Order-Optimal Data Collection in Wireless Sensor Networks: Delay and Capacity

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 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 different communication scenarios (single sink or multiple sinks, regularly-deployed or randomly-deployed sinks, with or without aggregation) under protocol interference model. For each scenario, we first propose a new collection/aggregation method and analyze its performance in terms of delay and capacity, then theoretically prove that our method can achieve the optimal order (i.e., its performance is within a constant factor of the optimal). Particularly, with a single sink, the capacity of data collection is in order of $\Theta(W)$ where W is the fixed data-rate on individual links. With k sinks, the capacity of data collection is increased to $\Theta(kW)$ when $k = O(\frac{n}{\log n})$ or $\Theta(\frac{n}{\log n}W)$ when $k = \Omega(\frac{n}{\log n})$. If each sensor can aggregate its receiving packets into a single packet to send, the capacity of data collection with a single sink is also increased to $\Theta(\frac{n}{\log n}W)$.**

I. Introduction

A wireless sensor network (WSN) consists of a set of sensor devices which spread over a geographical area. These sensors are able to perform processing as well as sensing and are additionally capable of communicating with each other. Due to its wide-range potential applications such as battlefield, emergency relief, environment monitoring, and so on, sensor network has recently emerged as a premier research topic. For wireless sensor networks, often the ultimate goal is to collect sensing data from all sensors to certain sink nodes and then perform further analysis at these sink nodes. 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.

The work of Y. Wang is supported in part by the US National Science Foundation (NSF) under Grant No. CNS-0721666 and funds provided by the University of North Carolina at Charlotte. The work of X.-Y. Li is partially supported by NSF under Grant No. CNS-0832120 and CCF-0515088, 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, Hong Kong RGC under Grant HKUST 6169/07 and HKBU 2104/06E, and CERG under Grant PolyU-5232/07E. Part of the work was done when X.-Y. Li visited Microsoft Research Asia, Beijing, China.

We consider a dense wireless sensor network where n sensors are randomly deployed in a finite geographical region. Each sensor measures independent field values at regular time intervals and sends these values to sinks. The union of all sensing values from n sensors at particular time is called a *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. As in the literature, the classical *protocol interference model* is used in our analysis, while all analysis results can also be extended to *physical interference model* by applying the technique introduced in [20], [26]. We also assume that a successful transmission over a link has a fixed data-rate of 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 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, large delay rate is desired. When multiple snapshots from sensors are generated continuously, data transport can be pipelined in the sense that further snapshot 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. Notice that the capacity is always larger than or equal to the delay rate. Both *delay rate* and *capacity* reflect that 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 performance of wireless sensor networks. In this paper, we are particularly interested in how the delay rate and capacity of data collection vary as the number of sensors n increases.

Capacity limits of data collection in random wireless sensor networks have been studied in the literature [1]–[4]. In [1], [2], 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. El Gamal [3] 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 Rong [4] 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 [3] and [4] assumed complex physical layer techniques, such as antenna sharing, channel coding and cooperative beamforming. More related work is reviewed in Section III. In this paper, we focus on deriving capacity bounds of data collection under different modalities of communication and/or assumptions, such as with single sink or multiple sinks, grid or random sink deployment, with or without aggregation, etc.

Main Contributions: In this paper, we make the following contributions:

- For sensor networks with a single sink under protocol interference model, we propose a new data collection method whose delay rate and capacity are both $\Theta(W)$ which match the theoretical upper bounds.
- When sensor networks have k sinks (which are either regularly deployed on a grid or randomly deployed in the field), we prove that the upper bounds of the capacities increase 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.
- One variation of data collection problem is data aggregation [5], [6] in which sensors can cooperate to aggregate information as it is transmitted to the sink. With data aggregation, communication overhead is reduced and the capacity is increased. Ideally, the sink can still receive enough information to compute the desired measurements. Here, we only consider the simplest set of aggregation functions by which multiple packets can be merged into a single packet at each sensor, such as the functions of maximum, summation, or mean. 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 $\Theta(\sqrt{\frac{n}{\log^3 n}})$.

For all above cases, we propose our own collection or aggregation methods which can achieve the optimal order (i.e., their performances are within a constant factor of upper bounds).

The rest of this paper is organized as follows. In Section II, we provide our network model and formal definitions of delay rate and capacity. We briefly review related work on capacities of data collection in wireless networks in Section III. Capacities of data collection with a single sink or multiple sinks are studied in Section IV and Section V respectively under protocol interference model. The capacity of data aggregation is then studied in Section VI. Finally, Section VII concludes the paper by pointing out possible future work.

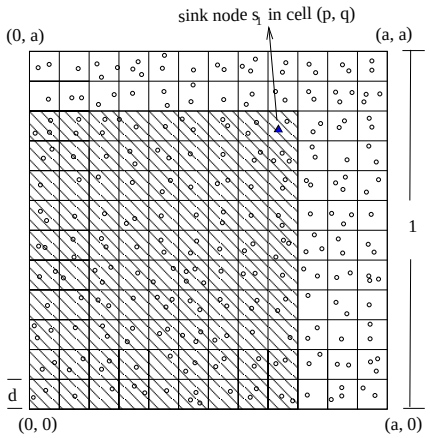


Fig. 1. Grid partition of the sensor network: a^2 cells with cell size of $d \times d$.

II. Preliminaries

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

A. Network Model

In this paper, we focus on the capacity bound of data collection in wireless sensor networks. Thus, for simplicity, we assume a simple and yet general enough model that is widely used in the community. We consider a sensor network which includes n wireless sensor nodes $V = \{v_1, v_1, \cdots, v_n\}$ and k sink nodes $S = \{s_1, s_2, \cdots, s_k\}$. Here, we assume that both sensor nodes and sink nodes are uniformly deployed in a square of unit area. 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 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 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 also assume that all packets have the unit size of b bits. Time is slotted into 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 *protocol interference model* in our analysis. In protocol interference model, 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. Here, for simplicity, we assume that $\frac{R}{r}$ is a constant α which is larger than 1.

B. Our Partition Method

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

The following lemma gives a guidance of the cell size.

Lemma 1: [7] 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 we set $d = \sqrt{3\frac{\log n}{n}}$ (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, we can still make the above equation hold by adjusting the deployment density (*i.e.*, n).

We then 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: [8], [9] 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.

C. Capacity and Delay

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

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. 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 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 paper, we analyze both the delay rate and capacity for data collection in random sensor networks.

III. Related Work

Gupta and Kumar initiated the research on capacity of random wireless networks by studying the unicast capacity in the seminal paper [10]. A number of following papers studied capacity under different communication scenarios: unicast [11], [12], multicast [13]–[15], broadcast [16], [17], [21]. In this paper, we focus on the capacity of data collection or data aggregation in a many-to-one communication scenario for random sensor networks.

Capacity of data collection in wireless sensor networks has been studied in [1]–[4], [18], [19], [22]. In [1], [2], Duarte-Melo *et al.* first studied the many-to-one transport capacity in dense and random sensor networks. They only considered the case with a single sink under protocol interference model. Using a different method, they showed that the overall capacity of data collection is $\Theta(W)$. They also studied how to compress the data to improve the capacity in [2]. El Gamal [3] studied the capacity of data collection subject to a total average transmitting power constraint. They relaxed the assumption that every node can only receive from one source node at a time. It was shown that the capacity of random networks scales as $\Theta(\log n)$ when n goes to infinity and the total average power remains fixed. Their method uses antenna sharing and channel coding. Recently, Barton and Rong [4], [18] also investigated the capacity of data collection under more complex physical layer models (non-cooperative SINR model and cooperative time reversal communication (CTR) model) where the data rate of individual link is not fixed as a constant W but depends on the level of interference which is decided by transmitting powers and transmitting distances of all simultaneous transmissions. They first demonstrated that $\Theta(\log n)$ is optimal and achievable using CTR for a regular grid network in [18], then showed that the capacity of $\Theta(\log n)$ and $\Theta(1)$ are optimal and achievable by CTR when operating

in fading environments with power path-loss exponents that satisfy $2 < \beta < 4$ and $\beta \geq 4$ for random networks [4]. Liu *et al.* [19] recently studied 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 constructive lower bounds for such problem. Notice that data collection is a special case for their problem when $s(n) = n$ and $d(n) = 1$. However, their results have a gap between the upper and lower bounds due to the random selection of the sources and designations. In addition, Zhu *et al.* [22] studied how to schedule data collection in an arbitrary sensor network, where sensors are not randomly distributed, such that delay or latency is minimized. Notice that random network is a subset of arbitrary network. However, they only adopted a simpler protocol interference model where $R = r$ and each node has two individual channels. Two constant approximation algorithms were presented in [22].

There is not much work on capacity of data aggregation in wireless sensor networks, except recent papers [5] and [6]. Giridhar and Kumar [5] investigated a more general aggregation problem in random sensor network where a symmetric function of sensor measurements is used for data aggregation. It was shown that for random planar multihop network, the maximum rate for computing divisible functions (a subset of symmetric functions) is $\Theta(\frac{1}{\log n}W)$. Notice that they defined the maximum rate as the data rate per sensor. By our definition, their result is $\Theta(\frac{n}{\log n}W)$ which matches our result in Section VI, since our aggregation function (e.g. maximum) is a divisible function. In addition, using a technique called block-coding, they further showed that type-threshold functions can be computed at a rate of $\Theta(\frac{1}{\log \log n}W)$. Moscibroda [6] then further studied the aggregation capacity for arbitrarily deployed networks (he called it as *worst-case capacity*) under both protocol interference model and physical interference model. He showed that the worst-case capacities of data aggregation are $\Theta(\frac{1}{n}W)$ and $\Omega(\frac{1}{\log^2 n}W)$ respectively. Finally, there are also some results [23]–[25] on how to schedule data aggregation in arbitrary sensor networks to minimize delay. Such problem has been proved NP-hard [23], and several approximation algorithms were proposed.

IV. 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.

As shown in Figure 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 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 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 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 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 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,

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

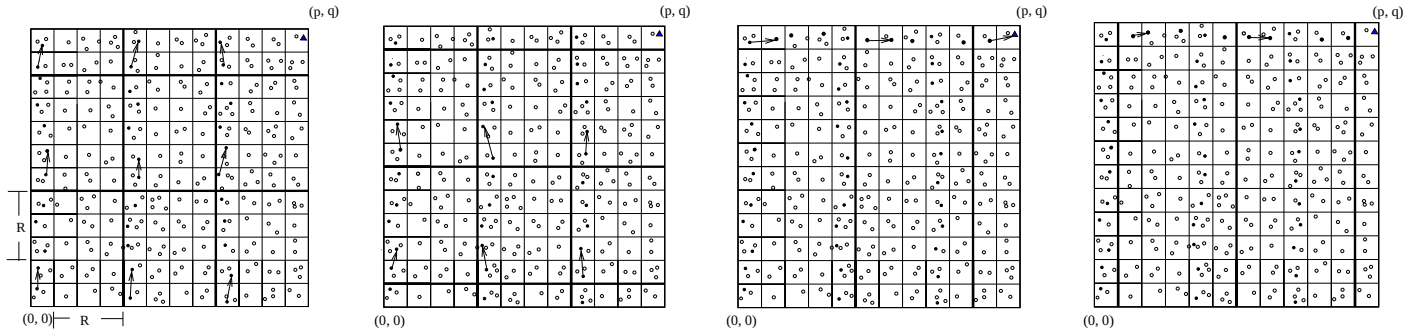
In the beginning of Phase II, all data are already at cells of the top row. The sink s_1 lies in the same row with these cells. We now estimate the time T_2 needed for sending all data to s_1. Each cell in the top row has at most $qO(\log n)$ nodes' data and the interference block is $1 \times (\frac{R}{d}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



(a) Phase I, 1st time slot (b) Phase I, 2nd time slot (c) Phase II, 1st time slot (d) Phase II, 2nd time slot

Fig. 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.

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

$$T_2' \leq (\frac{R}{d} + 2) \times t \times p = O(t\sqrt{\frac{n}{\log n}}).$$

Therefore, the total time for sink to receive $p \cdot b$ data is $T_1' + T_2' = O(t\sqrt{\frac{n}{\log n}})$. Thus, the capacity of our method with pipelining is still

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

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 $\Theta(W)$.

V. 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 IV). 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.

A. 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: $k \leq \frac{n}{15(\alpha+1)^2 \log n}$ or $k > \frac{n}{15(\alpha+1)^2 \log n}$.

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 $k < \frac{n}{15(\alpha+1)^2 \log n}$, every sink performs the collection method to collect their sub-areas. When $k \geq \frac{n}{15(\alpha+1)^2 \log n}$, $\frac{1}{(R+r)^2}$ sinks can be selected and 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.

Therefore, we have our second theorem.

Theorem 2: 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

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

B. Randomly Deployed Multiple Sinks

Consider the scenario when k sinks are randomly distributed in the network. It is clear that if k is very small, the capacity of collection should also be $\Theta(kW)$, since there will be no interference among sub-areas around neighboring sinks with high probability. Here we can use Voronoi diagram to divide the network into sub-areas. However, when k is very large, the capacity is still bounded by the interference area.

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 $2R \times 2R$. Thus, there are $B = \frac{n}{60\alpha^2 \log n}$ interference blocks. We then consider three cases when we randomly put k sinks into B interference blocks:

Case 1: When $k = o(\frac{n}{\log n})$. We calculate the probability that an interference block has more than two sinks.

$$\begin{aligned} & Pr(\text{an interference block has more than 2 sinks}) \\ = \quad & \binom{k}{2} \cdot (\frac{1}{B})^2 = \Theta((\frac{k}{\frac{n}{\log n}})^2). \end{aligned}$$

When $n \longrightarrow \infty$, this probability goes to 0 since $k = o(\frac{n}{\log n})$. In other words, no interference block has more than two sinks. Therefore, each sink can collect data at the same time with high probability. For this case, the capacity of data collection is bounded by $\Theta(kW)$ from above. 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.

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

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 $\Theta(kW)$).

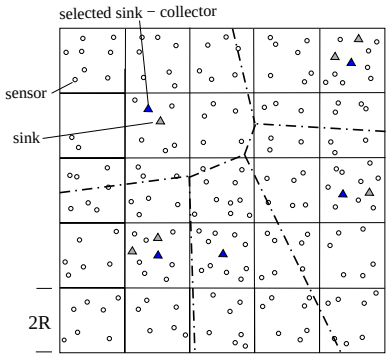


Fig. 3. **Voronoi partition of the sensor network:** each square is a $2R \times 2R$ block; one collector is selected per block if the block has some sinks; the dash lines are Voronoi diagram of collectors.

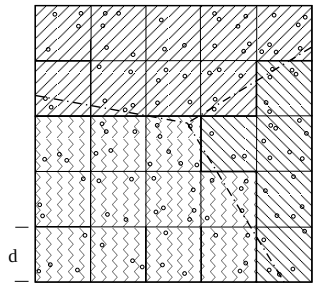


Fig. 4. **Micro cell assignment for Voronoi diagram:** each square is a $d \times d$ micro cell; each cell is assigned to one Voronoi region; the dash lines are original edges in Voronoi diagram; here three Voronoi regions are presented.

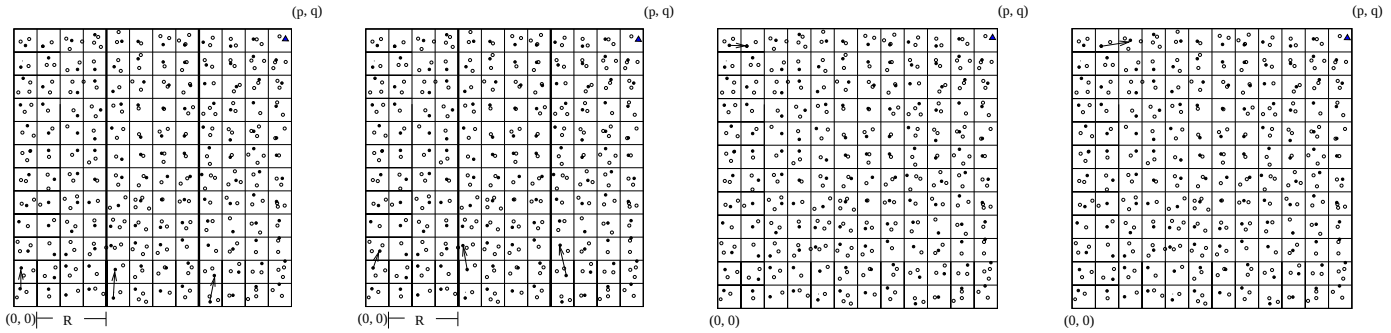
Case 3: When $k = \omega(\frac{n}{\log n})$. We also consider the probability that an arbitrary interference block has at least one sink.

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

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.

To achieve these capacity bounds, our data collection algorithm works as follows. We first use $2R \times 2R$ to partition the network into interference blocks. For each block, we choose only one sink to be the collector if the block is occupied



(a) Phase II, 1st time slot (b) Phase II, 2nd time slot (c) Phase III, 1st time slot (d) Phase III, 2nd time slot

Fig. 5. 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.

by some sinks. Then we use the Voronoi diagram defined by all selected collectors to partition the field, as shown in Figure 3. Each collector will collect data from sensors inside its Voronoi region using the collection method in Section IV for the single sink case. Note that a micro cell could lie in multiple Voronoi regions. However, we assign the micro cell to the region which covers the largest area inside the cell. See Figure 4 for illustration. By this way, a simple scheduling can avoid the interference among adjacent cells and the capacity of data collection is still in order of the theoretical bound.

Theorem 3: 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

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

In summary, with multiple sinks (either grid or random deployment of k sinks), the capacity of data collection is increased 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 [19] for many-to-many capacity.

VI. Data Aggregation with Single Sink

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 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, the delay rate is $\frac{nb}{\Delta}$ and the capacity is $\frac{nb}{T}$.

A. Delay Rate

We assume that a single sink s is located in cell (p, q) and we 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 IV.

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 Figure 5 (a) and (b) 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 5 (a) and (b).

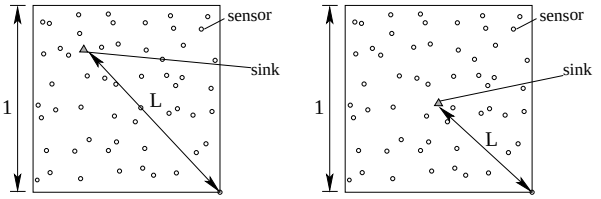
Third, as shown in Figure 5 (c) and (d), the information is aggregated via cells one by one in the top row. The time needed T_3'' is at most

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

Therefore, the total delay $\Delta_{agg} \leq T_1'' + T_2'' + T_3'' = O(\sqrt{\frac{n}{\log n}}t)$. 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 in order of the 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, $T_{farthest} \leq T_{complete}$. 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 $\frac{\sqrt{2}}{2}$ (when the sink is in the center of the field), *i.e.* $L \geq \frac{\sqrt{2}}{2}$. See Figure 6 for illustrations.



(a) $L > \frac{\sqrt{2}}{2}$ (b) $L = \frac{\sqrt{2}}{2}$

Fig. 6. Minimum value of L is $\frac{\sqrt{2}}{2}$.

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

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

Therefore, the delay rate of data aggregation is at most

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

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

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

For Phase II and Phase III if the aggregated values in previous snapshot are one interference block ahead (above or right in Figure 5), 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, 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 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.

In summary, we have the last theorem as follows.

Theorem 4: 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 $\Theta(\sqrt{\frac{n}{\log^3 n}})$.

So far, we only consider data aggregation capacity with a single sink, but results for the case with multiple sinks are easy to derive using the similar analysis in Section V. Due to space limit, here we just present the conclusion. The delay rate and capacity of data aggregation in random sensor networks with k sinks are

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

[1] We can also think 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.

VII. Conclusion

In this paper, we study theoretical limitations of data collection and data aggregation in terms of delay and capacity for random sensor networks. For different communication scenarios, we prove the asymptotical upper bound of delay rate and capacity and then propose a collection method to achieve the upper bound within a constant fact. Summary of all results is shown in Table I. These results can lead to better network planning and performance for data collection or data aggregation in wireless sensor network applications.

TABLE I

Summary of Results on Delay Rate and Capacity

Data Collection Problem	Delay Rate Γ	Capacity C
data collection, $k = 1$	$\Theta(W)$	$\Theta(W)$
data collection, $k = O(\frac{n}{\log n})$	$\Theta(kW)$	$\Theta(kW)$
data collection, $k = \Omega(\frac{n}{\log n})$	$\Theta(\frac{n}{\log n}W)$	$\Theta(\frac{n}{\log n}W)$
data aggregation, $k = 1$	$\Theta(\sqrt{n \log n}W)$	$\Theta(\frac{n}{\log n}W)$
data aggregation, $k = O(\frac{n}{\log n})$	$\Theta(k\sqrt{n \log n}W)$	$\Theta(\frac{kn}{\log n}W)$
data aggregation, $k = \Omega(\frac{n}{\log n})$	$\Theta(\frac{n\sqrt{n}}{\sqrt{\log n}}W)$	$\Theta((\frac{n}{\log n})^2W)$

Here, we only consider capacities for random sensor networks where sensors are uniformly distributed in the field. It is interesting to study the capacities of data collection for general networks (arbitrarily deployed networks), as in [6], [22], [24]. We leave this as one of our future work. Another direction is to investigate the achievable capacity and delay rate when the energy consumption of data collection is considered, since energy issue is very critical for wireless sensor networks. Recently, Li *et al.* [27] showed the trade-offs between the delay and energy consumption of data collection or data aggregation in sensor networks.

References

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

[2] Mingyan Liu David L. Neuhoff Daniel Marco, Enrique J. Duarte-Melo, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *Proc. of International Workshop on Information Processing in Sensor Networks*, 2003.

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

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

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

[6] Thomas Moscibroda, "The worst-case capacity of wireless sensor networks," in *IPSN '07: Proceedings of the 6th international conference on Information processing in sensor networks*, 2007.

[7] 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.

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

[9] Martin Raab and Angelika 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.

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

[11] Matthias Grossglauser and David Tse, "Mobility increases the capacity of ad-hoc wireless networks," in *Proc. of IEEE Conference on Computer Communications (Infocom)*, 2001.

[12] Benyuan Liu, Patrick Thiran, and Don Towsley, "Capacity of a wireless ad hoc network with infrastructure," in *Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2007.

[13] Xiang-Yang Li, Shao-Jie Tang, and Ophir Frieder, "Multicast capacity for large scale wireless ad hoc networks," in *Proc. of ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2007.

[14] Xufei Mao, Xiang-Yang Li, and Shaojie Tang, "Multicast capacity for hybrid wireless networks," in *Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2008.

[15] Srinivas Shakkottai, Xin Liu, and R. Srikant, "The multicast capacity of large multihop wireless networks," in *Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2007.

[16] Alireza Keshavarz-Haddad, Vinay Ribeiro, and Rudolf Riedi, "Broadcast capacity in multihop wireless networks," in *Proc. of ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2006.

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

[18] R.J. Barton and R. Zheng, "Order-optimal data aggregation in wireless sensor networks using cooperative time-reversal communication," in *Proc. of 40th Annual Conference on Information Sciences and Systems*, 2006.

[19] B. Liu, D. Towsley, and A. Swami, "Data gathering capacity of large scale multihop wireless networks," in *Proceedings of Fifth IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2008.

[20] A. Agarwal and P.R. Kumar, "Capacity bounds for ad hoc and hybrid wireless networks," *ACM SIGCOMM Computer Communication Review*, vol. 34, no.3, 71–81, 2004.

[21] R. Gandhi, S. Parthasarathy, and A. Mishra, "Minimizing broadcast latency and redundancy in ad hoc networks," in *Proceedings of ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2003.

[22] X. Zhu, B. Tang, and H. Gupta, "Delay efficient data grathering in sensor networks," in *Proc. of IEEE 1st Int'l Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, 2005.

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

[24] 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 *Proc. of IEEE Conference on Computer Communications (Infocom)*, 2007.

[25] 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, 626–636, 2008.

[26] 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 Fifth IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS)*, 2008.

[27] X.-Y. Li, Y. Wang, and Y. Wang, "Complexity of Data Collection, Aggregation, and Selection for Wireless Sensor Networks," Submitted for publication, 2009.