Data Collection Capacity of Random-Deployed Wireless Sensor Networks

Siyuan Chen^{*} Yu Wang^{*} Xiang-Yang Li[†] Xinghua Shi[‡]

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

[†]Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois, USA

[‡]Department of Computer Science, University of Chicago, Chicago, Illinois, USA

Abstract—Data collection is one of the most important functions provided by wireless sensor networks. In this paper, we study the theoretical limitations of data collection and data aggregation in terms of delay and capacity for a wireless sensor network where n sensors are randomly deployed. We consider two different communication scenarios (with or without aggregation) under physical interference model. For each scenario, we first propose a new collection method and analyze its performance in terms of delay and capacity, then theoretically prove that our method can achieve the optimal order. Particularly, the capacity of data collection is in order of $\Theta(W)$ where W is the fixed data-rate on individual links. If each sensor can aggregate its receiving packets into a single packet to send, the capacity of data collection increases to $\Theta(\frac{n}{\log n}W)$.

I. INTRODUCTION

For wireless sensor networks, often the ultimate goal is to collect the sensing data from all sensors to a sink node and then perform further analysis at the sink node. Thus, data collection is one of the most common services used in sensor network applications. In this paper, we study some fundamental capacity problems arising from different types of data collection scenarios in wireless sensor networks. For each problem, we will derive the asymptotic upper bound of transport capacity and present efficient algorithms to achieve such upper bound with certain constant factor.

We consider a dense sensor network where n sensors are randomly deployed in a finite geographical region. Each sensor measures independent field values at regular time intervals and send these values to the sink. The union of all sensing values from n sensors at particular time is called *snapshot*. The task of data collection is to deliver these snapshots to the sinks. Due to spatial separation, several sensors can successfully transmit at the same time if these transmissions do not cause any destructive wireless interferences. We assume that a successful transmission over a link has a fixed data-rate W bit/second.

The performance of data collection in sensor networks can be characterized by the rate at which sensing data can be collected and transmitted to the sink node. In particular, the theoretical measures that capture the possibilities and limitations of collection processing are delay and capacity for the many-to-one data collection. The delay of data collection is the time to transmit one single snapshot to the sink 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. When multiple snapshots from sensors are generated continuously, data transport can be pipelined in the sense that further snapshot may begin to transport before the sink receiving the prior snapshot. The maximum data rate at the sink to continuously receive the snapshot data from sensors is defined as the *capacity* of data collection. Both delay rate and capacity reflect that how fast the sink 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 performance of wireless sensor networks. In this paper, we are particularly interested in how delay rate and capacity of data collection vary as the number of sensors n increases.

Gupta and Kumar initiated the research on capacity of random wireless networks in the seminal paper [1]. A number of following papers studied capacity under different communication scenarios [2]-[4]. Recently, capacity limits of data collection in random wireless sensor networks have been studied in the literature [5]-[11]. In [5], [6], 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 protocol interference model. Chen et al. [7] studied capacity of data collection with multiple sinks in random networks under protocol interference model. Liu et al. [8] recently studied the capacity of a general someto-some communication paradigm under protocol interference model in random networks where there are multiple randomly selected sources and destinations. All these works are based on protocol interference model. El Gamal [9] 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. Barton and Rong [10], [11] then investigated data collection capacity under more complex physical layer models (non-cooperative SINR model and cooperative time reversal communication model) where the data rate of indi-

The work of Y. Wang is supported in part by the US National Science Foundation (NSF) under Grant No. CNS-0721666 and No. CNS-0915331. The work of X.-Y. Li is partially supported by the US NSF under Grant No. CNS-0832120, the National Natural Science Foundation of China under Grant No. 60828003, No.60773042 and No.60803126, the Natural Science Foundation of Zhejiang Province under Grant No. Z1080979, the National Basic Research Program of China (973 Program) under Grant No. 2006CB30300, the National High Technology Research and Development Program of China (863 Program) under Grant No. 2007AA01Z180, the Hong Kong RGC under Grant HKUST 6169/07, HKBU 2104/06E, and CERG under Grant PolyU-5232/07E. X.-Y. Li is also with Institute of Computer Application Technology, Hangzhou Dianzi University, Hangzhou, China.

vidual link is not fixed as a constant W but depended on the level of interference. In this paper, we study *capacity of data collection under physical interference model for random sensor networks with a single sink.*

The major contributions of this paper are as follows. (1) For sensor networks without data aggregation, we propose a new data collection method whose delay rate and capacity under physical interference model are both $\Theta(W)$ which match the theoretical upper bounds in order. (2) By using data aggregation [12], [13] where sensors can cooperate to aggregate information towards the sink, the communication overhead is reduced and the capacity is increased. 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. For both cases, our proposed collection/aggregation methods achieve the optimal orders (i.e., within a constant factor of upper bounds).

II. PRELIMINARIES

A. Network Model

We consider a sensor network which includes n wireless sensor nodes $V = \{v_1, v_1, \dots, v_n\}$ and a single sink node 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. At regular time intervals, each sensor node measures the field value at its position and transmits the value to the sink node. We adopt a fixed data-rate channel model where each wireless node can transmit at W bits/second over a common wireless channel. Under such channel model, we assume that every node has a fixed transmission power. Then a fixed transmission range r can be defined such that a node v_i can successfully receive the signal sent by node v_i only if the distance between them is less or equal to r. We also assume that all packets have unit size b bits. The time is slotted into time slots with t = b/W seconds. Thus, only one packet can be transmitted in a time slot between two neighboring nodes.

As in the literature, we consider the interference modeled by *physical interference model* in our analysis. 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

$$\frac{P_i \cdot l(||v_i - v_j||)}{B \cdot N_0 + \sum_{k \in I} P_k \cdot l(||v_k - v_j||)} \ge \eta.$$

Here $||v_i - v_j||$ is the Euclidean distance between v_i and v_j , l(x) is the transmission loss during a path of length x, 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, and P_k is node v_k 's transmission power. In this paper, we consider the attenuation function $l(x) = min\{1, x^{-\beta}\}$ where $\beta > 2$ is the path loss exponent. Hereafter, 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 $\frac{P \cdot r^{-\beta}}{BN_0} \ge \eta$. Thus, $r \le (\frac{P}{B \cdot N_0 \cdot \eta})^{1/\beta}$.



Fig. 1. Grid partition of the sensor network: a^2 cells with cell size of $d \times d$.

B. Capacity and Delay

We now formally define delay and capacity of data collection. Recall that each sensor at regular time intervals generates a field value with b bits and wants to transport it to the sink. We call the union of all values from all n sensors at particular sampling time a *snapshot* of the sensing data. The goal of data collection is to collect these snapshots from all sensors as quick 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 sink has 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 .

On the other hand, the data transport can be pipelined in the sense that further snapshots may begin to transport before the sink receiving the prior snapshots. Therefore, we need to define a new data rate of data collection under pipelining.

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

Thus, the time used by the sink to successfully receive a snapshot is $T = U \times t$. Due to pipelining, $T \leq \Delta$. 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 sink.

Thus, the capacity \hat{C} is the maximum data rate at the sink to continuously receive the snapshot data from sensors. Clearly, C is at least as large as the delay rate Γ , and usually substantially larger. In this paper, we analyze both delay rate and capacity for data collection in random sensor networks.

III. DATA COLLECTION WITHOUT AGGREGATION

In this section, we consider data collection without aggregation where each data packet generated from a sensor needs to individually reach the sink s. We first construct a data collection scheme whose delay rate is $\Omega(W)$, and then prove that it is order-optimal. Our data collection scheme is based on the following grid partition method.

A. Our Partition Method

We first introduce a grid partition method which is essential for our data collection methods and their theoretical analysis. As shown in Fig. 1, 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 *j*th row and *i*th column.

The following lemma gives a guidance of the cell size.

Lemma 1: [14] Given n random nodes in a $\sqrt{n} \times \sqrt{n}$ square, dividing the square into micro cells of the size $\sqrt{3\log n} \times \sqrt{3\log n}$, every micro cell is occupied with probability at least $1 - \frac{1}{n^2}$.

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 (the probability converges to one as $n \longrightarrow \infty$).

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

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

Proof: The proof is straightforward from results of the balls into bins problem [15] and thus ignored here.

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 $r = \sqrt{5}d = \sqrt{15 \log n}$.

B. Our Data Collection Scheme

As shown in Fig. 1, we can consider data collection of nodes from four different directions (i.e., quadrants) to the sink s located in cell (p,q). For the purpose of analysis, we only concentrate on the direction which has the largest number of sensors, e.g., the shaded rectangle in Fig. 1, since the sink can perform collection on each direction in turn and it only adds a constant 4 in the analysis. Furthermore, here we only consider the worst case where p = q = m, that is, 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 Fig. 2. 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 Fig. 2. 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 interference distance L larger than certain value.

Next, we derive the lower bound of interference distance such that all simultaneous transmissions as shown in Fig. 2 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.



Fig. 2. Grid partition of interference blocks with size of $L \times L$.

Based on physical interference model, its SINR is

$$\frac{P \cdot l(r)}{B \cdot N_0 + \sum_{\text{all blocks } (i,j) \text{ except } (0,0)} P \cdot l(d_{i,j})}$$

Here, $d_{i,j}$ is the distance from the sender in block (i, j) to the receiver in block (0,0). By a simple geometric calculation, $d_{i,j} = \sqrt{(iL - d)^2 + (jL - 2d)^2}$. Remember $r = \sqrt{15 \log n}$ thus both r and $d_{i,j}$ are larger than 1. Therefore, we need to derive L such that

$$\frac{P \cdot r^{-\beta}}{B \cdot N_0 + \sum_{(i,j)/(0,0)} P \cdot ((iL-d)^2 + (jL-2d)^2)^{-\beta/2}} \ge \eta.$$

In other words, we need

$$\sum_{(i,j)/(0,0)} ((iL-d)^2 + (jL-2d)^2)^{-\beta/2} \le \frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}.$$

Notice that:

=

$$\sum_{\substack{(i,j)/(0,0)\\ \leq}} ((iL-d)^2 + (jL-2d)^2)^{-\beta/2}$$

$$\leq \sum_{\substack{(i,j)/(0,0)\\ =}} ((iL-2id)^2 + (jL-2jd)^2)^{-\beta/2}$$

$$= (L-2d)^{-\beta} \sum_{\substack{(i,j)/(0,0)\\ (i^2+j^2)^{-\beta/2}}.$$

Instead of considering interference from all blocks (i, j)around (0,0), we relax it to 4 times the interference from blocks in the right up direction.

$$\sum_{\substack{(i,j)/(0,0)}} ((iL-d)^2 + (jL-2d)^2)^{-\beta/2}$$

$$\leq (L-2d)^{-\beta} \sum_{\substack{(i,j)/(0,0)}} (i^2 + j^2)^{-\beta/2}$$

$$\leq 4(L-2d)^{-\beta} \sum_{\substack{(i\geq 0,j\geq 0)/(0,0)}} (i^2 + j^2)^{-\beta/2}$$

$$= 4(L-2d)^{-\beta} (\sum_{\substack{(i\geq 1,j\geq 1)}} (i^2 + j^2)^{-\beta/2} + 2\sum_{\substack{(i\geq 1,j=0)}} i^{-\beta}).$$



Fig. 3. 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.

Notice that for i > 1, j > 1 and $\beta > 2$, $(\frac{1}{i^2 + j^2})^{\beta/2} \le (\frac{1}{4}(\frac{1}{i^2} + \frac{1}{j^2}))^{\beta/2} \le \frac{1}{2^{\beta}}(\frac{1}{i^{\beta}} + \frac{1}{j^{\beta}})$. Then we have

$$\begin{split} &\sum_{(i,j)/(0,0)} ((iL-d)^2 + (jL-2d)^2)^{-\beta/2} \\ &\leq 4(L-2d)^{-\beta} (\sum_{(i\geq 1,j\geq 1)} (i^2+j^2)^{-\beta/2} + 2\sum_{(i\geq 1,j=0)} i^{-\beta}) \\ &\leq 4(L-2d)^{-\beta} (\sum_{(i\geq 1,j\geq 1)} \frac{1}{2^{\beta}} (\frac{1}{i^{\beta}} + \frac{1}{j^{\beta}}) + 2\sum_{(i\geq 1,j=0)} i^{-\beta}) \\ &= 4(L-2d)^{-\beta} (\frac{1}{2^{\beta}} \sum_{i\geq 1} \frac{1}{i^{\beta}} + \frac{1}{2^{\beta}} \sum_{j\geq 1} \frac{1}{j^{\beta}} + 2\sum_{i\geq 1} i^{-\beta}) \\ &= 4(L-2d)^{-\beta} (\frac{1}{2^{\beta-1}} + 2) \sum_{i\geq 1} i^{-\beta} \\ &\leq 4(L-2d)^{-\beta} (\frac{1}{2^{\beta-1}} + 2) \sum_{i\geq 1} i^{-2} \\ &\leq \frac{2\pi}{3} (L-2d)^{-\beta} (\frac{1}{2^{\beta-1}} + 2) \text{ (since } \sum_{i=1}^{\infty} i^{-2} = \frac{\pi}{6}). \end{split}$$

Therefore, if $\frac{2\pi}{3}(L-2d)^{-\beta}(\frac{1}{2^{\beta-1}}+2) \leq \frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}$, we can sure that the SINR at the receiver in the center is at least η . This can be satisfied by setting

$$L \ge \left(\frac{3 \cdot 2^{(\beta-2)}}{\pi(1+2^{\beta})} \cdot \left(\frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}\right)\right)^{-\frac{1}{\beta}} + 2d$$

Remember $r \leq (\frac{P}{B \cdot N_0 \cdot \eta})^{1/\beta}$, this makes sure we can find such suitable *L*. We can further select $L = (\frac{3 \cdot 2^{(\beta-2)}}{\pi(1+2^{\beta})} \cdot (\frac{r^{-\beta}}{\eta} - \frac{BN_0}{P}))^{-\frac{1}{\beta}} + 2d$. Since $r = \sqrt{5}d$,

$$\frac{L}{d} = \left(\frac{3 \cdot 2^{(\beta-2)}}{\pi(1+2^{\beta})} \cdot \left(\frac{(\sqrt{5}d)^{-\beta}}{\eta d^{-\beta}} - \frac{BN_0}{Pd^{-\beta}}\right)\right)^{-\frac{1}{\beta}} + 2 \\
= \left(\frac{3 \cdot 2^{(\beta-2)}}{\pi(1+2^{\beta})} \cdot \left(\frac{5^{-\beta/2}}{\eta} - \frac{BN_0d^{\beta}}{P}\right)\right)^{-\frac{1}{\beta}} + 2.$$

When $n \to \infty$, this ratio goes to a constant, denoted by α . After having interference blocks, we can now present our collection algorithm which is quite simple and straightforward. It has two phases. In Phase I, every sensor sends its data up to the highest cell in its column (in the *a*th row) as shown in Fig. 3(a) and Fig. 3(b), and in Phase II, all data is sent via cells in the *a*th row to the sink as shown in Fig. 3(c) and Fig. 3(d).

In each phase, the data transmissions in all interference blocks are performed in parallel.

C. Analysis of Delay Rate

Now we analysis the delay rate of our data collection scheme above. We define the time needed for the 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. To avoid interference, every interference block can only have one node send a packet to its upper neighbor in every time slot t during Phase I. In Fig. 3, bold lines show the interference blocks. Remember that $\frac{L}{d}$ is a constant α , thus the number of cells in the interference block is $(\frac{L}{d})^2 = \alpha^2$. And the packet in the lowest row (*i.e.* cell (0, k)) has to walk m cells to reach nodes in the highest cell in the rectangle. Hence,

$$T_1 \leq (\frac{L}{d})^2 \times t \times O(\log n) \times m$$

$$\leq O(t \log n)m = O(t \log n)\sqrt{\frac{n}{3\log n}} = O(t\sqrt{n\log n})$$

In the beginning of Phase II, all data are already at cells of the top row. The sink s lies on the same row with these cells. We now estimate T_2 needed for sending all data to s. Each cell in the top row has at most $mO(\log n)$ nodes' data and the number of cells in each interference block is $\frac{L}{d}$. Similarly, we can get

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

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 [5], [6]. It is obvious that the sink cannot receive at 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 for each individual sensor the lowest achievable delay rate of our method is $\Theta(W/n)$ which also meets the upper bound.

D. Capacity of Data Collection

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 transmission of previous snapshot. Whenever the cells in the top row receive $m \cdot b$ 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 $m \cdot b$ data in Phase I is

$$T'_1 \le (\frac{L}{d})^2 \times t \times O(\log n) = O(t \log n)$$

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

$$T'_2 \le \frac{L}{d} \times t \times m = O(t\sqrt{\frac{n}{\log n}}).$$

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

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

This also meets the upper bound W in order.

In summary, we have the following theorem:

Theorem 1: 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 $\Theta(W)$.

IV. DATA COLLECTION WITH AGGREGATION

In this section, we investigate a different data collection scenario 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. 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 II. 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, delay rate is $\frac{nb}{\Delta}$ and capacity is $\frac{nb}{T}$.

There is not much work on capacity of data aggregation in wireless sensor networks, except for [12] and [13]. Giridhar and Kumar [12] investigated a more general aggregation problem in random sensor network where a symmetric function of



Fig. 4. 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.

the sensor measurements is used for data aggregation. Moscibroda [13] then further studied the aggregation capacity for arbitrarily deployed networks under both protocol interference model and physical interference model.

A. Analysis of Delay Rate

We again assume that the sink s is located in cell (m, m). Our aggregation scheme has three phases and uses the same partition method in Section III.

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 2, 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 distance R, T''_1 is at most $(\frac{L}{d})^2 \cdot O(\log n) \cdot t$.

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 it to its upper cell. See Fig. 4(a). 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 $m \times t \times (\frac{L}{d}) = \Theta(\sqrt{\frac{n}{\log n}t})$, since every $\frac{L}{d}$ columns only one node can transmit due to interference, as shown in Fig. 4(a).

Third, as shown in Fig. 4(b), the information is aggregated via cells one by one in the top row. The time needed T_3'' is at most $m \times t = \Theta(\sqrt{\frac{n}{\log n}}t)$.

Therefore, the total delay $\Delta_{agg} \leq T_1'' + T_2'' + T_3'' = O(\sqrt{\frac{n}{\log n}}t)$. 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 the 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 reaching the sink. Since to compute the aggregated value, all values from the snapshot are needed, $T_{farthest} \leq T_{complete}$. Based on the network model,

the farthest node from the sink locates on one corner of the field. We denote the distance between the farthest node and the sink as R. It is easy to show that the minimum value of R is $\frac{\sqrt{2a}}{2}$ (when the sink is in the center of the field), *i.e.* $R \geq \frac{\sqrt{2n}}{2}$. The data in the farthest node needs at least $\frac{R}{r}$ time slots to reach the sink, for the transmission range is r. Hence,

$$T_{farthest} \ge \frac{R}{r} \cdot t = \frac{R}{r} \cdot \frac{b}{W} \ge \frac{\frac{\sqrt{2n}}{2}}{r} \cdot \frac{b}{W} = \sqrt{\frac{n}{30 \log n}} \cdot \frac{b}{W}$$

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

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

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

B. Capacity of Data Aggregation

We now describe our aggregation algorithm with pipelining. In the above algorithm, until the sink receives the aggregated value for all data in the previous snapshot, sensors begin to send data in the next snapshot. However, with pipelining, a sensor can start sending (or aggregating) data in the next snapshot before the aggregated value of the previous snapshot reaches the sink. Actually, it can initiate sending if the aggregated data of the previous snapshot are far away enough. Thus, all three phases in the algorithm can perform in pipelining.

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 $(\frac{L}{d})^2 \cdot O(\log n) \cdot t = O(t \log n)$.

For Phase II and Phase III if the aggregated values in previous snapshot are one interference block ahead (above or right in Fig. 4), the values from next snapshot can be sent or aggregated. The time difference between such two snapshots is bounded by $(\frac{L}{d})^2 \cdot t$. This is much smaller than the time used for aggregation of data in a cell ($O(t \log n)$). Thus, in a cell, when the aggregation of data from one snapshot finishes, the aggregation 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, every $O(t \log n)$ the sink can collect one snapshot data with pipelining. Then the capacity of our data aggregation method is $\frac{nb}{O(t \log n)} = \Omega(\frac{n}{\log n}W)$. Next, we prove that the upper bound of data aggregation

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 order of the optimal. Because *n* sensors are randomly distributed in the $\sqrt{n} \times \sqrt{n}$ square, if we divide the region into disks with radius $\frac{L}{2} = \alpha\sqrt{3\log n}/2$, every such disk has average $\frac{3\pi\alpha^2\log n}{4}$ 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., $T_{agg} \ge \Theta(\log nt)$. Thus, the capacity C_{agg} is less than or equal to $O(\frac{n}{\log n}W)$ for sure.

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

Theorem 2: 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.

Notice that for data collection the delay rate and the capacity are in the same order (Theorem 1), i.e., the pipelining can only improve 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 $\Theta(\sqrt{\frac{n}{\log^2 n}})$.

V. CONCLUSION

In this paper, we study the theoretical limitations of data collection in terms of delay and capacity for random sensor networks. For communication scenarios with or without aggregation, we prove that the asymptotical upper bound of delay rate and capacity, and propose a collection method to achieve the upper bound within a constant fact. These results can lead to better network planning and performance for data collection in wireless sensor network applications. For future work, it is interesting to study (1) data collection capacity with multiple sinks under physical inference model and (2) data collection capacity for arbitrary networks.

REFERENCES

- P. Gupta and P.R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, 46(2):388-404, 2000.
- [2] M. Grossglauser and D. Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Proc. of IEEE INFOCOM*, 2001.
- [3] X.-Y. Li, S. Tang, and O. Frieder, "Multicast capacity for large scale wireless ad hoc networks," in *Proc. of ACM MobiCom*, 2007.
- [4] A. Keshavarz-Haddad, V. Ribeiro, and R. Riedi, "Broadcast capacity in multihop wireless networks," in *Proc. of ACM MobiCom*, 2006.
- [5] E. J. Duarte-Melo and M. Liu, "Data-gathering wireless sensor networks: Organization and capacity," *Computer Networks*, 43:519-537, 2003.
- [6] M. Liu D. L. Neuhoff D. Marco, E. J. Duarte-Melo, "On the manyto-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proc. of ACM IPSN*, 2003.
- [7] S. Chen, Y. Wang, X.-Y. Li, and X. Shi, "Order-Optimal Data Collection in Wireless Sensor Networks: Delay and Capacity," in *Proc. of IEEE* SECON, 2009.
- [8] B. Liu, D. Towsley, and A. Swami, "Data gathering capacity of large scale multihop wireless networks," in *Proc. of IEEE MASS*, 2008.
- [9] H.E.Gamal, "On the scaling laws of dense wireless sensor networks: the data gathering channel," *IEEE Trans. IT*, 51(3):1229-34, 2005.
 [10] R. Zheng and R.J. Barton, "Toward optimal data aggregation in random
- [10] R. Zheng and R.J. Barton, "Toward optimal data aggregation in random wireless sensor networks," in *Proc. of IEEE INFOCOM*, 2007.
- [11] R.J. Barton and R. Zheng, "Order-optimal data aggregation in wireless sensor networks using cooperative time-reversal communication," in *Proc. of Annual Conf on Information Sciences and Systems*, 2006.
- [12] A. Giridhar and P.R. Kumar, "Computing and communicating functions over sensor networks," *IEEE J. Sel. Areas in Com.*, 23(4):755-764,2005.
- [13] T. Moscibroda, "The worst-case capacity of wireless sensor networks," in *Proc. of ACM JPSN* 2007
- in *Proc. of ACM IPSN*, 2007.
 [14] S.R.Kulkarni, P.Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. IT*, 50(6):1041-9, 2004.
- [15] S. Rao, "The *m* balls and *n* bins problem," Lecture Note for Lecture 11, CS270, Univ. of California, Berkeley, 2003.