under physical interference model later. Recall that in protocol interference model [23, 24, 27], all nodes are assumed to have 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 is transmitting simultaneously. And $\frac{R}{r}$ is a constant α which is larger than 1.

3.1.2 A Grid-Partition Method

We introduce a classical grid-partition method which is essential for our data collection methods and theoretical analysis. As shown in Figure 3.1, the network (e.g., the unit square) is divided into a^2 micro cells of the size $d \times d$. Here $a = 1/d$.

sink node s_l in cell (p, q)

(0, a)   (a, a)

1

$\overline{d}$

(0, 0)   (a, 0)
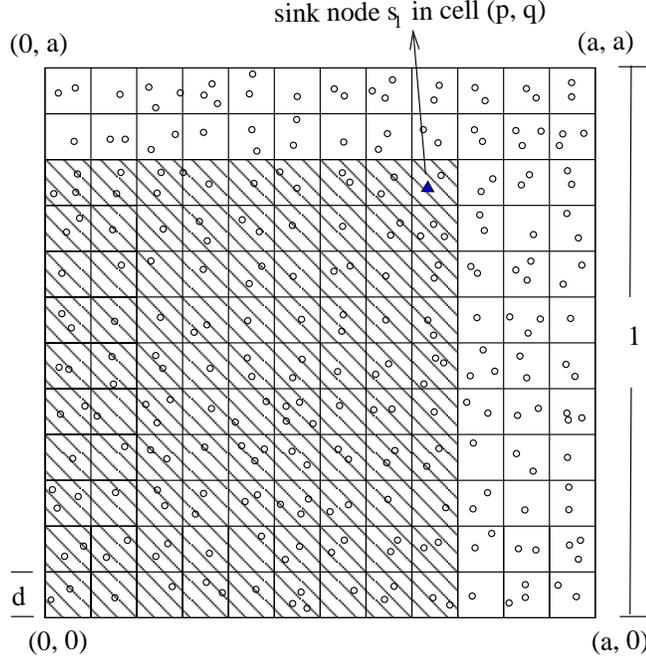
Figure 3.1: Grid partition of the WSN: a^2 cells with cell size of $d \times d$.

We assign each cell a coordinate (i, j), where i and j are between 1 and a, indicating its position at jth row and ith column.

The following lemma gives a guidance of the cell size.

Lemma 1. [43] Given n random nodes in a unit square, dividing the square into micro cells of the size $\sqrt{3\frac{\log n}{n}} \times \sqrt{3\frac{\log n}{n}}$, every micro cell is occupied with probability at least $1 - \frac{1}{n^2}$.

Therefore, if $d = \sqrt{3\frac{\log n}{n}}$ (i.e., $a = \sqrt{\frac{n}{3\log n}}$), every micro cell has at least one node with high probability (the probability converges to one as $n \longrightarrow \infty$).

In order to make the whole network connected, the transmission range r need to be equal or larger than $\sqrt{5}d$ so that any two nodes from two neighboring cells are inside each other's transmission range. Hereafter, we set $r = \sqrt{5}d = \sqrt{15\frac{\log n}{n}}$. In practice, the transmission range of a sensor device may be fixed. In such case, The above equation can still be hold by adjusting the deployment density (*i.e.*, n).

Then we can derive the upper bound of the number of nodes inside a single cell.

Lemma 2. Given n random nodes in a unit square, dividing the unit square into micro cells of the size $\sqrt{3\frac{\log n}{n}} \times \sqrt{3\frac{\log n}{n}}$, the maximum number of nodes in any cell is $O(\log n)$ with probability at least $1 - \frac{3\log n}{n}$.

Proof. The proof is straightforward from Lemma 3, thus we ignore the detail. Note that the number of balls $\gamma = n$ and the number of bins $\delta = a^2 = \frac{n}{3\log n}$. □

Lemma 3. [44, 45] Randomly putting γ balls into δ bins, with probability at least $1 - \frac{1}{\delta}$, the maximum number of balls in any bin is $O(\frac{\gamma}{\delta} + \log \delta)$.

Lemma 2 indicates the number of nodes inside any cell is bounded from above by $O(\log n)$ with high probability.

3.2   Data Collection with Single Sink

In this section, we consider the simplest situation: data collection under protocol interference model in a sensor network where a single sink s_1 located in cell (p, q) is used as the collector to collect all sensing data. We first construct a data collection scheme whose delay and delay rate are $O(nt)$ and $\Omega(W)$ respectively, and then prove that these values are order-optimal.

3.2.1   Data Collection without Aggregation

As shown in Figure 3.1, we consider the data collection of nodes from four different directions (i.e., quadrants) to s_1. For the purpose of analysis, we only concentrate on the direction which has the largest number of sensors, *e.g.*, the shaded rectangle in Figure 3.1, since the sink can perform collection on each direction in turn and it only adds a constant 4 in the analysis. Our collection algorithm has two phases. In the first phase (Phase I), every sensor sends its data up to the highest cell in its column (in the pth row) as shown in Figure 3.2 (a) and (b), and in the second phase (Phase II), all data is sent via cells in the pth row to the sink as shown in Figure 3.2 (c) and (d). We define the time needed for these two phases as T_1 and T_2, respectively.

By Lemma 2, the number of nodes in each cell is at most $O(\log n)$. Every node

(a) Phase I, 1st time slot

(b) Phase I, 2nd time slot

(c) Phase II, 1st time slot

(d) Phase II, 2nd time slot
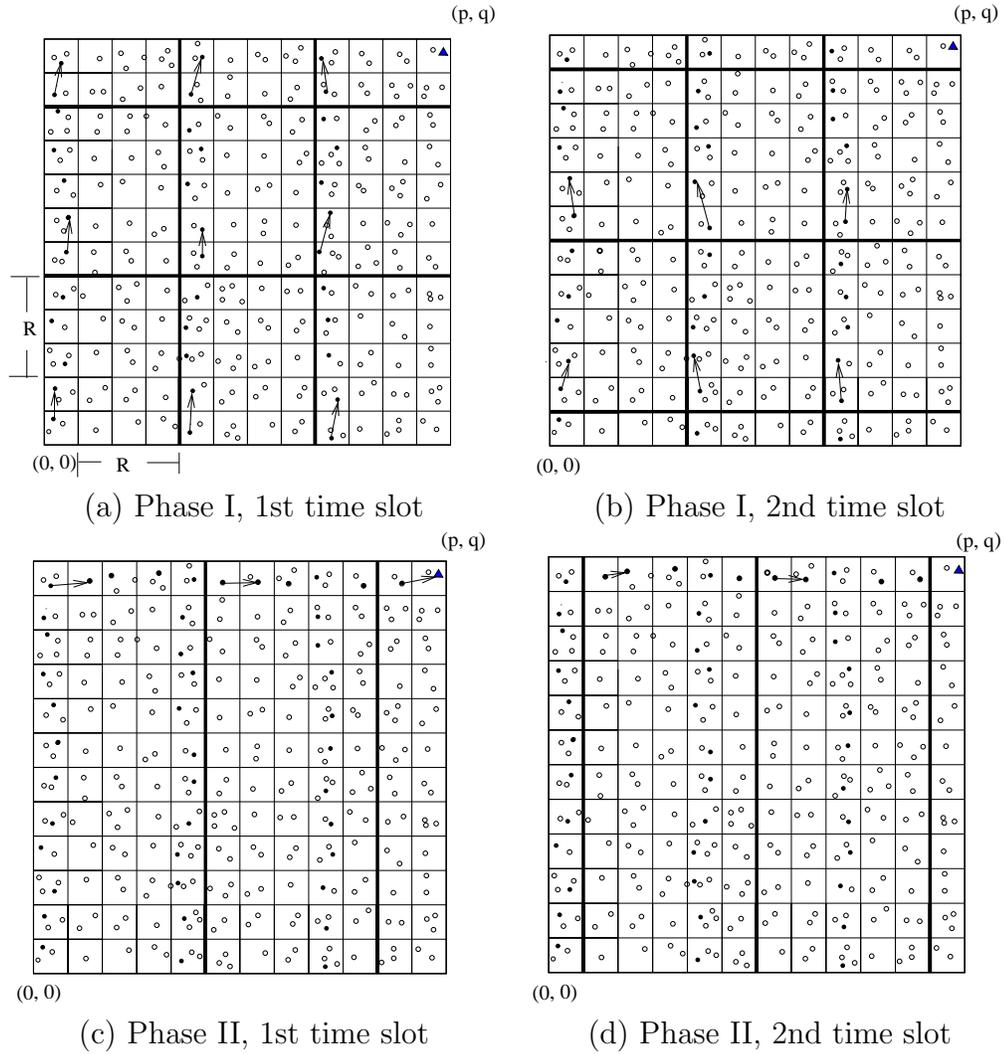
Figure 3.2: Our collection method: [Phase I] each node sends its data to its upper cell; [Phase II] each node in the top row sends its data to its right cell.

needs one time-slot t to send one packet to its neighbor in the next cell. However, due to wireless interference, when a node v_i transmits a packet to v_j, the nodes within R distance from v_j can not transmit any packets in the same time slot. Thus, every $(\frac{R}{d} + 2) \times (\frac{R}{d} + 1)$ cells (we call it an interference block hereafter) can only have one node send a packet to its upper neighbor in every time slot t during Phase I. In Figure 3.2, bold lines show interference blocks. Remember that $\frac{R}{r} = \alpha$ and $\frac{r}{d} = \sqrt{5}$, so $\frac{R}{d}$ is also a constant $\sqrt{5}\alpha$. And a packet in the lowest row (*i.e.* cell $(0, k)$) has to walk q cells to reach nodes in the highest cell in the rectangle. Hence,

$$
\begin{aligned}
T_1 &\leq (\frac{R}{d} + 2) \times (\frac{R}{d} + 1) \times t \times O(\log n) \times q \\
&= tO(\log n)q \leq O(t \log n)a \\
&= \sqrt{\frac{n}{3 \log n}} O(t \log n) = O(t\sqrt{n \log n}).
\end{aligned}
$$

In the beginning of Phase II, all data are already at cells of the top row. The sink s_1 lies in the same row with these cells. We now estimate the time T_2 needed for sending all data to s_1. Each cell in the top row has at most $qO(\log n)$ nodes' data and the interference block is $1 \times (\frac{R}{d} + 2)$ now. Similarly, we can get

$$
\begin{aligned}
T_2 &\leq (\frac{R}{d} + 2) \times t \times qO(\log n) \times p \\
&= O(t \log n)qp \leq a^2 O(t \log n) \\
&= \frac{n}{3 \log n} O(t \log n) = O(nt).
\end{aligned}
$$

Therefore, the total time needed to collect b-bits information from every sensor in the shaded rectangle to the sink is $T_1 + T_2 = O(nt)$. The other three directions need at most 3 times of such time. Thus, the total delay Δ_{col} for the sink to receive a complete snapshot is at most $O(nt)$. Consequently, the total delay rate of this

collection scheme is

$$\Gamma_{col} = \frac{nb}{\Delta_{col}} = \Omega(\frac{nb}{nt}) = \Omega(W).$$

It has been proved that the upper bound of delay rate or capacity of data collection is W [23, 24]. It is obvious that the sink cannot receive at a rate faster than W since W is the fixed transmission rate of individual link. Therefore, the delay rate of our collection scheme achieves the order of the upper bound, and the delay rate of data collection is $\Theta(W)$. Notice that even for individual sensors the lowest achievable delay rate of our method is $\Theta(W/n)$ which also meets the upper bound. In other words, our approach can achieve the order-optimal capacity for each individual sensor too.

Next, we consider the situation with pipelining. It is clear the upper bound of capacity is still W. Since our above scheme already reaches the upper bound, the pipelining operation can only improve the capacity within a constant factor.

With pipelining, in Phase I, the sensor can begin to transfer the data to its up-cell from next snapshot after sensors in its interference block finish their transmissions of previous snapshot. Whenever the cells in the top row receive $p \cdot b$ data (every cell in the top row receives a data from its lower cell), Phase II can begin at the top row. We consider the improvements of pipelining on both phases. With the pipelining, the time T_1' for the highest cell to receive a new set of $p \cdot b$ data in Phase I is

$$T_1' \leq (\frac{R}{d} + 2) \times (\frac{R}{d} + 1) \times t \times O(\log n) = O(t \log n).$$

And the time T_2' for the sink to receive a new set of $p \cdot b$ data in Phase II is

$$T_2' \leq \max\{nt, (\frac{R}{d} + 2) \times t \times p\} = O(nt).$$

Therefore, the total time for sink to receive $p \cdot b$ data is $T_1' + T_2' = O(nt)$. Thus, the

capacity of our method with pipelining is still

$$C_{col} = \frac{p \cdot b}{T_1' + T_2'} = \Omega(W).$$

This also meets the upper bound W in order.

In summary, we have the following theorem:

Theorem 1. Under protocol interference model, the delay rate Γ and the capacity C of data collection in random sensor networks with a single sink are both $\Theta(W)$.

3.2.2 Data Collection with Aggregation

In this section, we investigate a different data collection problem where each sensor can aggregate its received data (multiple packets) into a single packet. For example, if the sink just wants to know the maximal temperature in the deployed field, then each sensor can send out the maximal sensing value towards the sink instead of all values which it receives from other sensors. Hereafter, we will use this example as the running example of our analysis.

Here, we study both *delay rate* and *capacity* of data aggregation with a single sink. The definitions of delay rate and capacity are similar to those of data collection in Section 3.1. Notice that when the sink receives the maximal value (just b bits) of a snapshot of the field (n sensors), we still count the size of all values from that snapshot as the size of the received data. Thus, the delay rate is $\frac{nb}{\Delta}$ and the capacity is $\frac{nb}{T}$.

It seems that data aggregation is quite similar to broadcast, since both usually use a tree structure and the aggregation tree can be treated as a reversed broadcast tree. However, the capacity of data aggregation is completely different with the capacity of broadcast. Notice that in a broadcast tree a node can send its packet to all its children within one slot, while in an aggregation tree the children need multiple slots to send data to its parent due to interference among them. Therefore, the capacity

(a) Phase II   (b) Phase III
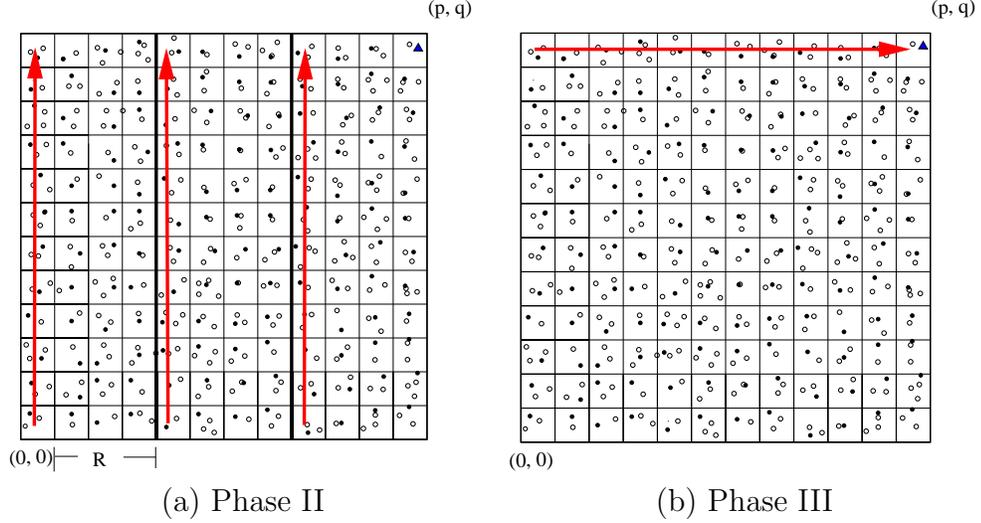
Figure 3.3: Our aggregation method: [Phase II] each selected node aggregates data to its upper cell; [Phase III] each selected node in the top row aggregates data to its right cell.

of aggregation is much smaller than the broadcast capacity $\Theta(nW)$ [21, 22, 46].

3.2.2.1 Delay Rate of Data Aggregation

We first consider the delay rate of data aggregation. We assume that a single sink s is located in cell (p, q) and only need to consider data aggregation from the direction which has the largest number of sensors. Our aggregation scheme has three phases and uses the same partition method as in Section 3.2.

First, each micro cell chooses a sensor which collects data from all the other sensors in the same micro cell and aggregates into one packet. Based on Lemma 3, each micro cell has at most $O(\log n)$ nodes. Assume that T_1'' is the time needed to collect data inside each cell. Because of the interference range R, T_1'' is at most

$$(\frac{R}{d} + 1)^2 \cdot O(\log n) \cdot t.$$

Second, every selected node waits for all data in the same snapshot from cells, which are below its own cell and within the same column, and then aggregates them with its value into a single packet and sends the packet to its upper cell. See Fig-

ure 3.3(a) for illustrations. At the end of this phase, all value has been aggregated at the top row where the sink sits. The time needed for this phase T_2'' is bounded from above by

$$(q-1) \times t \times (\frac{R}{d} + 1) = \Theta(\sqrt{\frac{n}{\log n}} t),$$

since for every $\frac{R}{d} + 1$ columns only one node can transmit due to interference, as shown in Figure 3.3(a).

Third, as shown in Figure 3.3(b), the information is aggregated via cells one by one in the top row. The time needed T_3'' is at most

$$(p-1) \times t = \Theta(\sqrt{\frac{n}{\log n}} t).$$

Therefore, the total delay $\Delta_{agg} \le T_1'' + T_2'' + T_3'' = O(\sqrt{\frac{n}{\log n}} t)$. The delay rate is

$$\Gamma_{agg} = \frac{nb}{\Delta_{agg}} = \Omega(\sqrt{n \log n} \cdot W).$$

Next, we prove that this delay rate is order optimal. Notice that for one snapshot data aggregation is completed when the sink has the aggregated value of all data in the snapshot. Let $T_{complete}$ denote the time that all data of one snapshot are aggregated in the sink and $T_{farthest}$ be the time needed for the value of the farthest node reach the sink. To compute the aggregated value, all values from the snapshot is needed. Therefore, $T_{farthest} \le T_{complete}$. Based on the network model, the farthest node from the sink is located in one corner of the field with high probability. We denote the distance between the farthest node and the sink as L. It is easy to show that the minimum value of L is $\frac{\sqrt{2}-d}{2}$ (when the sink is in the center of the field), i.e. $L \ge \frac{\sqrt{2}-d}{2}$. See Figure 3.4 for illustrations. Notice that when n goes to infinite, $L \ge \frac{\sqrt{2}}{2}$.

Since the transmission range is r, the data in the farthest node needs at least $\frac{L}{r}$

(a) $L > \frac{\sqrt{2}-d}{2}$ (b) $L = \frac{\sqrt{2}-d}{2}$
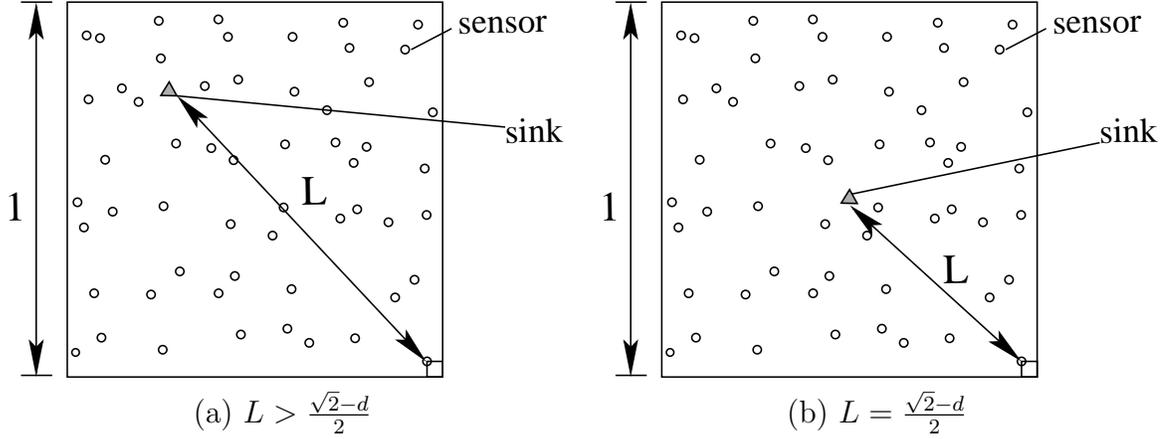
Figure 3.4: Minimum value of L is $\frac{\sqrt{2}-d}{2}$.

time slots to reach the sink. Hence,

$$T_{farthest} \geq \frac{L}{r} \cdot t = \frac{L}{r} \cdot \frac{b}{W} \geq \frac{\frac{\sqrt{2}}{2}}{r} \cdot \frac{b}{W} = \sqrt{\frac{n}{30 \log n}} \cdot \frac{b}{W}.$$

Consequently, we have

$$T_{complete} \geq T_{farthest} \geq \sqrt{\frac{n}{30 \log n}} \cdot \frac{b}{W}$$

Therefore, the delay rate of data aggregation is at most

$$\frac{nb}{T_{complete}} \leq \frac{nb}{\sqrt{\frac{n}{30 \log n}} \cdot \frac{b}{W}} = \Theta(\sqrt{n \log n} \cdot W).$$

Therefore, our data aggregation algorithm can achieve the upper bound of delay rate $\Theta(\sqrt{n \log n} \cdot W)$.

3.2.2.2 Capacity with Pipelining

We now describe our aggregation algorithm with pipelining. In the above algorithm, sensors will not start sending data in the next snapshot until the sink receives the aggregated value for all data in the previous snapshot. However, with pipelining,

a sensor can begin to send (or aggregate) data in the next snapshot before the aggregated value of the previous snapshot reaches the sink. Actually, it can begin to send if the aggregated data of the previous snapshot are far away enough. Thus, all three phases in the above algorithm can be pipelined.

At the beginning of each snapshot, each micro cell will choose a node to collect data from all the other nodes in the same micro cell and aggregates into one packet. The time required is $(\frac{R}{d} + 1)^2 \cdot O(\log n) \cdot t = O(t \log n)$.

For Phase II and Phase III if the aggregated values in previous snapshot are one interference block ahead (above or right in Figure 3.3), the values from next snapshot can be sent or aggregated. The time difference between such two snapshots will be bounded by $(\frac{R}{d} + 1)^2 \cdot t$.[1] This is much smaller than the time used for the aggregation of data in a cell ($O(t \log n)$). Thus, in a cell, when the aggregation of data from one snapshot finishes, the aggregated values of previous snapshot are already far away from this cell and can not cause any interference with current transmissions originated from this cell.

Therefore, at most every $\Theta(t \log n)$ the sink can collect one snapshot data with pipelining. Then the capacity of our data aggregation method is $\frac{nb}{\Theta(t \log n)} = \Theta(\frac{n}{\log n} W)$.

Next, we prove that the upper bound of data aggregation with pipelining is $O(\frac{n}{\log n} W)$. In other words, our schemes achieves the optimal order.

Consider n sensors are randomly distributed in the unit square. If we divide the region into disks with radius $\frac{R}{2} = \alpha \sqrt{\frac{15 \log n}{4n}}$, every such disk has average $\frac{15 \pi \alpha^2 \log n}{4}$ sensors. Due to Pigeonhole principle, there exists some disks that have $\Theta(\log n)$ sensors. Now let D be such a disk. When one sensor in D sends its data packet to a destination, all of the other $\Theta(\log n)$ sensors cannot send their data. The aggregation of these $\Theta(\log n)$ sensors will cost at least $\Theta(\log nt)$, i.e., $T_{agg} \geq \Theta(\log nt)$. Thus, the capacity C_{agg} is less than or equal to $O(\frac{n}{\log n} W)$ for sure.

---

[1] We can also think of this as the case where each cell has a single sensor. Then the rate of receiving data at the sink is a constant dependent on R.

In summary, we have the last theorem as follows.

Theorem 2. Under protocol interference model, the delay rate Γ and the capacity C of data aggregation in random sensor networks with a single sink are $\Theta(\sqrt{n \log n} W)$ and $\Theta(\frac{n}{\log n} W)$ respectively.

Notice that for data collection the delay rate and the capacity are in the same order (Theorem 1), i.e., pipelining can improve only a constant factor of the data rate. However, for data aggregation, it is very interesting to see that pipelining can increase the data rate in order of $\Theta(\sqrt{\frac{n}{\log^3 n}})$. This is because there is room for capacity improvement with data aggregation. Notice that the throughput in the non-aggregation case and aggregation case are limited differently.

3.3   Data Collection with Multiple Sinks

Now we consider networks with multiple sinks (e.g., k sinks). With more sinks, the collection task can be divided into small sub-tasks (i.e., collections in sub-areas) and each sub-task will be assigned to a single sink. Multiple sinks can collect data from their areas simultaneously if they are not interfering with each other. This can increase the capacity and decrease the delay of data collection. We will derive the bounds of data collection for multiple sinks using the results in the case with a single sink (Section 3.2). Since in both cases the delay rate and the capacity are always in the same order, here we will not distinct them and only use the term of capacity. Two scenarios are studied: sinks are regularly deployed on a grid or are randomly deployed in the field.

3.3.1   Multiple Sinks on Grid

When sinks are displayed regularly on a $\sqrt{k} \times \sqrt{k}$ grid, the capacity of collection depends on the number of sinks k. Here, we divide the unit area into k sub-areas which are $\frac{1}{\sqrt{k}} \times \frac{1}{\sqrt{k}}$ squares. There are two cases: $k < \frac{n}{15(\alpha+1)^2 \log n}$ or $k \geq \frac{n}{15(\alpha+1)^2 \log n}$.

Case 1: When $k < \frac{n}{15(\alpha+1)^2 \log n}$, $k < \frac{1}{(R+r)^2}$ since $R = \alpha r$ and $r = \sqrt{15 \frac{\log n}{n}}$. Thus, the area of each sub-area assigned to a sink is larger than or equal to $(R+r)^2$.

Therefore, we can perform the data collection in each sub-area without interfering with neighboring sub-areas. Since we have k sub-areas, the total delay rate and the total capacity of the whole area are at most $k \cdot \Theta(W) = \Theta(kW)$.

Case 2: When $k \geq \frac{n}{15(\alpha+1)^2 \log n}$, $k \geq \frac{1}{(R+r)^2}$. Thus the area of each sub-area is smaller than $(R+r)^2$, which indicates that there will be interference between neighboring sub-areas. Therefore, the total delay rate or capacity is bounded by $\frac{1}{(R+r)^2} \cdot \Theta(W) = \Theta(\frac{n}{\log n}W)$ from above, due to interference.

To achieve these upper bounds, the collection method for a single sink case can be used. When $k < \frac{n}{15(\alpha+1)^2 \log n}$, we partition the field into k sub-areas with size of $\frac{1}{\sqrt{k}} \times \frac{1}{\sqrt{k}}$ and every sink performs the collection method to collect their sub-areas. When $k \geq \frac{n}{15(\alpha+1)^2 \log n}$, we partition the field into $\frac{1}{(R+r)^2}$ sub-areas with size of $(R+r) \times (R+r)$ as shown in Figure 3.5. Then $\frac{1}{(R+r)^2}$ sinks can be selected to perform the collection method. Note that one selected sink may still cause interference with other selected sink in an adjacent block. However, the number of such adjacent selected sinks is bounded by eight. Thus, a simple scheduling can avoid the interference and the capacity of data collection is still in order of the theoretical bound. Figure 3.5 shows a possible scheduling where only one of nine selected sinks collects data from its surrounding blocks.

Therefore, we have our second theorem.

Theorem 3. Under protocol interference model, the delay rate Γ and the capacity C of data collection in random sensor networks with k regularly-deployed sinks are

$$
\begin{cases}
\Theta(kW), & \text{when } k < \frac{n}{15(\alpha+1)^2 \log n} \\
\Theta(\frac{n}{\log n}W), & \text{when } k \geq \frac{n}{15(\alpha+1)^2 \log n}.
\end{cases}
$$

Since when $k = \Theta(\frac{n}{\log n})$, the capacity (or delay rate) of two cases are all equal to
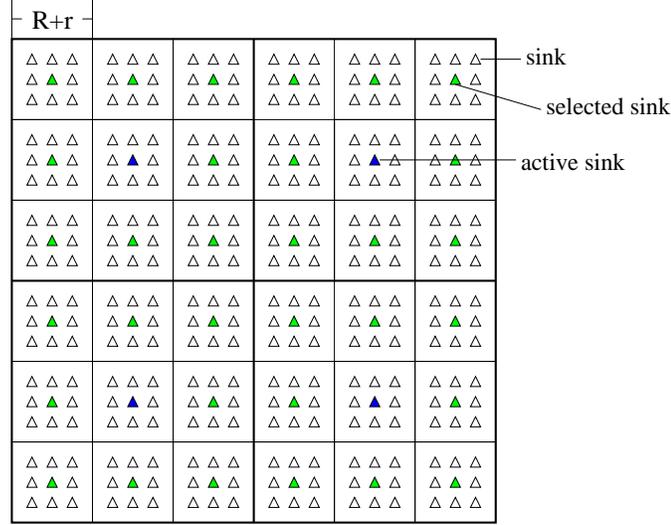
Figure 3.5: When k is large, we partition the field into $\frac{1}{(R+r)^2}$ sub-areas. Each subarea selects one sink as its selected sink (shown as a green triangle). Only one selected sink inside nine subareas is active for data collection, shown as a blue triangle. It will collect data from the surrounding 9 subareas using the method for single sink case. Notice that the adjacent 9-subareas will not interfere with each other when applying the collection method.

$\Theta(kW) = \Theta(\frac{n}{\log n}W)$. Therefore, the above equations can also be written as follows:

$$
\begin{cases}
\Theta(kW), & \text{when } k = O(\frac{n}{\log n}) \\
\Theta(\frac{n}{\log n}W), & \text{when } k = \Omega(\frac{n}{\log n}).
\end{cases}
$$

3.3.2 Randomly Deployed Multiple Sinks

Consider the scenario when k sinks are randomly distributed in the network. It is clear that if k is very large, the capacity is still bounded by the interference area. However, when the k is very small, the achievable capacity of collection may not reach the upper bound of $\Theta(kW)$ since the distribution of k sinks could be unbalanced in the field. In that case, even though the two neighboring sinks may not interfere with each other, they cannot fully operate over the whole period since some of them may finish their collection earlier and have no data to collect.

We first derive the upper bound of data collection capacity. Since the interference

range $R = \alpha r = \alpha \cdot \sqrt{15 \frac{\log n}{n}}$, we partition the whole area into interference blocks with size of $(R + r) \times (R + r)$. Thus, there are $B = \frac{n}{15(1+\alpha)^2 \log n}$ interference blocks. We then consider three cases when we randomly put k sinks into B interference blocks:

Case 1: When $k = o(\frac{n}{\log n})$. For this case, the capacity of data collection is bounded by $\Theta(kW)$ from above since the collection rate of each sink is bounded by W. Notice that data collection with a single sink is a special case when $k = 1$.

Case 2: When $k = \Theta(\frac{n}{\log n})$. We calculate the probability that an arbitrary interference block has at least one sink.

$$Pr(\text{an interference block has at least 1 sink})$$
$$= 1 - (1 - \frac{1}{B})^k = 1 - (1 - \frac{1}{\Theta(\frac{n}{\log n})})^k$$
$$= 1 - (1 - \frac{1}{\Theta(\frac{n}{\log n})})^{\Theta(\frac{n}{\log n})}$$

When $n \longrightarrow \infty$, this probability equals to $1 - \frac{1}{e}$. Let Pr be this probability. Then we define the number of interference blocks occupied by at least one sink as a random variable X. The expectation and variance of X are $E[X] = Pr \times B = (1 - \frac{1}{e})\frac{n}{60\alpha^2 \log n}$ and $\sigma^2 = Pr \times (1 - Pr) \times B = \frac{1}{e}(1 - \frac{1}{e})\frac{n}{60\alpha^2 \log n}$. Based on Chebyshev inequality, we have the following:

$$Pr(|X - E[X]| \geq \zeta\sigma) \leq \frac{1}{\zeta^2}.$$

Let $\zeta = \frac{1}{2} \cdot \sqrt{\frac{(1-\frac{1}{e})\frac{n}{60\alpha^2 \log n}}{\frac{1}{e}}}$, we have

$$Pr(|X - E[X]| \geq \frac{1}{2}E[X]) \leq \frac{4 \cdot \frac{1}{e}}{(1 - \frac{1}{e})\frac{n}{60\alpha^2 \log n}}$$

which goes to 0 when $n \longrightarrow \infty$. That means $\frac{1}{2}E[X] \leq X \leq \frac{3}{2}E[X]$ with high probability. In other words, the number of occupied interference blocks is $\Theta(\frac{n}{\log n})$. Therefore, the capacity of data collection is bounded by $\Theta(\frac{n}{\log n}W)$ (which is also

$\Theta(kW))$.

Case 3: When $k = \omega(\frac{n}{\log n})$. We also consider the probability that an arbitrary interference block has at least one sink.

$$Pr(\text{an interference block has at least 1 sink})$$

$$= 1 - (1 - \frac{1}{\Theta(\frac{n}{\log n})})^k$$

$$= 1 - (1 - \frac{1}{\Theta(\frac{n}{\log n})})^{\Theta(\frac{n}{\log n}) \cdot \frac{k}{\Theta(\frac{n}{\log n})}}$$

$$= 1 - (1 - \frac{1}{\Theta(\frac{n}{\log n})})^{\Theta(\frac{n}{\log n}) \cdot \frac{\Omega(\frac{n}{\log n})}{\Theta(\frac{n}{\log n})}}.$$

When $n \longrightarrow \infty$, this probability goes to 1. In other words, every interference block has at least one sink with high probability. Thus, we can select only one sink in each block to collect data at the same time. Then the capacity of data collection is bounded by $\Theta(\frac{n}{\log n}W)$ from above.

From the above analysis, we find that the capacity upper bounds for randomly distributed case are the same with the ones for regularly distributed case. Next, we present lower bounds of data collection capacity by giving our data collection methods.

When $k = O(\frac{n}{\log n})$, we first partition the network into interference blocks with size $\sqrt{3\frac{\log k}{k}} \times \sqrt{3\frac{\log k}{k}}$. From Lemma 1, we know that each of the block is occupied by at least one sink with high probability. Since $k = O(\frac{n}{\log n})$, the size of a block is $\sqrt{3\frac{\log k}{k}} > R + r$. Thus, we select one sink for each block, and use the same technique for grid-deployed sinks (Section 3.3.1) to schedule a subset of selected sinks to collect data from their surrounding area. The achieved capacity is $\Theta(\frac{k}{\log k}W)$ since the number of selected sinks are $\Theta(\frac{k}{\log k})$. Notice that there is a gap between this lower bound and the upper bound $\Theta(kW)$. This is due to possible uneven distribution of k sinks in this case, thus each sink may not have the same amount sensors (or areas)

to perform collection in order to achieve $\Theta(kW)$ capacity in total.

When $k = \omega(\frac{n}{\log n})$, we first partition the network into interference blocks with size $(R + r) \times (R + r)$. As show in Case 3, with high probability, each block has at least one sink. Using the same collection method, the achievable capacity is $\Theta(\frac{n}{\log n}W)$, which meats the upper bound perfectly.

Theorem 4. Under protocol interference model, the delay rate Γ and the capacity C of data collection in random sensor networks with k randomly-deployed sinks are

$$
\begin{cases}
\Theta(\frac{k}{\log k}W) \leq C \leq \Theta(kW), & \text{when } k = O(\frac{n}{\log n}) \\
C = \Theta(\frac{n}{\log n}W), & \text{when } k = \omega(\frac{n}{\log n}).
\end{cases}
$$

In summary, with multiple sinks (either grid or random deployment of k sinks), the capacity of data collection increases from that of the single sink case. When the capacity is constrained by the number of sinks (*i.e.*, $k = O(\frac{n}{\log n})$), it is beneficial to add more sinks. However, when the capacity is constrained by the interference among sinks (*i.e.*, $k = \omega(\frac{n}{\log n})$), adding more sinks has no substantial capacity improvement. Similar observations have been obtained in [27] for many-to-many capacity.

3.3.3   Data Collection with Aggregation

We have already studied data collection capacity without aggregation, however, results for the case with aggregation are easy to derive using the similar analysis in Section 3.3. Here we just present the conclusion. The delay rate of data aggregation in random sensor networks with k regularly-deployed sinks are

$$
\begin{cases}
\Gamma = \Theta(k\sqrt{n \log n}W), & \text{when } k = O(\frac{n}{\log n}) \\
\Gamma = \Theta(\frac{n\sqrt{n}}{\sqrt{\log n}}W), & \text{when } k = \Omega(\frac{n}{\log n}).
\end{cases}
$$

The capacity of data aggregation of this case are

$$
\begin{cases}
C = \Theta(\frac{kn}{\log n}W), & \text{when } k = O(\frac{n}{\log n}) \\
C = \Theta((\frac{n}{\log n})^2 W), & \text{when } k = \Omega(\frac{n}{\log n}).
\end{cases}
$$

The delay rate of data aggregation in random sensor networks with k randomly-deployed sinks are

$$
\begin{cases}
\Theta(\frac{k\sqrt{n \log n}}{\log k}W) \leq \Gamma \leq \Theta(k\sqrt{n \log n}W), & \text{when } k = O(\frac{n}{\log n}) \\
\Gamma = \Theta(\frac{n\sqrt{n}}{\sqrt{\log n}}W), & \text{when } k = \omega(\frac{n}{\log n}).
\end{cases}
$$

The capacity of data aggregation of this case are

$$
\begin{cases}
\Theta(\frac{kn}{\log k \log n}W) \leq C \leq \Theta(\frac{kn}{\log n}W), & \text{when } k = O(\frac{n}{\log n}) \\
C = \Theta((\frac{n}{\log n})^2 W), & \text{when } k = \omega(\frac{n}{\log n}).
\end{cases}
$$

3.4 Data Collection under Physical Interference Model

So far, we only consider the protocol interference model, which is an ideal but unrealistic in wireless sensor networks, where the interference is modeled as a localized phenomenon. However, a receiver can be interfered by a group of active transmitting sensors even its location is extremely far away from the group sensors. For example, when the scale of network is large and many sensors are sending signals at the same time, several concurrent transmissions may interfere with another sensor which has a long distance from all these active sensors. Thus, we need more accurate models to reflect the influence of interference. In this section, we will extend our analysis to a more general communication model: *physical interference model*, where SINR is considered to capture the physical interference constraints in real communication environment. In physical interference model, node v_j can correctly receive the signal

from the sender v_i if and only if, given a constant $\eta > 0$, the SINR

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

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. And we assume that each sensor uses the same transmission power P, and all N_0, β and η are fixed constants. Notice that values of P, N_0, η, and transmission range r should satisfy that $\frac{P \cdot r^{-\beta}}{B N_0} \geq \eta$. Thus, $r \leq (\frac{P}{B \cdot N_0 \cdot \eta})^{1/\beta}$.

We again consider a sensor network with n sensor nodes $V = \{v_1, v_1, \cdots, v_n\}$ and a single sink s. Here, we assume that both sensor nodes and the sink node are uniformly deployed in a square region with side-length $a = \sqrt{n}$, by use of Poisson distribution with density 1 instead of n. This network model is called *random extended network*[42] which is different from the random dense network we used in previous sections. We still use the fixed data-rate channel model where each wireless node can transmit at W bits/second over a common wireless channel.

3.4.1 Data Collection without Aggregation

We first consider data collection without aggregation by constructing a data collection scheme whose delay rate is order-optimal. Our data collection scheme is again based on a grid partition method which is very similar to the one used in Section 3.1.2.

3.4.1.1 Partition Method

As shown in Figure 3.6, the network (e.g., the $a \times a$ square) is divided into m^2 micro cells of the size $d \times d$. Here $m = a/d$. We assign each cell a coordinate (i, j), where i and j are between 1 and m, to indicate its position at jth row and ith column. Similarly to Lemma 1 and Lemma 2 in Section 3.1, we can get the following two lemmas:

Lemma 4. [43] Given n random nodes in a $\sqrt{n} \times \sqrt{n}$ square, dividing the square into
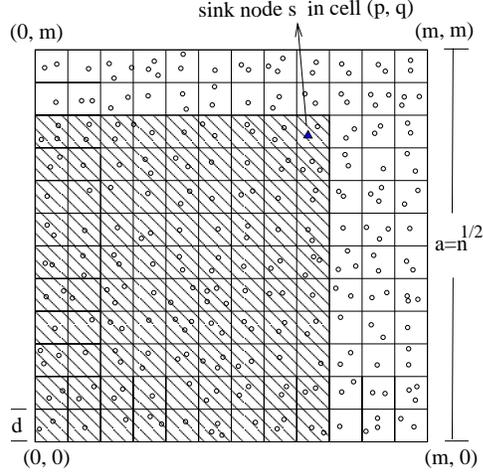
Figure 3.6: Grid partition of the sensor network: a^2 cells with cell size of $d \times d$.

micro cells of the size $\sqrt{3 \log n} \times \sqrt{3 \log n}$, every micro cell is occupied with probability at least $1 - \frac{1}{n^2}$.

Lemma 5. Given n random nodes in a $\sqrt{n} \times \sqrt{n}$ square, dividing the square into micro cells of the size $\sqrt{3 \log n} \times \sqrt{3 \log n}$, the maximum number of nodes in any cell is $O(\log n)$ with probability at least $1 - \frac{3 \log n}{n}$.

Therefore, if we set $d = \sqrt{3 \log n}$ (i.e., $m = \sqrt{\frac{n}{3 \log n}}$), every micro cell has at least one node with high probability and its maximum number of nodes is $O(\log n)$ with high probability. In order to make the whole network connected, the transmission range r need to be equal or larger than $\sqrt{5}d$ so that any two nodes from two neighboring cells are inside each other's transmission range. Hereafter, we set $r = \sqrt{5}d = \sqrt{15 \log n}$.

3.4.1.2 Data Collection Scheme

Same as the analysis in Section 3.2, we only concentrate one direction with the largest number of sensors and assume that the sink is in the upper right corner of the field. For our collection scheme, we first divide the field into big blocks with size $L \times L$ as shown in Figure 3.7. We call these blocks *interference blocks* and L *interference distance*. Thus, the number of interference blocks is $\lceil \frac{a^2}{L^2} \rceil$. We label each block with (i, j) where i and j are the indexes of the block as in Figure 3.7. In our collection
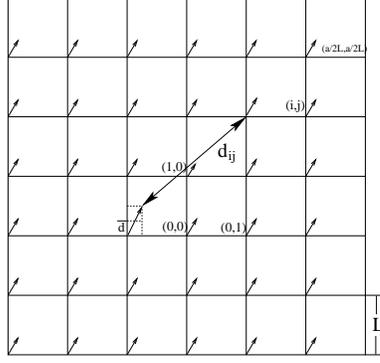
Figure 3.7: Grid partition of interference blocks with size of $L \times L$.

scheme, we schedule data transmission in parallel at all blocks but make sure that there is only one sensor in each interference block transferring at any time. To avoid interference from senders in other interference blocks, we need interference distance L larger than certain value.
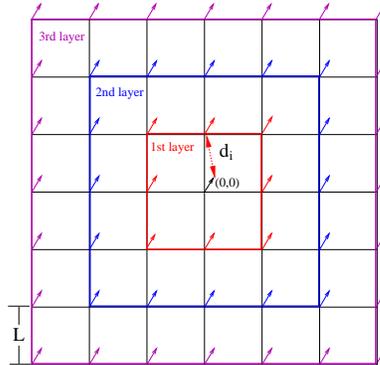
Figure 3.8: Simultaneous transmissions are around the center by layers.

Next, we derive the lower bound of interference distance such that all simultaneous transmissions as shown in Figure 3.7 can be successfully received. Here, we consider the SINR at the receiver in interference block $(0,0)$ (which is in the center of the field) since it has the minimum SINR among all receivers. Similar with the technique used in [47], we now label all simultaneous transmissions by layers from position $(0,0)$, as shown in Figure 3.8. Based on physical interference model, its SINR is at least

$$\frac{P \cdot r^{-\beta}}{B \cdot N_0 + \sum_{\text{all layers } i \geq 1} c_i P \cdot (d_i)^{-\beta}}.$$

Here, d_i is the minimum distance from a transmitter on ith layer to the receiver in block $(0,0)$ and c_i is the number of transmitters on ith layer. Therefore, we need to derive L such that $SINR \geq \eta$, i.e.,

$$\sum_{\text{all layers } i \geq 1} c_i(d_i)^{-\beta} \leq \frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}.$$

Notice that $d_i = iL - 2d$ and $c_i = 8i$. For example, there are 8 transmitters at the first layer with distance at least $L - 2d$ and 16 transmitters at the second layer with distance at least $2L - 2d$, and so on. Thus,

$$
\begin{aligned}
\sum_{i \geq 1} c_i(d_i)^{-\beta} &= \sum_{i \geq 1} 8i(iL - 2d)^{-\beta} \\
&\leq \sum_{i \geq 1} 8i(iL - 2id)^{-\beta} \\
&= 8(L - 2d)^{-\beta} \sum_{i \geq 1} i^{-(\beta-1)}.
\end{aligned}
$$

Since $\beta > 2$, $\sum_{i \geq 1} i^{-(\beta-1)}$ converges to a constant, let it be denoted by ϕ. Then we only need

$$8\phi(L - 2d)^{-\beta} \leq \frac{r^{-\beta}}{\eta} - \frac{BN_0}{P},$$

to guarantee that the SINR at the receiver in the center is at least η. This can be satisfied by setting

$$L \geq \left(\frac{1}{8\phi} \cdot \left(\frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}\right)\right)^{-\frac{1}{\beta}} + 2d.$$

Remember $r \leq \left(\frac{P}{B \cdot N_0 \cdot \eta}\right)^{1/\beta}$, this makes sure we can find such suitable L. We can further select $L = \left(\frac{1}{8\phi} \cdot \left(\frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}\right)\right)^{-\frac{1}{\beta}} + 2d$. Since $r = \sqrt{5}d$,

$$
\begin{aligned}
\frac{L}{d} &= \left(\frac{1}{8\phi} \cdot \left(\frac{(\sqrt{5}d)^{-\beta}}{\eta d^{-\beta}} - \frac{BN_0}{Pd^{-\beta}}\right)\right)^{-\frac{1}{\beta}} + 2 \\
&= \left(\frac{1}{8\phi} \cdot \left(\frac{5^{-\beta/2}}{\eta} - \frac{BN_0 d^{\beta}}{P}\right)\right)^{-\frac{1}{\beta}} + 2.
\end{aligned}
$$

When $n \to \infty$, this ratio goes to a constant, denoted by α.

(a) Phase I, 1st time slot

(b) Phase I, 2nd time slot

(c) Phase II, 1st time slot

(d) Phase II, 2nd time slot
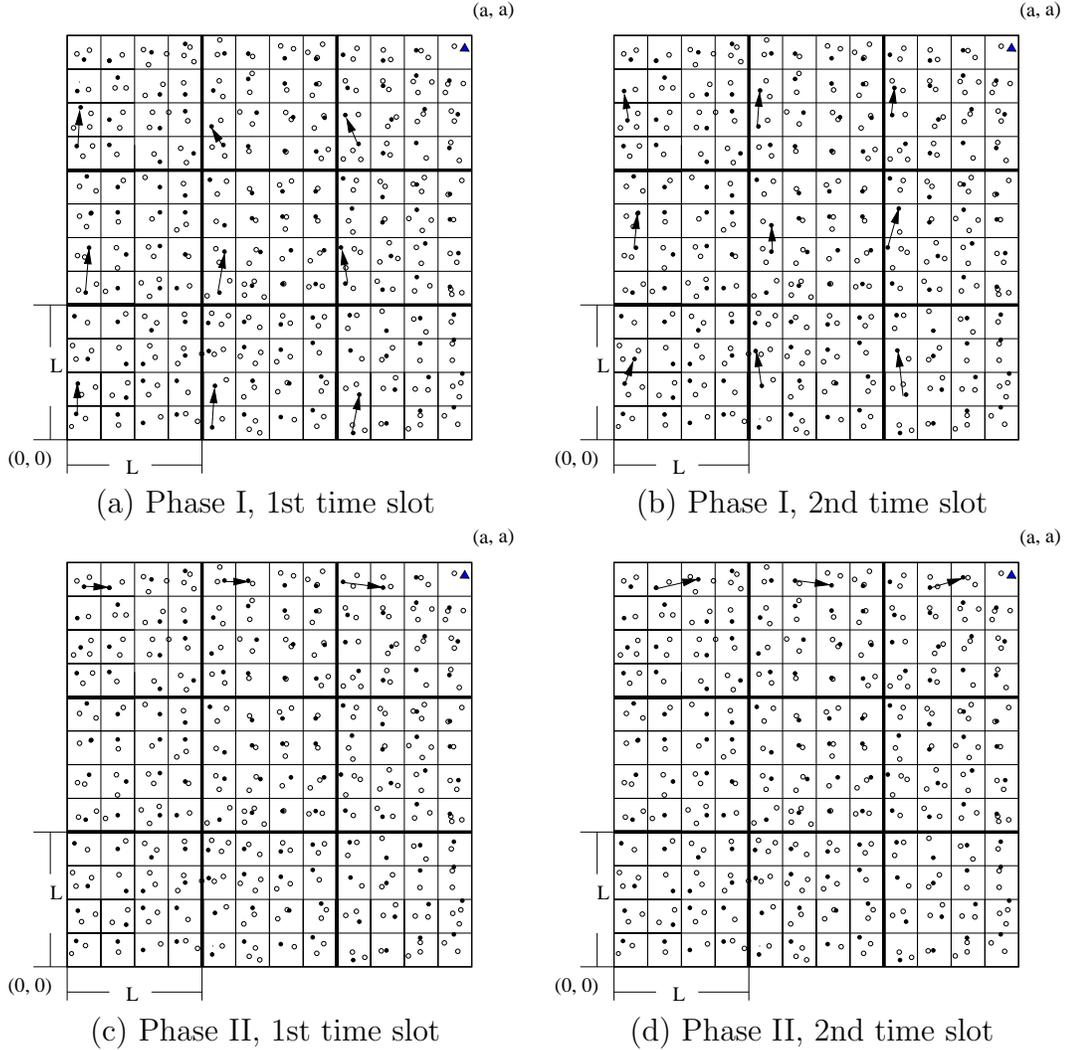
Figure 3.9: Our collection method: [Phase I] each node send its data to its upper cell; [Phase II] each node in the top row send its data to its right cell.

After having interference blocks, we can simply use the data collection scheme which is already presented in Section 3.2.1. Such collection process is illustrated in Figure3.9.

3.4.1.3 Analysis of Delay Rate and Capacity

Similar with the analysis of delay rate under protocol interference model, we also define the time needed for the two phases as T_1 and T_2, respectively.

By Lemma 5, we can get that:

$$
\begin{aligned}
T_1 & \leq \ (\frac{L}{d})^2 \times t \times O(\log n) \times m \\
& \leq \ O(t \log n)m = O(t \log n)\sqrt{\frac{n}{3 \log n}} = O(t\sqrt{n \log n}).
\end{aligned}
$$

and

$$
T_2 \leq \frac{L}{d} \times t \times mO(\log n) \times m \leq m^2 O(t \log n) = O(nt).
$$

Therefore, the total time needed to collect b-bits information from every sensor in the field to the sink is $T_1 + T_2 = O(nt)$. Thus, the total delay Δ_{col} for the sink to receive a complete snapshot is at most $O(nt)$. Consequently, the total delay rate of this collection scheme is

$$
\Gamma_{col} = \frac{nb}{\Delta_{col}} = \Omega(\frac{nb}{nt}) = \Omega(W).
$$

It has been proved that the upper bound of delay rate or capacity of data collection is W [23, 24]. Therefore, the delay rate of our collection scheme achieves the order of the upper bound, and the delay rate of data collection is $\Theta(W)$.

With the pipelining, the time T_1' for the highest cell to receive a new set of $m \cdot b$ data in Phase I is

$$
T_1' \leq (\frac{L}{d})^2 \times t \times O(\log n) = O(t \log n).
$$

And the time T_2' for the sink to receive a new set of $a \cdot b$ data in Phase II is

$$
T_2' \leq \frac{L}{d} \times t \times m = O(t\sqrt{\frac{n}{\log n}}).
$$

Therefore, the total time for sink to receive $m \cdot b$ data is $T_1' + T_2' = O(t\sqrt{\frac{n}{\log n}})$. Thus,

the capacity of our method with pipelining is still

$$C_{col} = \frac{m \cdot b}{T_1' + T_2'} = \Omega(W).$$

We summarize the results in the following theorem:

Theorem 5. Under physical interference model, the delay rate Γ and the capacity C of data collection in random sensor networks with a single sink are both $\Theta(W)$.

3.4.2 Data Collection with Aggregation

We then investigate the data aggregation scenario where each sensor can aggregate its received packets into a single packet. Here, we study both *delay rate* and *capacity* of data aggregation with a single sink. The definitions of delay rate and capacity are similar to those of data collection in Section 1.3.

We again assume that the sink s is located in cell (m, m). Our aggregation scheme is the same with the one in Section 3.2.2, which includes three phases. The only difference is that now the interference block is with size of L instead of $R+r$. However, since L is in the same order of r, the previous analysis in Section 3.2.2 will stay the same.

In summary, we have the following theorem for data aggregation.

Theorem 6. Under physical interference model, the delay rate Γ and the capacity C of data aggregation in random sensor networks with a single sink are $\Theta(\sqrt{n \log n}W)$ and $\Theta(\frac{n}{\log n}W)$ respectively.

3.5 Data Collection under 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 wire-

less communication. We now discuss the capacity bounds in random wireless sensor networks 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

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

Where N_0 is the background Gaussian noise, I is the set of actively transmitting nodes when node v_i is transmitting, $l(v_i, v_j) = \min\{1, ||v_i - v_j||^{-\beta}\}$ is the pass loss exponent and $\beta > 2$, and P is the fixed transmission power.

We consider a sensor network with n sensor nodes $V = \{v_1, v_1, \cdots, v_n\}$ and a single sink node s uniformly deployed in a square region with side-length $a = \sqrt{n}$ with density 1. At regular time intervals, each sensor node measures the field value at its position and transmits the value to the sink node.

First, we give a lemma to derive the upper bound of capacity of data collection under Gaussian channel model.

Lemma 6. An upper bound for data collection capacity under Gaussian channel model is at most $O((\log n)W)$.

Proof. We first order all the incoming links of sink s according to their length as follows: $||v_1, s|| \leq ||v_2, s|| \leq \cdots \leq ||v_{n'}, s||$. Here n' is the number of incoming links at sink s and $n' \leq n$. Next, we try to bound the SINR of the sink node s. For any

link $v_i s$ $(i \neq 1)$, its SINR

$$
\begin{aligned}
SINR_{is} &\leq \frac{P \cdot l(v_i, s)}{N_0 + \sum_{k=1}^{i-1} P \cdot l(v_k, s)^{-\beta}} \\
&\leq \frac{P \cdot l(v_i, s)}{N_0 + \sum_{k=1}^{i-1} P \cdot l(v_i, s)^{-\beta}} < \frac{1}{i-1}
\end{aligned}
$$

Therefore, for $i \neq 1$,

$$
W_{is} = W \log_2(1 + SINR_{is}) < W \log_2(\frac{i}{i-1}).
$$

So the maximum rate at sink v_0 is at most

$$
\begin{aligned}
W_{1s} &+ \sum_{i=2}^{n'} W \log_2(\frac{i}{i-1}) \\
&= W_{1s} + W \log_2(\Pi_{i=2}^{n'} \frac{i}{i-1}) \\
&\leq \max_i(W_{is}) + W \cdot \log_2(n') \\
&\leq \max_i(W_{is}) + W \cdot \log_2(n)
\end{aligned}
$$

And since $\max_i(W_{is}) \leq O(W)$, the upper bound of capacity is $O((\log n)W)$. $\quad \Box$

In order to study the lower bound, we still need to introduce a partition method which we already used in previous sections. As shown in Figure 3.6, the network (e.g., the $a \times a$ square) is divided into m^2 micro cells of the size $d \times d$. Here $m = a/d$. We set $d = \sqrt{3 \log n}$ (i.e., $m = \sqrt{\frac{n}{3 \log n}}$), such that every micro cell has at least one node and at most $O(\log n)$ nodes with high probability.

The data collection scheme under Gaussian channel model is still the same with those under protocol or physical interference models but with different size of the interference block. We now divide the field into big blocks with certain size $L(d) \times L(d)$ as shown in Figure 3.7. We call suck blocks *interference blocks*. Thus, the number
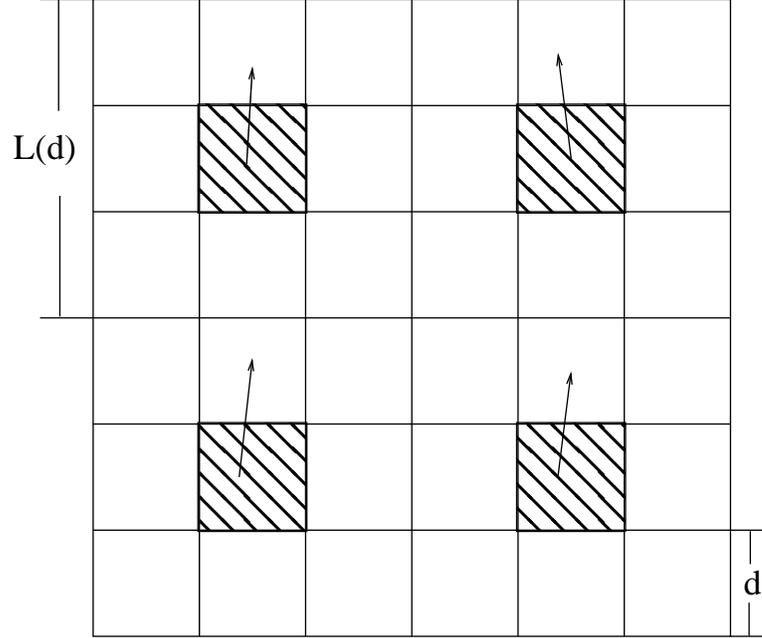
Figure 3.10: Grid partition of interference blocks with size of $L(d) \times L(d)$.

of interference blocks is $\lceil \frac{a^2}{L(d)^2} \rceil$. We label each block with (i,j) where i and j are the indexes of the block as in Figure 3.7. In our collection scheme, we schedule data transmission in parallel at all blocks but make sure that there is only one sensor in each interference block transferring at any time.

We now prove that the transmission rate of each transmitting sensor node in such data collection scheme is at least $\Omega(Wd^{-\beta})$, if $L(d) = \mathcal{C}d$ and $\mathcal{C}$ is a constant and bigger than 4.

Lemma 7. Given n random nodes in a $\sqrt{n} \times \sqrt{n}$ square, in each interference block $\mathcal{C}d \times \mathcal{C}d$, there exists a node that can transmit at rate $\Omega(Wd^{-\beta})$ to any destination in its adjacent cell.

Proof. Let us focus on one given sensor node v_i which transmits to a destination v_j in v_i's adjacent cell. Its transmission rate is:

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

Since the distance between v_i and v_j is at most $\sqrt{5}d$, $P \cdot l(v_i, v_j) \geq P \cdot \sqrt{5}d = \Omega(d^{-\beta})$.

We also need to find the upper bound of the interference at the receiver v_j. As shown in the Figure 3.10, the transmitters in the adjacent eight interference blocks are located at most $L(d) - 2d$ from v_j. Therefore,

$$
\begin{aligned}
\sum_{k \in I} P \cdot l(v_k, v_j) \ &\leq\ \sum_{m=1}^{\infty} 8mP(m \cdot L(d) - 2d)^{-\beta} \\
&\leq\ \sum_{m=1}^{\infty} 8mP(m\mathcal{C} - 2)^{-\beta}d^{-\beta} \\
&\leq\ 8P \cdot d^{-\beta} \cdot \sum_{m=1}^{\infty} m(m\mathcal{C} - 2)^{-\beta}.
\end{aligned}
$$

Since $\beta > 2$, the summation $\sum_{m=1}^{\infty} m(m\mathcal{C}-2)^{-\beta}$ converges to a constant $\mathcal{L}$. Therefore,

$$
\sum_{k \in I} P \cdot l(v_k, v_j) \leq 8P\mathcal{L} \cdot d^{-\beta}\ =\ \Omega(d^{-\beta}).
$$

And when $n \to \infty$, $d \to \infty$, hence,

$$
\frac{P \cdot l(v_i, v_j)}{N_0 + \sum_{k \in I} P \cdot l(v_k, v_j)} = \Omega(d^{-\beta}).
$$

Therefore, the transmission rate

$$
W_{ij} = \Omega(W \cdot d^{-\beta}).
$$

$\square$

We use the same data collection scheme in Section 3.2. The total time we need to collect all the n packets is

$$
T \leq t \cdot \mathcal{C}\sqrt{\frac{n}{3 \log n}} \cdot O(\log n) \cdot \sqrt{\frac{n}{3 \log n}} \cdot \frac{b}{\Omega(W \cdot d^{-\beta})} = O(nt \cdot (\log n)^{\frac{\beta}{2}}).
$$

Thus, the achieved capacity of data collection under Gaussian channel model is $\Omega((\log n)^{-\frac{\beta}{2}}W)$.

In summary, the bounds of collection capacity could be revised as the following:

Theorem 7. Under Gaussian channel model, the upper bound and the lower bound of data collection capacity for random wireless sensor networks are $O((\log n)W)$ and $\Omega((\log n)^{-\frac{\beta}{2}}W)$.

3.6 Summary

The summaries of data collection results under protocol interference model, physical interference model and Gaussian channel model in this section are shown in Table 3.1, Table 3.2 and Table 3.3 respectively.

From the results in Table 3.1, the capacity of random wireless sensor networks indeed increases with the number of sinks raising. But when the number of sinks reaches a certain threshold $((k = \Theta(\frac{n}{\log n})$, the value of capacity trends toward a fixed scale. And if the data can be aggregated, the capacity of collection can be enlarged in order. These results can lead to better network planning and algorithm designing in random wireless sensor networks' applications.

From the results in Table 3.2, both the delay rate and the capacity under physical interference model are the same with those results under protocol interference model.

From the results in Table 3.3, there is a gap between the upper bound and lower bound of capacity under Gaussian channel model.

Table 3.1: Summary of Delay Rate and Capacity in RWSN under ProIM

Task	Sink Number k	Delay Rate and Capacity
data collection	$k = 1$	$\Gamma = \Theta(W)$ $C = \Theta(W)$
data collection	$k = O(\frac{n}{\log n})$ and regularly-deployed	$\Gamma = \Theta(kW)$ $C = \Theta(kW)$
data collection	$k = \Omega(\frac{n}{\log n})$ and regularly-deployed	$\Gamma = \Theta(\frac{n}{\log n}W)$ $C = \Theta(\frac{n}{\log n}W)$
data collection	$k = O(\frac{n}{\log n})$ and randomly-deployed	$\Theta(\frac{k}{\log k}W) \leq \Gamma \leq \Theta(kW)$ $\Theta(\frac{k}{\log k}W) \leq C \leq \Theta(kW)$
data collection	$k = \omega(\frac{n}{\log n})$ and randomly-deployed	$\Gamma = \Theta(\frac{n}{\log n}W)$ $C = \Theta(\frac{n}{\log n}W)$
data aggregation	$k = 1$	$\Gamma = \Theta(\sqrt{n \log n}W)$ $C = \Theta(\frac{n}{\log n}W)$
data aggregation	$k = O(\frac{n}{\log n})$ and regularly-deployed	$\Gamma = \Theta(k\sqrt{n \log n}W)$ $C = \Theta(\frac{kn}{\log n}W)$
data aggregation	$k = \Omega(\frac{n}{\log n})$ and regularly-deployed	$\Gamma = \Theta(\frac{n\sqrt{n}}{\sqrt{\log n}}W)$ $C = \Theta((\frac{n}{\log n})^2 W)$
data aggregation	$k = O(\frac{n}{\log n})$ and randomly-deployed	$\Theta(\frac{k\sqrt{n \log n}}{\log k}W) \leq \Gamma \leq \Theta(k\sqrt{n \log n}W)$ $\Theta(\frac{kn}{\log k \log n}W) \leq C \leq \Theta(\frac{kn}{\log n}W)$
data aggregation	$k = \omega(\frac{n}{\log n})$ and randomly-deployed	$\Gamma = \Theta(\frac{n\sqrt{n}}{\sqrt{\log n}}W)$ $C = \Theta((\frac{n}{\log n})^2 W)$

Table 3.2: Summary of Delay Rate and Capacity in RWSN under PhyIM

Task	Sink Number k	Delay Rate and Capacity
data collection	$k = 1$	$\Gamma = \Theta(W)$ $C = \Theta(W)$
data aggregation	$k = 1$	$\Gamma = \Theta(\sqrt{n \log n}W)$ $C = \Theta(\frac{n}{\log n}W)$

Table 3.3: Summary of Capacity in RWSN under Gaussian Channel Model

Task	Upper Bound	Lower Bound
data collection	$O((\log n)W)$	$\Omega((\log n)^{-\frac{\beta}{2}}W)$

We have studied the capacity of data collection on large-scale random wireless sensor networks. However, all these results are based on a strong assumption that sensors are deployed randomly in an environment and the number of nodes n must be extremely large. But in many practical sensor applications, the sensor network is not deployed uniformly and the number of sensors may not be large enough. For example, wireless sensor networks are used to detect intruders and monitor the area by deploying along the border line [48]. In that case, the sensors must be deployed carefully on the border line to form a barrier. Our analysis result in Section 3 can not hold and data collection algorithm in Section 3.2 is infeasible. Therefore, the capacity of data collection in an arbitrary sensor network need to be reconsidered. In arbitrary sensor networks, sensors's distribution may be very strange and uneven, as show in Figure 4.1 (b). How to efficiently collect data in such arbitrary sensor networks seems more difficult than that in random sensor networks.
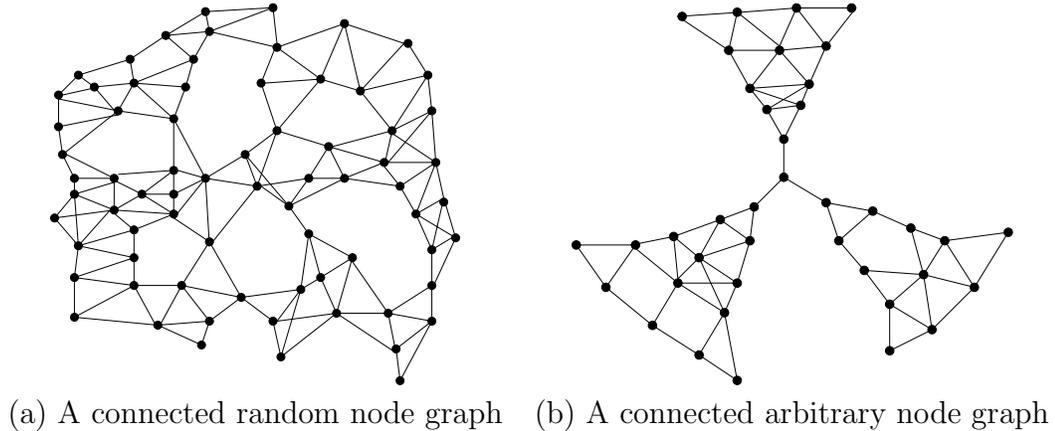

(a) A connected random node graph   (b) A connected arbitrary node graph

Figure 4.1: Random graph vs arbitrary graph.

4.1   Preliminaries

We use the network model mentioned in Section 1.1: a wireless sensor network consists n sensor $v_1, v_2, \cdots, v_n$ and a sink set S; both sensor nodes and sink nodes are deployed in a plain area; at regular time intervals, each sensor node measures the field value at its position and transmits the value to one of the sinks; a fixed data-rate channel model is adopted where each wireless node can transmit at W bits/second over a common wireless channel; all packets have a unit size of b bits and each sensor has a fixed transmission power P; time is partitioned into slots with $t = b/W$ seconds. The only difference between the model we used in this section and the model in Section 3 is that in this section we assume that these n sensors are arbitrarily distributed but not randomly and uniformly distributed.

Under protocol interference model, since every node has a fixed transmission power P, 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. We use protocol interference model for simplifying the analysis. Thus, all nodes have a uniform interference range R and node v_j can receive the signal successfully from node v_i if no node within a distance R from v_j is transmitting simultaneously. Similarly with Section 1.2, 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, $i = 0, 1, \cdots, n$.

4.2   Data Collection under Protocol Interference Model

We first consider the case that there only exists one sink in the sensor network. It has been proved that the upper bound of data collection capacity for random networks with a single sink is W in Section 3. It is obviously that this upper bound also holds
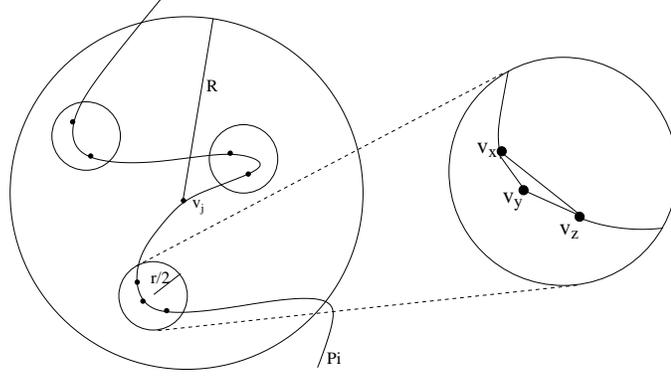
Figure 4.2: The number of interference nodes for a node v_j is bounded by a constant on a path P_i.

for any arbitrary network. The reason is that sink s_1 cannot receive at rate faster than W since W is the fixed transmission rate of individual link. Through this work, we plan to construct a scheduling scheme to achieve capacity in the same order of the upper bound, *i.e.* $\Theta(W)$

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.

4.2.1 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 s_1. 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 s_1. 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.

Lemma 8. 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 4.2 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 4.2. This is a contradiction with the construction of BFS tree. As shown in Figure 4.2, 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. □
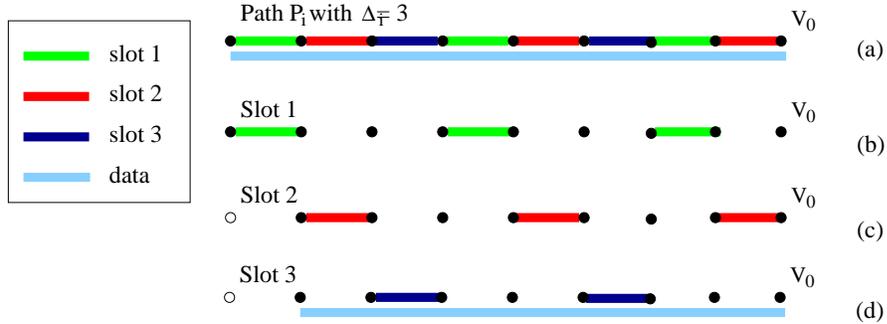
Figure 4.3: Scheduling on a path: after Δ_i slots the sink gets one data.

4.2.2   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 4.3 for illustration. 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 4.3(a). Notice that each node on the path has unit data to transfer. Links with the
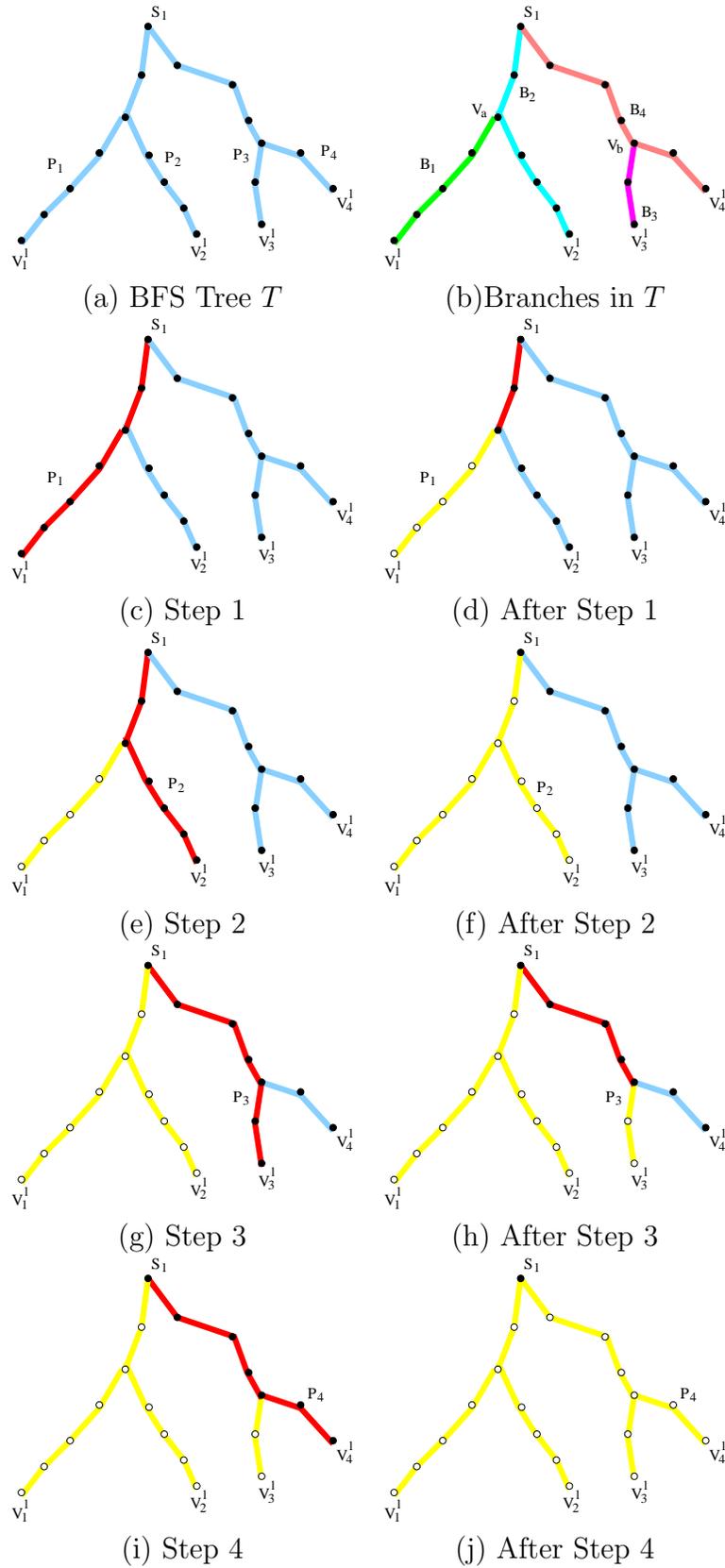
Figure 4.4: Illustrations of our scheduling on the data collection tree T.

same color are active in the same slot. After three slots (Figure 4.3(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 4.4(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 s_1, B_3 is from v_3^l to v_b, and B_4 is from v_4^l to s_1. Notice that the union of all branches is the whole tree T. Algorithm 1 shows the detailed branch scheduling algorithm. Figure 4.4(c)-(j) give an example of scheduling on T. In the first step (Figure 4.4(c)), all nodes on P_1 participate in the collection using the scheduling method for a single path (every Δ_1 slots, sink s_1 receives one data). Such collection stops until there is no data in this snapshot on branch B_1, as shown in Figure 4.4(d). Then Step 2 collects data on path P_2. This procedure repeats until all data in this snapshot reaches s_1 (Figure 4.4(j)).

Algorithm 1: Branch Scheduling on BFS Tree

---

1 **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 s_1 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**

4.2.3 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 in one snapshot.

Theorem 8. 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 8, 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 9. Under protocol interference model and disk graph model, data collection capacity for arbitrary wireless sensor networks is $\Theta(W)$.

4.3 Data Collection under Physical Interference Model

For the arbitrary wireless sensor networks, 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.* [46] 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

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

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

Proof. To give an upper bound on the capacity of data collection, we will show that we can set an artificial transmission range r_0 and an artificial interference range R_0 such that (1) the receiving node v_j of a sender v_i is within distance r_0, and (2) a transmitting node v_k will cause interference at node v_j within distance R_0. I.e., if there is any interference among nodes in protocol interference model with these artificial ranges, there is also interference among them in physical interference model.

Given P, N_0, and η, we choose artificial ranges as follows:

$$r_0 \leq (\frac{P}{B \cdot N_0 \cdot \eta})^{1/\beta} \ and \ R_0 < (\frac{\eta \cdot P}{P - B \cdot N_0 \cdot \eta})^{1/\beta}.$$

Notice that the definition of R_0 is valid since $P - B \cdot N_0 \cdot \eta > 0$.

First, if the receiving node v_j within distance of d from the sender v_i can correctly decode the signal, $\frac{P \cdot d^{-\beta}}{BN_0} \geq \eta$. Obviously if $d \leq r_0 \leq (\frac{P}{B \cdot N_0 \cdot \eta})^{1/\beta}$, $\frac{P \cdot d^{-\beta}}{BN_0} \geq \eta$ holds. Thus r_0 is the maximum distance of a successful communication.

Second, if a receiving node v_j is within a distance R_0 of a transmitting node v_k and v_i is the legitimate sender of v_j, the SINR at node v_j is at most $\frac{P}{BN_0 + P||v_k - v_j||^{-\beta}} \leq$

$\frac{P}{BN_0+PR_0^{-\beta}}$, since the maximum strength of the signal from v_i received at node v_j is at most P. Given $R_0 < (\frac{\eta \cdot P}{P - B \cdot N_0 \cdot \eta})^{1/\beta}$, thus the SINR $< \eta$. Therefore, the node v_j cannot receive data from v_i correctly due to the interference from v_k.

By artificially setting r_0 and R_0 (which are both constants), we convert the physical interference model into a protocol interference model. Using previous proofs in protocol interference model, it is straightforward to show that the upper bound on the capacity under disk graph model is bounded by $\Theta(W)$ where the constant behind the $\Theta()$ is related to $\frac{R_0}{r_0}$.

To give a lower bound on the capacity of data collection, we then set an artificial transmission range r_1 and an artificial interference range R_1 such that, when all simultaneously transmitting nodes are separated by a distance R_1, and the receiving nodes of a transmitting node is within r_1, the SINR of every receiving node is at least η. In other words, if there is no interference among nodes in the protocol interference model with artificial ranges r_1 and R_1, there is no interference among the nodes in the physical interference model as well.
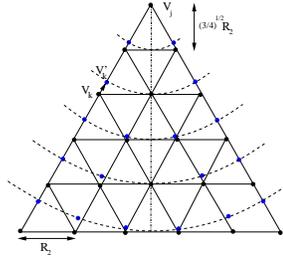
Figure 4.5: Illustration of positions of active transmitting nodes causing the maximum interference at v_j. Here only nodes in one $60°$ direction from v_j are shown.

Consider v_i transmits signal to v_j and there are other active transmitting nodes v_k. Thus, the interference at the receiver v_j by all other transmitting nodes $v_k \in I$ is $\sum_{v_k \in I} P \cdot ||v_k - v_j||^{-\beta}$. Since every pair of active transmitting nodes need to be separated by at least $R_2 = R_1 - r_1$, the maximum number of active transmitting nodes are distributed in a regular triangulation pattern as shown in Figure 4.5. To

calculate the upper bound of the maximum interference at v_j, we move every active transmitting node v_k to the position of v'_k on a circle centered at v_j, as shown in Figure 4.5. On the i-th circle from v_j (whose radius is $i\frac{\sqrt{3}}{2}R_2$), there are $6i$ active nodes. Therefore, the maximum interference at v_j is at most

$$\sum_{i=1}^{\infty} 6i \cdot P(i\frac{\sqrt{3}}{2}R_2)^{-\beta} = \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta} \sum_{i=1}^{\infty} \frac{1}{i^{\beta-1}} \leq \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta}\zeta.$$

Here ζ is a constant bound of $\sum_{i=1}^{\infty} \frac{1}{i^{\beta-1}}$ for $\beta > 2$. Thus, the SINR at v_j is at least

$$\frac{P \cdot ||v_i - v_j||^{-\beta}}{B \cdot N_0 + \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta}\zeta} \geq \frac{Pr_1^{-\beta}}{B \cdot N_0 + \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta}\zeta}.$$

The last inequality is due to $||v_i - v_j|| \leq r_1$. To make the SINR value $\geq \eta$, we need

$$r_1 \leq (\frac{P}{(B \cdot N_0 + \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta}\zeta) \cdot \eta})^{1/\beta}.$$

Then we can carefully choose r_1 and R_2 such that the above inequation holds. Notice that r_1 need to be larger than 1 to satisfy $P \cdot r_1^{-\beta} < P$. Since $\eta < \frac{P}{B \cdot N_0}$, we can choose sufficiently large constant R_2 such that $\frac{P}{B \cdot N_0 + \frac{6P}{(\frac{\sqrt{3}}{2}R_2)^\beta}\zeta} > \eta$. Then we can set $R_1 = r_1 + R_2$.

By artificially setting r_1 and R_1, we can convert the physical interference model into a protocol interference model. Using previous collection algorithms for protocol interference model, it can be shown that the lower bound $\Theta(W)$ on the capacity of data collection under disk graph model is achievable.

$\square$

4.4 Data Collection for General Graph Model

In previous part of Section 4, our collecting algorithm and analysis are based on a unit 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 $v_i v_j \notin G$). 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 4.6(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 4.6(b). Clearly, a scheduling that lets every node transfer data in order can lead to a capacity W which is much larger than $\frac{W}{\Delta} = \frac{W}{n}$. From these two examples, we find that the capacity problem for general graph model is

more complexity. We study the problem under both protocol interference model and Gaussian channel model.

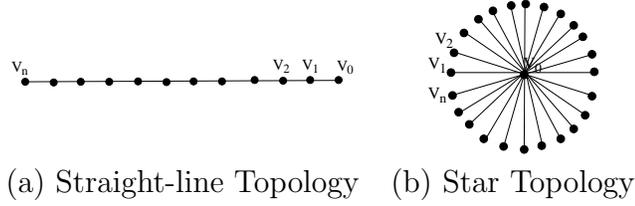(a) Straight-line Topology   (b) Star Topology

Figure 4.6: The optimum of BFS-based method under two extreme cases.

Under protocol interference model, we analyze the upper bound and lower bound of the collection capacity for the general graph model.

4.4.1   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 4.7 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
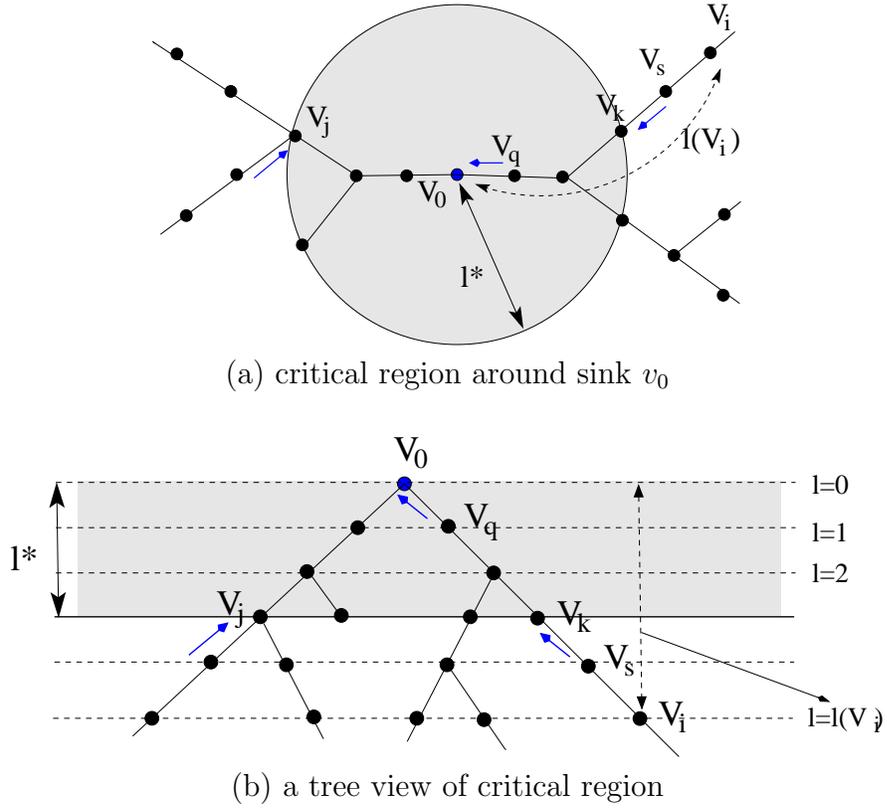
Figure 4.7: Illustration of the definition of critical region, *i.e.* l^*. The grey area is the critical region, where no any two nodes can receive a message in the same time slot due to interference around v_0.

Lemma 9. 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).

4.4.2 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 $\tilde{\Delta}$ is not bounded by a constant any more, and it could be $O(1)$ or $O(n)$. Thus, there is a gap between our lower bound of data collection $\frac{W}{\tilde{\Delta}}$ and the natural upper bound W. Considering both examples shown in Figure 4 of the paper, BFS-based method matches their tight upper bounds $\Theta(\frac{W}{n})$ and W. For the star topology, even though the sink has the maximal interference $\Delta = n$, each individual path has the path interference $\Delta_i = 1$ which leads to capacity of $\frac{W}{1} = W$. For the straight-line topology, the path interference of the single path $\Delta_i = n$, thus the capacity is $\frac{W}{n}$. In both cases, $\frac{W}{\tilde{\Delta}}$ matches the optimal capacity. However, similar to $\frac{W}{\Delta}$, $\frac{W}{\tilde{\Delta}}$ is still not a tight bound too. We will show such an example in Figure 4.8.

Now we modify the basic *Path Scheduling* of the BFS-method to achieve better collection capacity. Recall that in Section IV.B we claim that the path scheduling for a path P_i can be done in $\Delta_i \cdot |P_i|$ time slots. However, we can perform path scheduling in the following way to save more slots. Assume that path $P_i = v_0, v_1, v_2, \cdots, v_{|P_i|}$. Let $\delta_k^{P_i} = \max\{\delta^{P_i}(v_1), \cdots, \delta^{P_i}(v_k)\}$, *i.e.*, $\delta_k^{P_i}$ is the maximum interference number among *first* k nodes v_1 to v_k in path P_i. Clearly, $\delta_k^{P_i} \leq \delta_{k+1}^{P_i}$. In the first step, using $\delta_{|P_i|}^{P_i}$ slots, every node on the path transfers its data to its parent in the BFS tree. After the first step, the leaf $v_{|P_i|}$ already finishes its task in this round and has no data from current snapshot. In the second step, using $\delta_{|P_i|-1}^{P_i}$ slots, the current snapshot data

will move one more level up along the path in BFS tree. Repeat these steps until all data along this path reach the sink. It is easy to show that the total number of time slots used by the above procedure is $\sum_{k=1}^{|P_i|} \delta_k^{P_i}$. Since $\delta_k^{P_i} \leq \Delta_i$, $\sum_{k=1}^{|P_i|} \delta_k^{P_i} \leq \Delta_i \cdot |P_i|$.
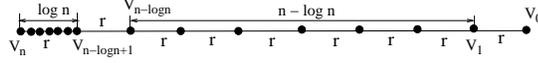

Figure 4.8: Illustration of the advantage of a new path scheduling.

Figure 4.8 shows an example where $\sum_{k=1}^{|P_i|} \delta_k^{P_i}$ is much smaller than $\Delta_i \cdot |P_i|$. Again we have n sensors and the sink distributed on a line P as shown in the figure. Assume that $R = r$. On the left side, there are $\log n$ nodes close to each other, thus their $\delta(v_i) = \log n$ except for $\delta(v_{n-\log n+1}) = \log n + 1$. On the right side, every node has $\delta(v_i) = 3$. Thus, $\Delta = \log n + 1$ and $\Delta \cdot |P| = \Theta(n \log n)$. In addition, $\delta_k^P = \log n + 1$ for $k = n - \log n + 1, \cdots, n$ and $\delta_k^P = 3$ for $k = 3, \cdots, n - \log n$, $\delta_2^P = 2$, and $\delta_1^P = 1$. Therefore, $\sum_{k=1}^{|P|} \delta_k^P = (\log n + 1) \log n + 3(n - \log n) - 3 = \Theta(n)$. It is obvious that $\sum_{k=1}^{|P|} \delta_k^P = \Theta(n)$ is smaller than $\Delta \cdot |P| = \Theta(n \log n)$ in order.

Using the above new path scheduling analysis, we now derive a tight lower bound for our BFS-based method. Recall that our method transfers data based on branches in BFS tree T. Given T, there are c paths P_i and c branches B_i as shown in Figure 3(a) and 3(b) in the paper. Then the total number of time slots used by Algorithm 1 with new path scheduling is at most

$$\sum_{i=1}^{c} \sum_{k=|P_i|-|B_i|+1}^{|P_i|} \delta_k^{P_i}.$$

It is clear that this number is much smaller than $\sum_{i=1}^{c} \Delta_i \cdot |B_i|$ from previous analysis. Notice that for path P_i our algorithm (Line 3-4 in Algorithm 1) will terminate the transmission until the branch B_i does not have data for current snapshot and switch

to next path P_{i+1}. Thus, the index of k is only from $|P_i|$ to $|P_i| - |B_i| + 1$. Therefore, the capacity achieved by our algorithm is at least

$$\frac{W}{\frac{\sum_{i=1}^{c} \sum_{k=|P_i|-|B_i|+1}^{|P_i|} \delta_k^{P_i}}{n}}.$$

Let $\Delta^{**} = \frac{\sum_{i=1}^{c} \sum_{k=|P_i|-|B_i|+1}^{|P_i|} \delta_k^{P_i}}{n}$ which can be derived given the BFS tree. We now have a new lower bound of collection capacity as $\frac{W}{\Delta^{**}}$. Here Δ^{**} is a kind of weighted-average of the maximum interference among paths P_i and branches B_i in the BFS tree. We then have the following relationship:

$$n \geq \Delta \geq \tilde{\Delta} \geq \Delta^{**} \geq 1,$$

among the maximum interference number Δ in the whole graph, the maximum interference number $\tilde{\Delta}$ in the paths/braches of BFS tree, and the "average" maximum interference Δ^{**} in the paths/branches of BFS tree. These three interference numbers can be different from each other in order.

Now we introduce a new greedy-based scheduling algorithm which is inspired by [40]. 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

---
1 **Input**: BFS tree T.
  1: Compute the priority $\rho_i = 1/l(v_i)$ of each message p_i.
  2: **for** each snapshot **do**
  3:    **while** $\exists p_j$ such that $v_j^\tau \neq v_0$ **do**
  4:       **for all** such p_i in decreasing order of priority ρ_i **do**
  5:          **if** sending p_i from node v_i^τ will not create interference with any higher-priority messages that are already scheduled for this time slot **then**
  6:            node v_i^τ sends p_i to its parent $par(v_i^\tau)$ in T.
  7:          **end if**
  8:       **end for**
  9:    $\tau = \tau + 1$.
  10:   **end while**
  11: **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.*, $D_j = t \cdot \min\{\tau : v_j^\tau = v_0\}$.
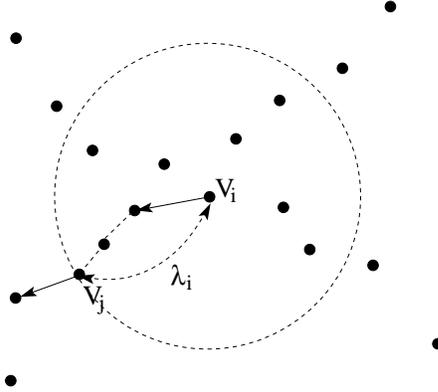

Figure 4.9: Illustration of the definitions of λ_i.

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 transmis-

sion from v_j to $par(v_j)\} + 1$. Here $par(v_i)$ is the parent of v_i in T. See Figure 4.9 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 $\lambda \geq \lambda^*$ in Lemma 10.

Lemma 10. 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 and let $\lambda = \max_i\{\lambda_i\}$, then $\lambda \geq \lambda^*$.

Proof. Let v_k be the node inside critical region with the largest level. We now consider two cases.

Case 1: If there is a node outside the critical region, as shown in Figure 5(a) in the paper, the transmission from v_s to v_k should interfere with the transmission from v_q to v_0. Thus, in the view of v_s, its $\lambda_s \geq l^* + 1 = \lambda^*$. Therefore $\lambda \geq \lambda^*$.

Case 2: If all nodes are inside the critical region, again consider the v_k with largest level. Then $\lambda = \lambda_k = l(v_k) + 1 > l(v_k) = \lambda^*$.

Consequently, we conclude $\lambda \geq \lambda^*$.

□

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 11. 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. $\square$

Lemma 12. 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 11 and $\lambda \geq \lambda^*$,

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

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 11. 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 4.6, the greedy method matches the optimal solutions in order. For the straight-line topology in Figure 4.6(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 4.6(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.

Compared with the lower bound $\frac{\lambda^*}{\lambda} \frac{W}{\Delta^*}$ which we derive from greedy scheduling

on BFS tree, this new lower bound $\frac{W}{\Delta^{**}}$ may be smaller in some cases. Consider the example in Figure 4.8, $\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*} = \Theta(\frac{W}{\log n})$, while $\frac{W}{\Delta^{**}} = \frac{W}{\Delta^*} = \Theta(W)$. However, the reason is mainly due to the rough relaxation in our capacity analysis of greedy scheduling.

In summary, the bounds of collection capacity could be revised as the following:

Theorem 12. Under protocol interference model and general graph model, data collection capacity for arbitrary sensor networks is at least $min(\frac{\lambda^*}{\lambda}\frac{W}{\Delta^*}, \frac{W}{\Delta^{**}})$ and at most $\frac{W}{\Delta^*}$.

4.5 Data Collection under Gaussian Channel Model

We can also study the capacity bounds in arbitrary distributed wireless sensor networks under *Gaussian channel model* for either disk graph model or general graph model. The upper bound of data collection (Lemma 6) we get for random networks still holds, since it is a very general result.

Theorem 13. An upper bound for data collection capacity under Gaussian channel model is at most $O((\log n)W)$.

The proof of this theorem is exactly same with the one of Lemma 6, thus we ignore it here. A lower bound of data collection capacity in this model is still open.

4.6 Summary

In this section, we study the theoretical limits of data collection in terms of capacity for arbitrary wireless sensor networks. For protocol interference model, 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 $\Theta(\frac{W}{\Delta^*})$. we also discuss the collection capacity under a more general model: physical

interference model. By setting artificial transmission range and interference range, we convert the physical interference model into a protocol interference model and get the same collection capacity with that under protocol interference model. For general graph model under protocol interference model, we derive the upper bound of collection capacity and introduce a new data collecting algorithm to approximate the upper bound. At last, we discuss the general upper bound of data collection under Gaussian channel model.

Table 4.1 summarizes our results. All of our methods can achieve these results for random networks too.

Table 4.1: Summary of Data Collection Capacity in Arbitrary WSNs

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	$C \leq O((\log n)W)$

CHAPTER 5:   CONCLUSION

In this dissertation, we study theoretical limitations of data collection in terms of capacity for both random and arbitrary wireless sensor networks under three different kinds of communication models. We briefly summarize our completed work.

We make the following contributions for randomly distributed wireless sensor networks:

For sensor networks with k sinks under protocol interference model, when sensor networks have regularly deployed sinks, we prove that the capacity increases to $\Theta(kW)$ if $k = O(\frac{n}{\log n})$ and $\Theta(\frac{n}{\log n}W)$ if $k = \Omega(\frac{n}{\log n})$. These results show that (1) when k is small the capacity is $\Theta(kW)$ since there will be no interference among neighboring sinks with high probability; however, (2) when k is large the capacity is bounded by the number of interference areas instead of k.

For sensor networks with k sink under protocol interference model, when sensor networks have randomly deployed sinks, we prove that the capacity is between $\Theta(\frac{k}{\log k}W)$ and $\Theta(kW)$ if $k = O(\frac{n}{\log n})$ and $\Theta(\frac{n}{\log n}W)$ if $k = \omega(\frac{n}{\log n})$. Notice that there is a gap between the upper bound and lower bound of data collection capacity when $k = O(\frac{n}{\log n})$.

For sensor networks with a single sink under protocol interference model, with data aggregation, the capacity of data collection is increasing. We theoretically prove that the delay rate and the capacity of data aggregation are $\Theta(\sqrt{n \log n}W)$ and $\Theta(\frac{n}{\log n}W)$ respectively. Thus, pipelining can increase the capacity in order of $\Theta(\sqrt{\frac{n}{\log^3 n}})$.

By penetrating the connection between protocol interference model and physical interference model, we successfully convert the physical interference model into the

protocol interference model so that our proposed data collection scheme still works for the case with a single sink. Both the delay rate and the capacity under physical interference model are the same with those under protocol interference model.

For sensor networks with a single sink under Gaussian channel model, we derive an upper bound of $O(\log n)W$ and a lower bound of $\Omega((\log n)^{-\frac{\beta}{2}}W)$ for data collection.

For arbitrarily distributed wireless sensor networks, we make the following contributions:

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.

For arbitrary sensor networks under physical interference model and disk graph model, we prove that the capacity of data collection is still in the same order as the one under protocol interference model.

Since the disk graph model is idealistic, we also consider a more practical network model: *general graph model*. 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 4.4.

Finally, we discuss the data collection capacity under Gaussian channel model (for disk or general graph model) and derive an upper bound of data collection capacity $O((\log n)W)$.

There are still several open problems left as our future work:

For random sensor network under Gaussian channel model, there is still a gap between the upper and lower bounds, we are interested in reducing or filling it in the future.

For arbitrary sensor network under protocol interference model and disk graph model, we are interested in the capacity problem of data collection if there are multiple sinks or the data can be aggregated.

For arbitrary sensor network under Gaussian channel model, we plan to design a data collection scheme to match or approximate the upper bound.

In all of our results, we assume that underlying communication is always reliable. However, recently, there is new results on capacity with probabilistic model []. It is interesting to further study various capacity problems under such model.

REFERENCES

[1] "http://fiji.eecs.harvard.edu/volcano."

[2] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 102 - 114, 2002.

[3] M. Tubaishat and S. Madria, "Sensor networks: An overview," *IEEE Potentials*, vol. 22(2), no. 20 - 23, 2003.

[4] F. Zhao and L. Guibas, *Wireless Sensor Networks: An Information Processing Approach*. Morgan Kaufmann, 2004.

[5] N. Xu, "A survey of sensor network applications," *IEEE Communications Magazine*, vol. 40, 2002.

[6] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*, ser. WSNA '02, 2002, pp. 88–97.

[7] R. Szewczyk, A. Mainwaring, J. Polastre, J. Anderson, and D. Culler, "An analysis of a large scale habitat monitoring application," in *Proceedings of the 2nd international conference on Embedded networked sensor systems*, ser. SenSys '04, 2004, pp. 214–226.

[8] G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, and W. Hong, "A macroscope in the redwoods," in *Proceedings of the 3rd international conference on Embedded networked sensor systems*, ser. SenSys '05, 2005, pp. 51–63.

[9] C. Hartung, R. Han, C. Seielstad, and S. Holbrook, "Firewxnet: a multi-tiered portable wireless system for monitoring weather conditions in wildland fire environments," in *Proceedings of the 4th international conference on Mobile systems, applications and services*, ser. MobiSys '06, 2006, pp. 28–41.

[10] P. Juang, H. Oki, Y. Wang, M. Martonosi, L. S. Peh, and D. Rubenstein, "Energy-efficient computing for wildlife tracking: design tradeoffs and early experiences with zebranet," in *Proceedings of the 10th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS-X, 2002, pp. 96–107.

[11] M. Welsh, G. Werner-Allen, K. Lorincz, O. Marcillo, J. Johnson, M. Ruiz, and J. Lees, "Sensor networks for high-resolution monitoring of volcanic activity," in *Proceedings of the twentieth ACM symposium on Operating systems principles*, ser. SOSP '05, 2005, pp. 1–13.

[12] G. Werner-Allen, K. Lorincz, M. Welsh, O. Marcillo, J. Johnson, M. Ruiz, and J. Lees, "Deploying a wireless sensor network on an active volcano," *IEEE Internet Computing*, vol. 10, pp. 18–25, March 2006.

[13] S. Kim, S. Pakzad, D. Culler, J. Demmel, G. Fenves, S. Glaser, and M. Turon, "Health monitoring of civil infrastructures using wireless sensor networks," in *Proceedings of the 6th international conference on Information processing in sensor networks*, ser. IPSN '07, 2007, pp. 254–263.

[14] "http://www.alertsystems.org/."

[15] P. Gupta and P. Kumar, "The capacity of wireless networks," *IEEE Transactions on Information Theory*, vol. 46, no. 2, pp. 388–404, 2000.

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

[17] B. Liu, P. Thiran, and D. Towsley, "Capacity of a wireless ad hoc network with infrastructure," in *MobiHoc '07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing.* New York, NY, USA: ACM, 2007, pp. 239–246.

[18] X.-Y. Li, S.-J. Tang, and O. Frieder, "Multicast capacity for large scale wireless ad hoc networks," in *MobiCom '07: Proceedings of the 13th annual ACM international conference on Mobile computing and networking.* New York, NY, USA: ACM, 2007, pp. 266–277.

[19] X. Mao, X.-Y. Li, and S. Tang, "Multicast capacity for hybrid wireless networks," in *MobiHoc '08: Proceedings of the 9th ACM international symposium on Mobile ad hoc networking and computing.* New York, NY, USA: ACM, 2008, pp. 189–198.

[20] S. Shakkottai, X. Liu, and R. Srikant, "The multicast capacity of large multihop wireless networks," in *MobiHoc '07: Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing.* New York, NY, USA: ACM, 2007, pp. 247–255.

[21] A. Keshavarz-Haddad, V. Ribeiro, and R. Riedi, "Broadcast capacity in multihop wireless networks," in *MobiCom '06: Proceedings of the 12th annual international conference on Mobile computing and networking.* New York, NY, USA: ACM, 2006, pp. 239–250.

[22] B. Tavli, "Broadcast capacity of wireless networks," *Communications Letters, IEEE*, vol. 10, pp. 68–69, 2006.

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

[24] 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. International Workshop on Information Processing in Sensor Networks*, 2003.

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

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

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

[28] R. Barton and R. Zheng, "Order-optimal data aggregation in wireless sensor networks using cooperative time-reversal communication," in *2006 40th Annual Conference on Information Sciences and Systems*, 2006, pp. 1050–1055.

[29] S. Li, Y. Liu, and X.-Y. Li, "Capacity of large scale wireless networks under gaussian channel model," in *MobiCom '08: Proceedings of the 14th ACM international conference on Mobile computing and networking*. ACM, 2008, pp. 140–151.

[30] X.-Y. Li, S. Tang, and X. Mao, "Capacity bounds for large scale wireless ad hoc networks under gaussian channel model," in *Proceedings of the 6th Annual IEEE communications society conference on Sensor, Mesh and Ad Hoc Communications and Networks*, ser. SECON'09, 2009, pp. 180–188.

[31] S. Ji, Y. Li, and X. Jia, "Capacity of dual-radio multi-channel wireless sensor networks for continuous data collection," in *Proceedings of IEEE INFOCOM*, 2011.

[32] S. Ji, R. Beyah, and Y. Li, "Continuous data collection capacity of wireless sensor networks under physical interference model," in *Proceedings of IEEE MASS*, 2011.

[33] S. Ji and Z. Cai, "Distributed data collection and its capacity in asynchronous wireless sensor networks," in *Proceedings of IEEE INFOCOM*, 2012.

[34] S. Ji, R. Beyah, and Z. Cai, "Snapshot/continuous data collection capacity for large-scale probabilistic wireless sensor networks," in *Proceedings of IEEE INFOCOM*, 2012.

[35] A. Giridhar and P. Kumar, "Computing and communicating functions over sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 755–764, 2005.

[36] T. Moscibroda, "The worst-case capacity of wireless sensor networks," in *IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks.* New York, NY, USA: ACM, 2007, pp. 1–10.

[37] S. C.-H. Huang, P.-J. Wan, C. T. Vu, Y. Li, and F. Yao, "Nearly constant approximation for data aggregation scheduling in wireless sensor networks," in *Proceedings of IEEE INFOCOM 2007*, 2007.

[38] X. Chen, X. Hu, and J. Zhu, "Minimum data aggregation time problem in wireless sensor networks," in *Proceedings of IEEE 1st Int'l Conference on Mobile Ad-hoc and Sensor Networks (MSN05)*, 2005.

[39] J. Zhu and X. Hu, "Improved algorithm for minimum data aggregation time problem in wireless sensor networks," *Journal of System Science and Complexity*, vol. 21, pp. 626–636, 2008.

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

[41] J.-C. Bermond, J. Galtier, R. Klasing, N. Morales, and S. Pereenes, "Hardness and approximation of gathering in static radio networks," *Parallel Processing Letters*, vol. 16, no. 2, pp. 165–183, 2006.

[42] C. Wang, X.-Y. Li, C. Jiang, and S. Tang, "General capacity scaling of wireless networks," in *Proceedings of IEEE INFOCOM*, 2011.

[43] S. R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1041–1049, 2004.

[44] S. Rao, "The m balls and n bins problem," Lecture Note for Lecture 11, CS270 Combinatiorial Algorithms and Data Structures, University of Berkeley, 2003.

[45] M. Raab and A. Steger, "Balls into bins - a simple and tight analysis," in *RANDOM '98: Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science*, 1998, pp. 159–170.

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

[47] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Information Theory*, vol. 53, pp. 1009–1018, 2007.

[48] A. Saipulla, B. Liu, G. Xing, X. Fu, and J. Wang, "Barrier coverage with sensors of limited mobility," in *Proceedings of the eleventh ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '10, 2010, pp. 201–210.