

Rabie A. Ramadan

Computer Engineering Department
Cairo University
Cairo, Egypt

G. Ramamurthy

Communication Research Center
IIIT Hyderabad
Andhra Pradesh, India

Chris Reade

Kingston University
London, United Kingdom

M. Rinaudo

OPERA Department
Université Libre de Bruxelles
Brussels, Belgium

and

Department of Information Engineering
University of Parma
Parma, Italy

Rawya Yehia Rizk

Electrical Engineering Department
Port Said University
Port Said, Egypt

Lei Shu

Guangdong Petrochemical Equipment Fault
Diagnosis Key Laboratory
Guangdong University of Petrochemical
Technology
Maoming, China

M. Sujeethnanda

Communication Research Center
IIIT Hyderabad
Andhra Pradesh, India

Asis Kumar Tripathy

Department of Computer Science and
Engineering
National Institute of Technology Rourkela
Odisha, India

S. Van Roy

OPERA Department
Université Libre de Bruxelles
Brussels, Belgium

Emmanouel (Manos) Varvarigos

Department of Computer Engineering and
Informatics
University of Patras
Patras, Greece

Pu Wang

School of Electrical and Computer
Engineering
Georgia Institute of Technology
Atlanta, Georgia

Yu Wang

Department of Computer Science
University of North Carolina at Charlotte
Charlotte, North Carolina

Jean-Lien C. Wu

Department of Computer and
Communication Engineering
St. John's University
Taipei, Taiwan

Mohamed Younis

Department of Computer Science and
Electrical Engineering
University of Maryland, Baltimore County
Baltimore, Maryland

DATA COLLECTION

I

Chapter 1

Data Collection in Wireless Sensor Networks: A Theoretical Perspective

Yu Wang

University of North Carolina at Charlotte

Contents

1.1	Introduction.....	4
1.1.1	Network Model	5
1.1.2	Communication Model	6
1.1.3	Capacity and Delay in Data Collection	7
1.1.4	Related Works	8
1.2	Data Collection in Random Sensor Networks	9
1.2.1	Preliminaries.....	9
1.2.2	Data Collection with a Single Sink.....	10
1.2.3	Data Collection with Multiple Sinks	13
1.2.3.1	Regularly Deployed Multiple Sinks	13
1.2.3.2	Randomly Deployed Multiple Sinks.....	15
1.2.4	Data Collection under the Physical and Generalized Physical Models.....	18
1.2.4.1	Data Collection under the Physical Model	18
1.2.4.2	Data Collection under the Generalized Physical Model.....	20
1.3	Data Collection in Arbitrary Sensor Networks.....	23
1.3.1	Data Collection under the Protocol Model.....	23
1.3.1.1	Data Collection Tree: BFS Tree	23
1.3.1.2	Branch Scheduling Algorithm	24
1.3.1.3	Capacity Analysis	25
1.3.2	Data Collection under the General Graph Model.....	27
1.3.2.1	Upper Bound of Collection Capacity.....	28
1.3.2.2	Lower Bound of Collection Capacity.....	30

1.3.3	Data Collection under the Physical and Generalized Physical Models.....	35
1.3.3.1	Data Collection under the Physical Model	35
1.3.3.2	Data Collection under the Generalized Physical Model.....	36
1.4	Conclusion.....	36
	Acknowledgments	38
	References	38

1.1 Introduction

A wireless sensor network (WSN) consists of a set of sensor devices that are spread over a geographical area [1]. These sensors are able to perform processing as well as sensing and are additionally capable of communicating with each other. Due to the wide range of its potential applications, such as in the battlefield, emergency relief, environment monitoring, and so on, sensor networking has recently emerged as a premier research topic. For WSNs, the ultimate goal is often to collect sensing data from all sensors to certain sink nodes and then perform further analyses at these sink nodes. Thus, data collection is one of the most common services used in sensor network applications. Figure 1.1 shows an example of the data collection process in a WSN, in which a single sink node s at the center collects sensing values from every sensor using a collection tree.

The performance of data collection in sensor networks can be characterized by the rate at which sensing data can be collected and transmitted to sink nodes. In particular, theoretical measures that capture the possibilities and limitations of collection processing in sensor networks are the *delay* and *capacity* for many-to-one data collection. The delay of data collection is the time to transmit one single snapshot to sinks from its generation 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, a large delay rate is desired. When multiple snapshots from sensors are generated continuously, data transport can be pipelined in the sense that further snapshots may begin to transport before sinks receive the prior snapshot. The maximum data rate at the sinks to continuously receive snapshot data from sensors is defined as the capacity of data collection. Note that the capacity is always larger than or equal to the delay rate. Both delay rate and capacity reflect how fast the sinks can collect sensing 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 the performance of WSNs. In this chapter, we are particularly interested in how the delay rate and capacity of data

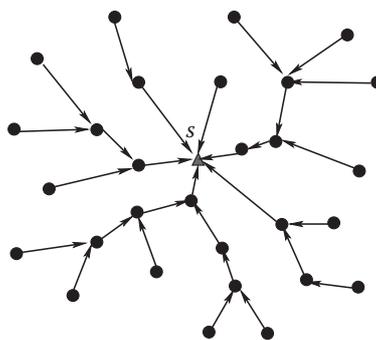


Figure 1.1 Data collection in a WSN with a single sink s .

collection vary in theory as the number of sensors increases. We will study some fundamental capacity problems arising from different types of data collection scenarios in WSNs. For each problem, we will introduce the asymptotic upper bound of transport capacity and present some efficient algorithms to achieve or approximate the upper bound.

1.1.1 Network Model

We focus on the theoretical capacity bound of data collection in WSNs. We consider a static sensor network, which includes n wireless sensor nodes $V = \{v_1, v_2, \dots, v_n\}$ and k sink nodes $S = \{s_1, s_2, \dots, s_k\}$ (when $k = 1$, we use s to denote the single sink). We assume that both sensor nodes and sink nodes are deployed in a two-dimensional square area. Two types of networks will be considered in this chapter: random networks and arbitrary networks (Figure 1.2 shows an example for each case). In random networks, sensor nodes are uniformly and randomly deployed in the area. Usually, under this model, the number of sensor nodes in the network is assumed to be very large. Such an assumption is useful to simplify the analysis and derive nice theoretical limits. Thus, the random network model has been widely used in the community for analyzing network performance. On the other hand, random networks may be invalid in many practical sensor applications in which the number of sensors is limited and the distribution of sensors is uneven inside the deployment region. In these cases, the arbitrary network model can be used. In arbitrary networks, sensors are deployed in any distribution and can form any network topology. Obviously, this model is more general and the random network model is just a special case of it.

Throughout this chapter, we assume each sensor node transmits at a fixed transmission power P . Then, a fixed transmission range r can be defined such that a node v_i can successfully receive the signal sent by node v_j only if $\|v_i - v_j\| \leq r$. Here, $\|v_i - v_j\|$ is the Euclidean distance between v_i and v_j . We can 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. This graph model is called a disk graph model. We assume the communication graph G is connected.

At regular time intervals, each sensor node measures the field value at its position and transmits the value to one of the sink nodes. We assume that the channel bandwidth for all wireless links is W bits per second. We also assume that all packets have the unit size of b bits. Time is divided into slots with $t = b/W$ seconds. Accordingly, only one packet can be transmitted in each time slot between two neighboring nodes. Time division multiple access (TDMA) scheduling is used at the media access control (MAC) layer.

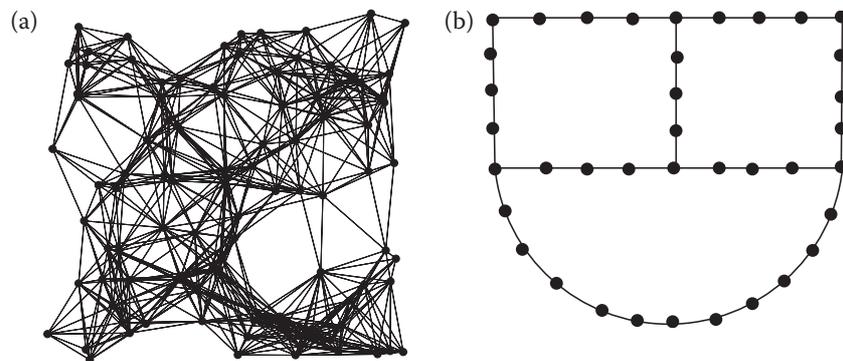


Figure 1.2 A random network (a) versus an arbitrary network (b).

1.1.2 Communication Model

Due to spatial separation, several sensors can successfully transmit at the same time if these transmissions do not cause any destructive wireless interference. There are three widely used communication models [2,3] to capture such interference constraints: the protocol model, the physical model, and the generalized physical model.

In the protocol model (also called protocol interference model), all nodes are assumed to have a uniform interference range R . When node v_i transmits to node v_j , node v_j can receive the signal successfully if no other node within a distance R of v_j is transmitting simultaneously. Figure 1.3 illustrates an example of the protocol model in which the transmission between v_k and v_q will cause interference at v_j . Usually, R/r is assumed as a constant α larger than 1. The protocol model is the simplest communication model considering the interference among nodes, and has been widely used in the literature. However, it is sometimes too simple to capture the complexity of interference.

In the physical model (also called 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 signal-to-noise ratio (SINR)

$$\frac{P \cdot l(v_i, v_j)}{N_0 + \sum_{k \in I} P \cdot l(v_k, v_j)} \geq \eta.$$

Here, $l(v_i, v_j)$ is the transmission loss between v_i and v_j , $N_0 > 0$ is the background Gaussian noise, I is the set of actively transmitting nodes when node v_i is transmitting, and P is the fixed transmission power. In this chapter, we consider the attenuation function $l(v_i, v_j) = \min\{1, \|v_i - v_j\|^{-\beta}\}$ where $\beta > 2$ is the path loss exponent and $\|v_i - v_j\|$ is the Euclidean distance between v_i and v_j . Hereafter, we assume that all P , N_0 , β , and η are fixed constants. Notice that the values of P , N_0 , η , and transmission range r should satisfy $\frac{P \cdot r^{-\beta}}{N_0} \geq \eta$. Thus, $r \leq \left(\frac{P}{N_0 \cdot \eta}\right)^{1/\beta}$.

For both the protocol model and the physical model, as long as the value of a given conditional expression (such as transmission distance or SINR value) reaches some threshold, the sender can

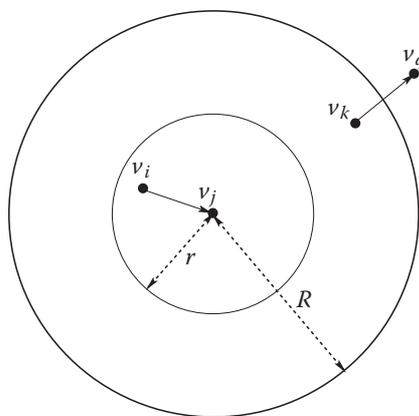


Figure 1.3 In the protocol model, node v_j can receive the signal successfully from v_i if v_i is within v_j 's transmission range r , and no other transmitting node within a distance R of v_j . In this example, v_k 's transmission will cause interference at node v_j .

send data successfully to a receiver at a specific constant rate W due to the fixed rate channel model. However, the fixed rate channel model may not capture the feature of wireless communication well. As a result, a more realistic model, the generalized physical model (also called Gaussian channel model) is introduced. Such a model 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

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

This model assigns a more realistic transmission rate at a larger distance than the fixed rate channel models (protocol model and physical model). In this chapter, we will cover the data collection in WSNs under all these communication models. Without specific notification, we use the protocol model as the default model in our analysis.

1.1.3 Capacity and Delay in Data Collection

We now formally define delay and capacity of data collection in WSNs. Recall that each sensor, at regular time intervals, generates an independent field value with b bits and wants to transport it to one of the sink nodes. The union of all sensing values from n sensors at a particular sampling time is called a *snapshot* of the sensing data. The task of data collection is to collect these snapshots from all sensors to sinks as quickly as possible.

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 nb 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 data collection under pipelining (sometimes called continuous 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 the next snapshot. ■

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 snapshot (i.e., 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 chapter, we analyze the delay rate and capacity for data collection in both random and arbitrary WSNs under various communication models. Notice that in our definitions, we require data from every sensor to reach the sink in the same rate; thus, fairness among all sensors is guaranteed.

1.1.4 Related Works

Gupta and Kumar [4] initiated research on the capacity of wireless ad hoc networks by studying fundamental capacity limits in their seminal article under both protocol and physical models. The following articles studied capacity under different communication scenarios in wireless networks: unicast [5,6], multicast [7–9], and broadcast [10,11] capacities. In this chapter, we focus on the capacity of data collection, which is an all-to-one communication scenario different from the unicast, multicast, and broadcast capacities.

The capacity of data collection in random WSNs has been previously studied [12–25]. In previous work by Duarte-Melo et al. [12,13], they first studied the many-to-one transport capacity in random sensor networks under the protocol model and gave the results of overall capacity of data collection as $\Theta(W)$. They also showed that compressing data is inefficient in improving capacity when the density of the sensor network increases to infinity [13]. El Gamal [14] studied data collection capacity subject to a total average transmitting power constraint. The assumption that every node can only receive a packet from one source node at a time was relaxed, and it was shown that the capacity of random networks scales as $\Theta((\log n) W)$ when n goes to infinity and the total average power remains fixed. Their methods used antenna sharing and channel coding. Barton and Zheng [15,16] also investigated data collection capacity under more complex physical models [a noncooperative SINR model and a cooperative time reversal (CTR) communication model]. They first demonstrated that $\Theta((\log n) W)$ is optimal and achievable by using CTR for a regular grid network [15], then showed that the capacities of $\Theta((\log n) W)$ 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 [16]. Liu et al. [17] 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. [18–21] also studied data collection methods in random WSNs under different communication models, such as dual-radio multichannel networks [18], asynchronous WSNs [20], or probabilistic network models [21].

This chapter mainly covers the recent results from the author and his colleagues [22–27] on data collection capacity in both random and arbitrary WSNs. However, readers are encouraged to further read the references listed above to get the complete picture of capacity research in wireless networks.

The rest of the chapter is organized as follows. We first discuss the study on data collection capacity of random WSNs under various communication models and network scenarios in Section 1.2. We then consider the data collection capacity for an arbitrary WSN under different models. Finally, we conclude this chapter with a short summary in Section 1.4.

1.2 Data Collection in Random Sensor Networks

In this section, we focus on data collection in a large-scale random WSN and study how fast it can be performed under the existence of interference among sensors. We consider both cases with a single sink or multiple sinks under the protocol model, and also discuss the cases under physical models.

1.2.1 Preliminaries

We consider a random WSN in which n sensor nodes are randomly and uniformly deployed in a square area with side length l . Two types of random network models [28] can be defined: random dense network and random extended network. In the random dense network, sensors are uniformly deployed in a unit square area ($l = 1$). Thus, its node density is n . In the random extended network, sensors are uniformly deployed in a square region with $l = \sqrt{n}$, thus its density is 1. In most of this chapter (except for Section 1.2.4), we use the random dense network.

We now introduce a classic grid-partition method that is essential for the proposed data collection methods and theoretical analysis. As shown in Figure 1.4, the network (e.g., the unit square) is divided into a^2 micro cells of the size $d \times d$. Here, $a = 1/d$. We assign each cell a coordinate (i, j) , where i and j are between 1 and a , indicating its position at the j th row and i th column.

The following lemma gives a guidance of the cell size.

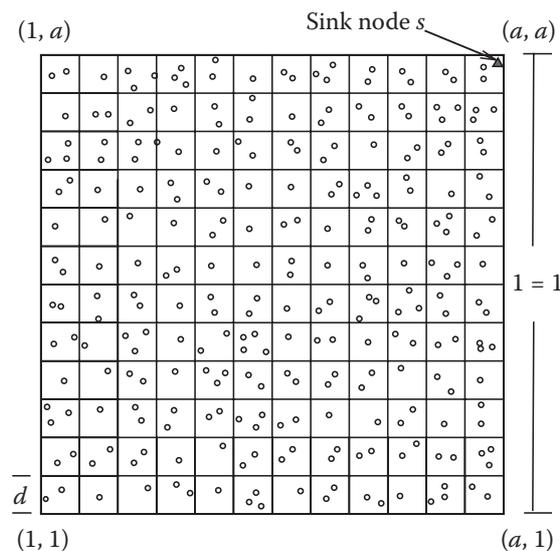


Figure 1.4 Grid partition of the WSN: a^2 cells with a cell size of $d \times d$.

Lemma 1

Given n random nodes in a unit square [29], 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 a probability of 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 a high probability (the probability converges to one as $n \rightarrow \infty$).

We can also derive the upper bound of the number of nodes inside a single cell.

Lemma 2

Given n random nodes in a unit square [22,23], 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 a probability of at least $1 - \frac{3\log n}{n}$. ■

The proof is straightforward from the following lemma when the number of balls $\gamma = n$ and the number of bins $\delta = a^2 = \frac{n}{3\log n}$. Lemma 2 indicates the number of nodes inside any cell is bounded from above by $O(\log n)$ with high probability.

Lemma 3

Randomly putting γ balls into δ bins [30], with a probability of at least $1 - \frac{1}{\delta}$, the maximum number of balls in any bin is $O\left(\frac{\gamma}{\delta} + \log \delta\right)$. ■

To make the whole network connected, the transmission range r needs to be equal to 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 $r = \sqrt{5}d = \sqrt{15\frac{\log n}{n}}$.

1.2.2 Data Collection with a Single Sink

In this subsection, we consider the simplest situation: data collection under the protocol model in a sensor network in which a single sink s is located in the upper right corner of the deployment region (i.e., cell (a,a) as shown in Figure 1.5). Notice that if the sink is located at the center of the region or anywhere in the region, it only adds a constant in the analysis. We first construct a data

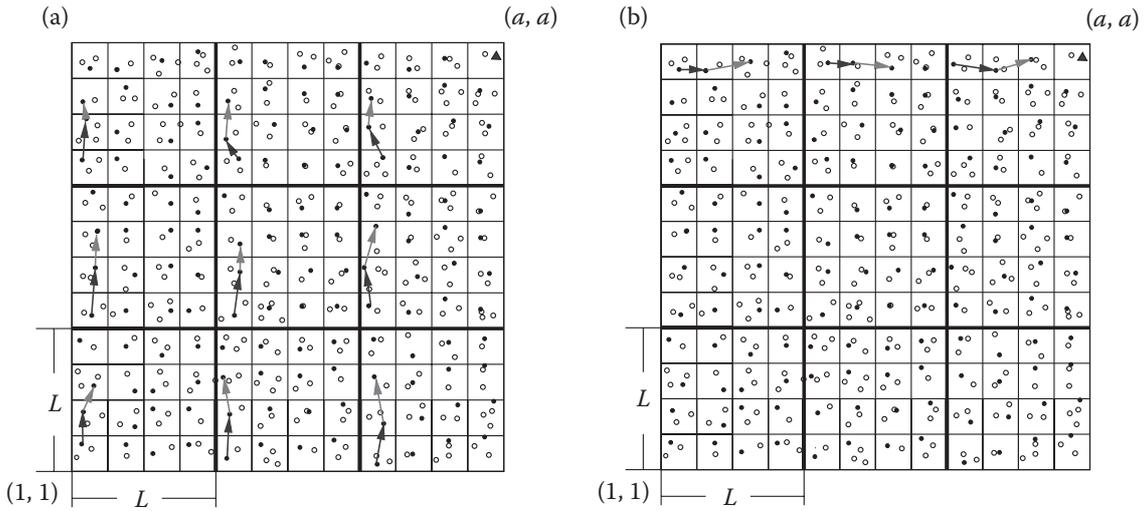


Figure 1.5 Our collection method: (a) every node sends its data to the upper cell in Phase I; (b) then each node in the top row sends its data to the cell to its right in Phase II.

collection scheme whose delay and delay rate are $O(nt)$ and $\Omega(W)$, respectively, and then prove that these values are order-optimal.

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 a th row) as shown in Figure 1.5a, and in the second phase (phase II), all data is sent via cells in the a th row to the sink as shown in Figure 1.5b. 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 needs one time slot t to send one packet to its neighbor in the next cell. However, due to wireless interference, when node v_i transmits a packet to v_j , nodes within R distance from v_i cannot transmit any packets in the same time slot. Let $L = \left(\frac{R}{d} + 2 \right)$. Thus, every $L \times L$ cell (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 1.5, bold lines show interference blocks. Remember that $\frac{R}{r} = \alpha$ and $\frac{r}{d} = \sqrt{5}$, so $\frac{R}{d}$ and L are also constants, and a packet in the lowest row (i.e., cell $(0, k)$) has to walk a cells to reach nodes in the highest cell in the rectangle. Hence,

$$T_1 \leq L \times L \times t \times O(\log n) \times a = O(t \log n) a = O(t \log n) \sqrt{\frac{n}{3 \log n}} = O\left(t \sqrt{n \log n}\right).$$

In the beginning of phase II, all data are already at the cells of the top row. The sink s lies in the same row with these cells. We now estimate the time T_2 needed for sending all data to s . Each cell in the top row has at most $a \times O(\log n)$ nodes' data and the interference block is now $1 \times L$. Similarly, we can get

$$T_2 \leq L \times t \times a \times O(\log n) \times a = O(t \log n) a^2 = \frac{n}{3 \log n} O(t \log n) = O(nt).$$

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

$$\Gamma = \frac{nb}{\Delta} = \Omega\left(\frac{nb}{nt}\right) = \Omega(W).$$

It has been proved that the upper bound of delay rate or capacity of data collection is W [12,13]. It is obvious that the sink cannot receive at a rate faster than W because W is the fixed transmission rate of an 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 was $\Theta(W/n)$, 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 that the upper bound of capacity is still W . Because our scheme above 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 $a \times b$ 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 $a \times b$ data in phase I is

$$T_1' \leq L \times L \times t \times O(\log n) = O(t \log n)$$

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

$$T_2' \leq \max\{nt, L \times t \times a\} = O(nt).$$

Therefore, the total time for sink to receive $a \times b$ data is $T_1' + T_2' = O(nt)$. Thus, the capacity of our method with pipelining is still

$$C = \frac{a \cdot b}{T_1' + T_2'} = \Omega(W).$$

This also meets the upper bound W in order.

In summary, we have the following theorem:

Theorem 1

Under the protocol model [22,23], the delay rate Γ and the capacity C of data collection in random sensor networks with a single sink are both $\Theta(W)$. ■

Notice that the scheduling algorithm presented here is order-optimal but the constant behind the big Θ could be large. There are different methods (such as those used by Ji et al. [19]) that could further improve the achieved capacity by constant times.

1.2.3 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 subtasks (i.e., collections in subareas) and each subtask can 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. Because the delay rate and the capacity are always of the same order in both cases, we will not distinguish between them and instead use only the term of capacity. Two scenarios are considered in the following subsections: sinks are regularly deployed on a grid or randomly deployed in the field.

1.2.3.1 Regularly Deployed Multiple Sinks

When sinks are displayed regularly on a $\sqrt{k} \times \sqrt{k}$ grid, the capacity of collection depends on the number of k sinks. Here, we divide the unit area into k subareas, 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 $k < \frac{n}{15(\alpha+1)^2 \log n}$, $k < \frac{1}{(R+r)^2}$ because $R = \alpha r$ and $r = \sqrt{15 \frac{\log n}{n}}$. Thus, each subarea assigned to a sink is larger than or equal to $(R+r)^2$. Therefore, we can perform the data collection in each subarea without interfering with the neighboring subareas. Because we have k subareas, the total delay rate and the total capacity of the whole area is 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 subarea is smaller than $(R+r)^2$, which indicates that there will be interference between neighboring subareas.

Therefore, the total delay rate or capacity is bounded by $\frac{1}{(R+r)^2} \cdot \Theta(W) = \Theta\left(\frac{n}{\log n} W\right)$ 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 subareas with the size of $\frac{1}{\sqrt{k}} \times \frac{1}{\sqrt{k}}$ and every sink performs the collection method to collect their subareas. When $k \geq \frac{n}{15(\alpha+1)^2 \log n}$, we partition the field into $\frac{1}{(R+r)^2}$ subareas with a size of $(R+r) \times (R+r)$ as shown in Figure 1.6. 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 sinks 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 the order of the theoretical bound. Figure 1.6 shows

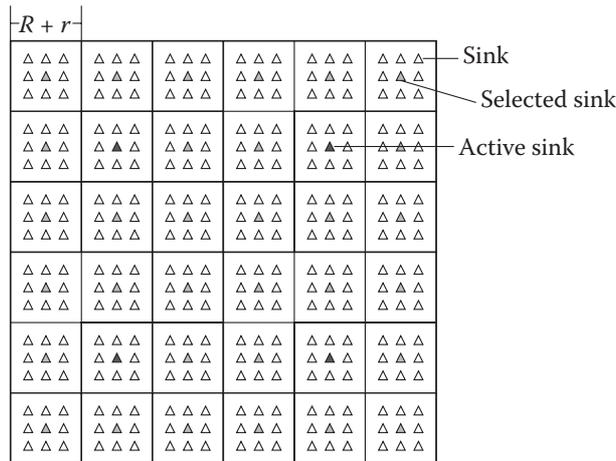


Figure 1.6 When k is large, we partition the field into $\frac{1}{(R+r)^2}$ subareas. Each subarea selects one sink as its selected sink (shown as a gray triangle). Only one selected sink inside nine subareas is active for data collection (shown as a black triangle). It will collect data from the surrounding nine subareas using the single sink method. Notice that the adjacent nine subareas will not interfere with each other when applying the collection method.

a possible scheduling in which only one of nine selected sinks collects data from its surrounding blocks, and thus, we have our second theorem.

Theorem 2

Under the protocol model [22,23], the delay rate Γ and the capacity C of data collection in random sensor networks with k regularly deployed sinks are

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

■

Because when $k = \Theta\left(\frac{n}{\log n}\right)$, the capacity (or delay rate) of two cases are all equal to $\Theta(kW) = \Theta\left(\frac{n}{\log n}W\right)$. Therefore, the above equations can also be written as follows:

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

1.2.3.2 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 $\Theta(kW)$ because 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 because some of them may finish their collection earlier and have no data to collect.

We first derive the upper bound of data collection capacity. Because the interference range is $R = \alpha r = \alpha \cdot \sqrt{15 \frac{\log n}{n}}$, we partition the whole area into interference blocks with a 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 $k = o\left(\frac{n}{\log n}\right)$. For this case, the capacity of data collection is bounded by $\Theta(kW)$ from above because 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 $k = \Theta\left(\frac{n}{\log n}\right)$. We calculate the probability that an arbitrary interference block has at least one sink.

$$\begin{aligned} \Pr(\text{an interference block has at least one sink}) &= 1 - \left(1 - \frac{1}{B}\right)^k = 1 - \left[1 - \frac{1}{\Theta\left(\frac{n}{\log n}\right)}\right]^k \\ &= 1 - \left[1 - \frac{1}{\Theta\left(\frac{n}{\log n}\right)}\right]^{\Theta\left(\frac{n}{\log n}\right)}. \end{aligned}$$

When $n \rightarrow \infty$, this probability is equal 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 $E[X] = \Pr \times B = \left(1 - \frac{1}{e}\right) \frac{n}{60\alpha^2 \log n}$ and $\sigma^2 = \Pr \times (1 - \Pr) \times B = \frac{1}{e} \left(1 - \frac{1}{e}\right) \frac{n}{60\alpha^2 \log n}$. Based on the Chebyshev inequality, we have the following:

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

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

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

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

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

$$\begin{aligned} \Pr(\text{an interference block has at least one sink}) &= 1 - \left[1 - \frac{1}{\Theta\left(\frac{n}{\log n}\right)}\right]^k \\ &= 1 - \left[1 - \frac{1}{\Theta\left(\frac{n}{\log n}\right)}\right]^{\Theta\left(\frac{n}{\log n}\right) \cdot \frac{k}{\Theta\left(\frac{n}{\log n}\right)}} \\ &= 1 - \left[1 - \frac{1}{\Theta\left(\frac{n}{\log n}\right)}\right]^{\Theta\left(\frac{n}{\log n}\right) \frac{\Omega\left(\frac{n}{\log n}\right)}{\Theta\left(\frac{n}{\log n}\right)}} \end{aligned}$$

When $n \rightarrow \infty$, 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 $\Theta\left(\frac{n}{\log n} W\right)$ from above.

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

When $k = O\left(\frac{n}{\log n}\right)$, we first partition the network into interference blocks with the size $\sqrt{3\frac{\log k}{k}} \times \sqrt{3\frac{\log k}{k}}$. From Lemma 1, we know that each of the blocks is occupied by at least one sink with a high probability. Because $k = O\left(\frac{n}{\log n}\right)$, 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 1.2.3.1) to schedule a subset of selected sinks to collect data from its surrounding area. The capacity achieved is $\Theta\left(\frac{k}{\log k}W\right)$ because the number of selected sinks is $\Theta\left(\frac{k}{\log k}\right)$. Notice that there is a gap between this lower bound and the upper bound $\Theta(kW)$. This is due to the possibly uneven distribution of k sinks in this case, thus each sink may not have the same number of sensors (or areas) to perform the collection to achieve $\Theta(kW)$ capacity in total.

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

Theorem 3

Under the protocol model [22,23], the delay rate Γ and the capacity C of data collection in random sensor networks with k randomly deployed sinks are

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

■

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., $k = O\left(\frac{n}{\log n}\right)$), it is beneficial to add more sinks. However, when the capacity is constrained by the interference among sinks (i.e., $k = \omega\left(\frac{n}{\log n}\right)$), adding more sinks

has no substantial capacity improvement. Similar observations were made by Liu et al. [17] for many-to-many capacity.

1.2.4 Data Collection under the Physical and Generalized Physical Models

Thus far, we only consider the protocol model, which is ideal but unrealistic in WSNs, in which the interference is modeled as a localized phenomenon. However, a receiver can be interfered with by a group of actively transmitting sensors even if its location is extremely far away from the group of sensors. Thus, we now consider more accurate models to reflect the influence of interference: the physical model and the generalized physical model. Please refer to Section 1.1.2 for their definitions.

Again, we consider a random sensor network with n sensor nodes and a single sink. We now use the random extended network model [28], in which all sensor nodes are uniformly deployed in a square region with side length $l = \sqrt{n}$, by use of Poisson distribution with density 1. The grid partition method we introduced in Section 1.2.1 is the same except for the size length of every cell and the transmission range of each sensor are \sqrt{n} times larger than those in Section 1.2.1.

1.2.4.1 Data Collection under the Physical Model

For the case of data collection under the physical model, our collection scheme and analysis are almost the same with the one under the protocol model (Section 1.2.2). The only difference is that a new size of interference block is used.

We first divide the field into big blocks with size $L \times L$ as shown in Figure 1.7. We call these blocks *interference blocks* and L *interference distance*. Thus, the number of interference blocks is $\frac{l^2}{L^2}$. We label each block with (i,j) where i and j are the indexes of the block as in Figure 1.7. 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. To avoid interference from senders in other interference blocks, we need an interference distance L that is larger than a certain value.

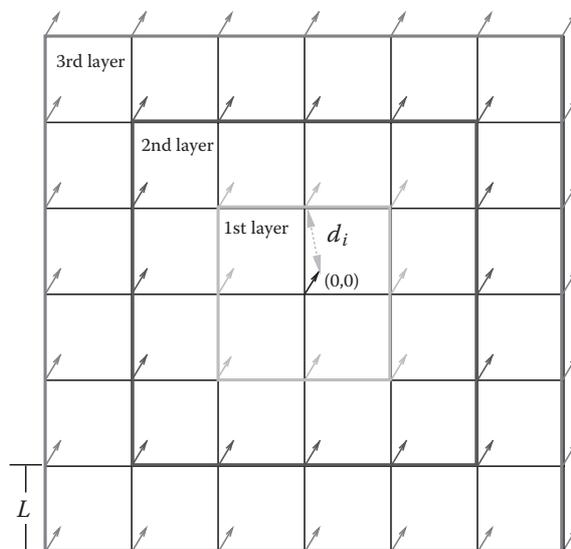


Figure 1.7 Grid partition of interference blocks with size of $L \times L$ and simultaneous transmissions around the center block $(0,0)$ by layers.

Next, we derive the lower bound of interference distance such that all simultaneous transmissions, as shown in Figure 1.7, 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, because it has the minimum SINR among all receivers. Similar to the technique used by Franceschetti et al. [31], we now label all simultaneous transmissions by layers from block (0,0), as shown in Figure 1.7.

Based on the physical interference model, its SINR is at least

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

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

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

Notice that $d_i \geq iL - 2d$ 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 a distance of at least $2L - 2d$, and so on. Thus,

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

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

$$8\phi(L - 2d)^{-\beta} \leq \frac{r^{-\beta}}{\eta} - \frac{N_0}{P},$$

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

$$L \geq \left[\frac{1}{8\phi} \cdot \left(\frac{r^{-\beta}}{\eta} - \frac{N_0}{P} \right) \right]^{\frac{1}{\beta}} + 2d.$$

Remember that $r \leq \left(\frac{P}{N_0 \cdot \eta} \right)^{\frac{1}{\beta}}$, this makes sure we can find such suitable L . We can further

select $L = \left[\frac{1}{8\phi} \cdot \left(\frac{r^{-\beta}}{\eta} - \frac{N_0}{P} \right) \right]^{\frac{1}{\beta}} + 2d$. Because $r = \sqrt{5}d$,