SelectCast: Scalable Data Aggregation Scheme in Wireless Sensor Networks

Cheng Wang*†, Shaojie Tang‡, Xiang-Yang Li‡§, Changjun Jiang*†
* Department of Computer Science, Tongji University, Shanghai, China
† Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai, China
‡ Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 60616
§ TNLIST, School of Software, Tsinghua University

***Abstract*—In this work, for a wireless sensor network (WSN) of n randomly placed sensors with node density $\lambda \in [1, n]$, we study the tradeoffs between the *aggregation throughput* and *gathering efficiency*. The gathering efficiency refers to the ratio of the number of the sensors whose data has been gathered to the total number of sensors. Specifically, we design two efficient aggregation schemes, called *single-hop-length* (SLH) scheme and *multiple-hop-length* (MLH) scheme. By novelly integrating these two schemes, we theoretically prove that our protocol achieves the optimal tradeoffs, and derive the optimal aggregation throughput depending on a given threshold value (lower bound) on gathering efficiency. Particularly, we show that under the MLH scheme, for a practically important set of symmetric functions called *perfectly compressible functions*, including the mean, max, or various kinds of indicator functions, *etc.*, the data from $\Theta(n)$ sensors can be aggregated to the sink at the throughput of a constant order $\Theta(1)$, implying that our MLH scheme is indeed scalable.**

***Index Terms*—Wireless sensor networks, Data Aggregation, Percolation theory, aggregation capacity.**

I. Introduction

Data aggregation is a key energy consuming functionality in wireless sensor networks (WSNs) for both data gathering applications and event-based applications, since the communication cost is often the higher-order of the computation cost [1]. It has been shown in the literature that the achievable minimum data rate among all sensor nodes is severely limited for random WSNs if we insist data from *all* sensors should be collected. In this paper, we design structure-based aggregation schemes for WSNs to achieve the optimal tradeoffs between the *aggregation throughput* and *gathering efficiency*. Here the gathering efficiency refers to the ratio of the number of the sensor nodes whose data were gathered successfully to the total number of sensor nodes in the network. Collecting data from a subset of sensor nodes is reasonable because of the potential spatial correlations among sensed environment. In our protocol, for the neighborhood of every node, we will approximately select Ψ portion of nodes and aggregate their data to the sink. Such sampling scheme will achieve high aggregation throughput while maintaining the spatial coverage by the sampled sensors.

For data gathering, we focus on an important set of symmetric functions called *perfectly compressible functions*, such as the mean, max, or kinds of indicator functions [2] that will be used to compute the data aggregation. Two characteristics of this work are extracted as following:

• To meet specific application requirement, e.g., full coverage, k-coverage, connectivity, etc the node density (number of nodes per unit area) can be treated as a varible within a large range. Thus we consider a random deployed WSN with a general density, where n sensors constitute a network with node density λ, $1 \leq \lambda \leq n$, rather than the special *random dense networks* or *random extended networks*, where $\lambda = n$ and $\lambda = 1$, respectively. Depending on the requirement of gathering efficiency, we determine the thresholds of the density λ by which the aggregation throughput and tradeoffs are divided into different regimes.

• As the node density is decreasing and the area of deployment region is thus increasing, we need to rely on some long links to ensure the network connectivity. For those long links, it is unrealistic to set the link rate be a constant order as under the *protocol model* and *physical model* [3]. Hence, we design efficient protocols under a more realistic model called *generalized physical model*, rather than under the protocol model [4], [5] or physical model [2].

Under the structure-based aggregation schemes for a random WSN, the aggregation throughput for a specific type of function is mainly limited by the following two factors:

• **Outliers**: In random networks, given a proper threshold (upper bound) on the length of links, there is a giant *connected component* in which any pair of nodes can be connected by the link of length below the threshold. While, there might be some nodes, called *outliers*, outside a specific connected component. To reach them, some links longer than the threshold are needed, which possibly leads to lower link rate.

• **Dense Components**: Given a deterministic routing, in the conflict graph modeling link interferences, there might be some cliques (complete subgraphs) of high-order size. Then, the scheduling of corresponding links might become a bottleneck.

To address these limitations and challenges, we design two efficient protocols to improve the tradeoffs between throughput and gathering efficiency.

• **Single-Hop-Length (SLH) Scheme**: The routing is non-hierarchical and consists of the links with similar lengths. By selecting a certain number of sensors in local regions depending on the given lower bound on gathering efficiency, we improve the throughput by deliberating the bottleneck produced by the second limitations, i.e., dense components.

• **Multiple-Hop-Length (MLH) Scheme**: The routing is hierarchical and consists of the links with various lengths. By selecting a fixed number of sensors from local regions and limiting the length of those long links, we improve the aggregation throughput by deliberating the bottleneck produced by both the outliers and dense components.

In summary, our main contributions are as follows:

• Scalability is an important metric when designing the network protocol. We prove that under the MLH scheme, the measurements from $\Theta(n)$ sensors can be aggregated into the sink at the throughput of order $\Theta(1)$, which means that the MLH aggregation scheme is indeed scalable. To the best of our knowledge, our MLH scheme is the first scalable structure-based aggregation scheme.

• Combining the schemes SLH and MLH, we derive the optimal tradeoffs between the aggregation throughput and gathering efficiency for perfectly compressible functions in the random WSN with general density λ, $1 \leq \lambda \leq n$, as illustrated in Fig.1. When we set the gathering efficiency be 1 and the node density λ be $\Theta(n)$, the resulted aggregation throughput is specified into the ordinary aggregation throughput for random dense WSNs [2], [4], [5].

The rest of the paper is organized as follows. In Section II, we introduce the system model and formulate the problem. In Section III, we propose two aggregation schemes for random WSNs with general density. We derive the achievable aggregation throughput and the tradeoffs between it and the gathering efficiency in Section IV. In Section V, we draw some conclusions and future perspective.

II. System Model

A. Network Model

Assume that the sensors are deployed on the 2-dimension plane according to a Poisson point process of density λ, where $\lambda = \Omega(1)$ and $\lambda = O(n)$. We consider a random network consisting of n (or $\Theta(n)$) sensors. Specifically, we focus on the square $\mathcal{A}(\lambda, n) = [0, \sqrt{n/\lambda}]^2$. Then, according to Chebyshev's inequality, the number of sensors in $\mathcal{A}(\lambda, n)$ is within $[(1-\varepsilon)\cdot n, (1+\varepsilon)\cdot n]$ with high probability. To simplify the description, we assume the number of nodes is exactly n, without changing the final results in order sense. Denote $\mathcal{S}(n) = \{s_0\} \cup \{s_1, s_2, \cdots, s_{n-1}\}$, where s_0 is the sink node and $s_i, i \in [1, n-1]$ are the ordinary sensor nodes. In the following, we denote such a random network by $\mathcal{N}(\lambda, n)$.

B. Aggregation Throughput for Wireless Sensor Networks

As in the models of most related works [2], [5], every sensor node s_i, $i \in [0, n-1]$, periodically generates measurements of the environment, which belong to a fixed finite set $\mathcal{M}$ with $|\mathcal{M}| = \mathfrak{m}$, and the function of interest is then required to be computed periodically for the measured data. Define the function of interest to sink node as $\mathbf{g}_n$:$\mathcal{M}^n \rightarrow \mathcal{G}_n$; furthermore, for any node $k \in [1, n]$, define the function of the sensor measurements as $\mathbf{g}_k$:$\mathcal{M}^k \rightarrow \mathcal{G}_k$, where $\mathcal{G}_k$ is the range of $\mathbf{g}_k$. Suppose that each sensor has an associated block of L readings, known a *priori* [5]. We call L rounds of measurements an *aggregation unit*. Notice that the aggregation operation can only be applied to the data from the same round. Before formulating the definition of aggregation throughput, we introduce some notations in Table I.

TABLE I
Some Notations.

Notations	Meaning
$\phi(n) \sim [\phi_0(n), \phi_1(n)]$	$\phi(n) = \Omega(\phi_0(n))$ and $\phi(n) = O(\phi_1(n))$.
$\phi(n) \sim (\phi_0(n), \phi_1(n))$	$\phi(n) = \omega(\phi_0(n))$ and $\phi(n) = o(\phi_1(n))$.
$\mathcal{A}(\lambda, n)$	the square region $[0, \sqrt{n/\lambda}]^2$.
$\mathcal{N}(\lambda, n)$	a random network composed of n sensors with density λ.
$\mathfrak{m} := \lvert\mathcal{M}\rvert$	the size of a fixed finite set $\mathcal{M}$ containing all measurements.
L	block-length, i.e., the size of aggregation units.
$\mathrm{M}^{n\times L} \in \mathcal{M}^{n\times L}$	a $n \times L$ matrix of measurements.
$\mathrm{M}^{n\times L}(i, j)$	the j-th measurement of sensor node s_i.
$\mathrm{M}^{n\times L}(i, \cdot)$	a block of L consecutive measurements of s_i.
$\mathrm{M}^{n\times L}(\cdot, j)$	a set of the j-th measurements of n sensors.
$\mathbf{g}_k(\mathrm{M}^k)$	$:= \mathbf{g}_k(\mathrm{M}_1, \mathrm{M}_2, ..., \mathrm{M}_k)$, for any k-vector $\mathrm{M}^k = [\mathrm{M}_1, \mathrm{M}_2, ...\mathrm{M}_k]^T \in \mathcal{M}^k$.
$\mathbf{g}_k^{\mathrm{L}}(\mathrm{M}^{k\times L})$	$:= \big(\mathbf{g}_k(\mathrm{M}^{k\times L}(\cdot, 1)), \cdots, \mathbf{g}_k(\mathrm{M}^{k\times L}(\cdot, L))\big)$, for a given matrix $\mathrm{M}^{k\times L}$.
$\Psi := \Psi(n)$	gathering efficiency.
$\mathcal{S}(\Psi n)$	a subset consisting of $\Psi \cdot n$ sensors.
$\mathscr{A}(n, L, \mathcal{S}(\Psi n))$	a scheme that can aggregate the measurements from sensors in $\mathcal{S}(\Psi n)$.
$\Lambda := \Lambda(\lambda, n)$	the achievable aggregation throughput.
$\Phi := \Phi(\lambda, n)$	Tradeoff between throughput and gathering efficiency.

1) Aggregation Functions of Interest: We focus on an important class of symmetric functions called *perfectly compressible* [2]. Functions such as the mean, max (or min), or various kinds of indicator functions all belong to this category.

Note that $|\mathcal{G}_k| = \Theta(\mathfrak{m})$ is not a sufficient condition ensuring the aggregation function $\mathbf{g}_k$ to be perfectly compressible [2], [5]. For simplicity, we assume that $|\mathcal{G}_k| = \mathfrak{m}$ for a perfectly compressible aggregation function, without changing the order of the derived throughput. Functions like max, min, *etc.*, belong to this category.

2) Achievable Aggregation Throughput: Denote an aggregation scheme as $\mathscr{A}(n, L, \mathcal{S}(\Psi \cdot n))$, where

- L denotes the block-length, which determines a sequence of message passings between sensors and computations at sensors;
- $\mathcal{S}(\Psi \cdot n) \subseteq \mathcal{S}(n)$, $\Psi \in (0, 1]$, is a subset of sensors which will be used to measure the gathering efficiency;
- input any $\mathrm{M}^{(\Psi\cdot n)\times L} \in \mathcal{M}^{(\Psi\cdot n)\times L}$ from all sensors in $\mathcal{S}(\Psi \cdot n)$, output $\mathbf{g}_{(\Psi\cdot n)}^{\mathrm{L}}(\mathrm{M}^{(\Psi\cdot n)\times L})$ at the sink node.

Next, we define the *achievable aggregation throughput*. All the logs in this paper are to the base 2.

Definition 1: For a given aggregation function: $\mathbf{g}_n$:$\mathcal{M}^n \rightarrow \mathcal{G}_n$, we say a throughput of $\Lambda(n) = \frac{L\cdot\log\mathfrak{m}}{T}$ bps Ψ-*achievable*, if there exists an aggregation scheme $\mathscr{A}(n, L, \mathcal{S}(\Psi \cdot n))$ under which there is a subset $\mathcal{S}(\Psi \cdot n)$ such that the corresponding $\mathrm{M}^{(\Psi\cdot n)\times L} \in \mathcal{M}^{(\Psi\cdot n)\times L}$ can be aggregated into

TABLE II
SOME PRE-DEFINED CONSTANT PARAMETERS.

Range	Conditions
$c \in (0,\infty)$	$c^2 > \ln 6$
$\kappa \in (0,\infty)$	$\kappa > 2/(c^2 - \ln 6)$
$\varpi \in (0,\infty)$	$\varpi < (\kappa \cdot (c^2 - \ln 6) - 2)/(\ln(1 - e^{-c^2}) + c^2)$
$\varepsilon_1 \in (0,\infty)$	arbitrary
$\varepsilon_2 \in (0,1)$	$\varepsilon_2 + (1-\varepsilon_2)\ln(1-\varepsilon_2) > 0$
$\varepsilon_3 \in (0,\infty)$	$(1+\varepsilon_3)\ln(1+\varepsilon_3) - \varepsilon_3 > 0$
$\varepsilon_4 \in (0,1)$	$\varepsilon_4 + (1-\varepsilon_4)\ln(1-\varepsilon_4) > 1/z$, for $z \in (0,\infty)$
$\varepsilon_5 \in (0,\infty)$	$(1+\varepsilon_5)\ln(1+\varepsilon_5) - \varepsilon_5 > 1/z$, for $z \in (0,\infty)$
$\varepsilon_6 \in (0,\infty)$	$\sigma\lambda \cdot ((1+\varepsilon_6)\cdot\ln(1+\varepsilon_6) - \varepsilon_6) + \ln(\sigma\lambda) = o(\ln n)$
$\varepsilon_7 \in (0,\infty)$	arbitrary
$\varepsilon_8 \in (0,1)$	$\varepsilon_8 = (1+\varepsilon_3)\cdot(1+\varepsilon_7)\cdot\left(e^{\varepsilon_6}/(1+\varepsilon_6)^{(1+\varepsilon_6)}\right)^{c^2}$
$\varepsilon_9 \in (0,1)$	$\varepsilon_9 \geq \varepsilon_2 + \varepsilon_8$

$\mathbf{g}^{\mathrm{L}}_{(\Psi\cdot n)}(\mathrm{M}^{(\Psi\cdot n)\times L})$ at the sink node within T seconds.

The ordinary achievable throughput [2], [5] is indeed 1-achievable. We say a Ψ-achievable throughput *asymptotically* 1-*achievable* if $\liminf_{n\to\infty} \Psi = 1$. We also directly call it *asymptotically achievable*. We call the ratio Ψ the *gathering efficiency* of a specific aggregation scheme $\mathscr{A}(n, L, \mathcal{S}(\Psi\cdot n))$.

3) Tradeoffs between Throughput and Gathering Efficiency: It is intuitive that there exists a tradeoff between the aggregation throughput $\Lambda(\lambda, n)$ and gathering efficiency $\Psi(n)$. Define such tradeoff as

$$\Phi(\lambda, n) = \Lambda(\lambda, n)\cdot\Psi(n).$$

Obviously, $\Phi(\lambda, n) = O(1)$. Particularly,

Definition 2: We say an aggregation scheme $\mathscr{A}(n, L, \mathcal{S}(\Psi\cdot n))$ *scalable* if

$$\Phi(\lambda, n) = \Lambda(\lambda, n)\cdot\Psi(n) = \Theta(1), \tag{1}$$

i.e., $\Lambda(\lambda, n) = \Theta(1)$ and $\Psi(n) = \Theta(1)$, where $\Lambda(\lambda, n)$ is the throughput derived by $\mathscr{A}(n, L, \mathcal{S}(\Psi n))$.

III. AGGREGATION SCHEMES FOR RANDOM WSNS

Two proposed aggregation schemes are both cell-based. To simplify the description, we recall a notion called *scheme lattice*.

Definition 3 (Scheme Lattice): Partition a square region $\mathcal{A} = [0, \mathfrak{a}]^2$ into a lattice consisting of square cells of side length $\mathfrak{l}$, we call the produced lattice *scheme lattice*, and denote it by $\mathbb{L}(\mathfrak{a}, \mathfrak{l}, \theta, \lambda)$, where $\theta \in [0, \frac{\pi}{4}]$ is the minimum angle between the boundaries of $\mathcal{A}$ and the sides of cells.

A. Single-Length-Hop (SLH) Aggregation Scheme

We design the scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi\cdot n))$ based on the scheme lattice $\mathbb{L}_1 = \mathbb{L}(\sqrt{n/\lambda}, \sqrt{z\cdot\ln n/\lambda}, 0, \lambda)$. For all cells in $\mathbb{L}_1$, the number of sensors inside each cell is w.h.p. within $[(1-\varepsilon_4)\cdot\ln n, (1+\varepsilon_5)\cdot\ln n]$, where ε_4 and ε_5 are some constants depending on z and are defined in Table.II. For simplicity, we ignore the details about the integer, and assume that the number of rows (or columns) $\sqrt{\frac{n}{z\cdot\ln n}}$ is always an integer, without changing on the results in order sense. Taking the cell in bottom left corner as the origin with a 2-dimensional index $(0,0)$, we give each cell in $\mathbb{L}_1$ an index in the order from left to right and bottom to top, i.e., the index of the cell in top right corner is (δ,δ), where $\delta = \delta(n) = \frac{\sqrt{n}}{\sqrt{z\cdot\ln n}} - 1$. We assume without loss of generality that the sink node s_0 is located in the cell (δ,δ). Choose randomly just one sensor from each cell as the *aggregation station*, we obtain a set, denoted by $\mathcal{B}$, consisting of $\frac{n}{z\cdot\ln n}$ sensors. Let $b_{i,j} \in \mathcal{B}$ denote the aggregation station in cell (i,j). Note that the sink node s_0 can be selected as the aggregation station of cell (δ,δ). Define a sequence of sets: $\mathcal{H}_{h,v} := \{b_{i,j} |\ (i \bmod 3 = h) \wedge (j \bmod 3 = v)\}$, where $h \in \{0,1,2\}$ and $v \in \{0,1,2\}$. Then, the aggregation scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi\cdot n))$ is as follows:

Algorithm 1: Horizontal Backbone Pipelined Aggregation

Input: L rounds of aggregated measurements at all aggregation stations.
Output: L rounds of aggregated data at all station $b_{i,\delta}$.
for $k = 1, 2, \cdots, L, L+1, \cdots, L+\delta-3$ **do**
  $k \to k'$;
  if $k > L$ **then** $L \to k$;
  else for $h = 0, 1, 2$ **do**
    for $v = 0, 1, 2$ **do**
      for $r = 1, \cdots, k$ **do**
        All $b_{i,j} \in \mathcal{H}_{h,v}$ are permitted to transmit;
        if *it holds that* $1 \leq j \leq \delta - 1$*, and (1)* $b_{i,j}$*,* $j \geq 1$*, has received the* r*-th round aggregated data from* $b_{i,j-1}$*, and (2)* $b_{i,j+1}$ *has not received the* r*-th round aggregated data from* $b_{i,j}$*,* **then** $b_{i,j}$ sends them to $b_{i,j+1}$;
        else if $j = 0$*, and* $b_{i,1}$ *has not received the* r*-th round aggregated measurements from* $b_{i,0}$*,* **then** $b_{i,0}$ sends them to $b_{i,1}$.
    $k' \to k$

• **Local Aggregation**: In each cell in $\mathbb{L}_1$, $\beta\cdot\ln n$ sensors are selected, if applicable, where

$$\beta = \max\{\Psi\cdot(1+\varepsilon_5),\ 1/\ln n\}. \tag{2}$$

L rounds of measurements from those sensors to be aggregated to the aggregation stations by a single hop; all transmissions are scheduled by a 4-TDMA scheme, as illustrated in Fig.2(a).

• **Horizontal Backbone Aggregation**: L rounds of data held by each aggregation station are aggregated to the adjacent aggregation stations in the order from left to right in a *pipelined fashion* (Algorithm 1); all transmissions are scheduled by a 9-TDMA scheme, as illustrated in Fig.2(b).

• **Vertical Backbone Aggregation**: L rounds of measurements held by each station in the δth-column are aggregated to the adjacent station in the order from bottom to top in a similar pipelined fashion to Algorithm 1; all transmissions are scheduled by a 3-TDMA scheme, as illustrated in Fig.2(b).

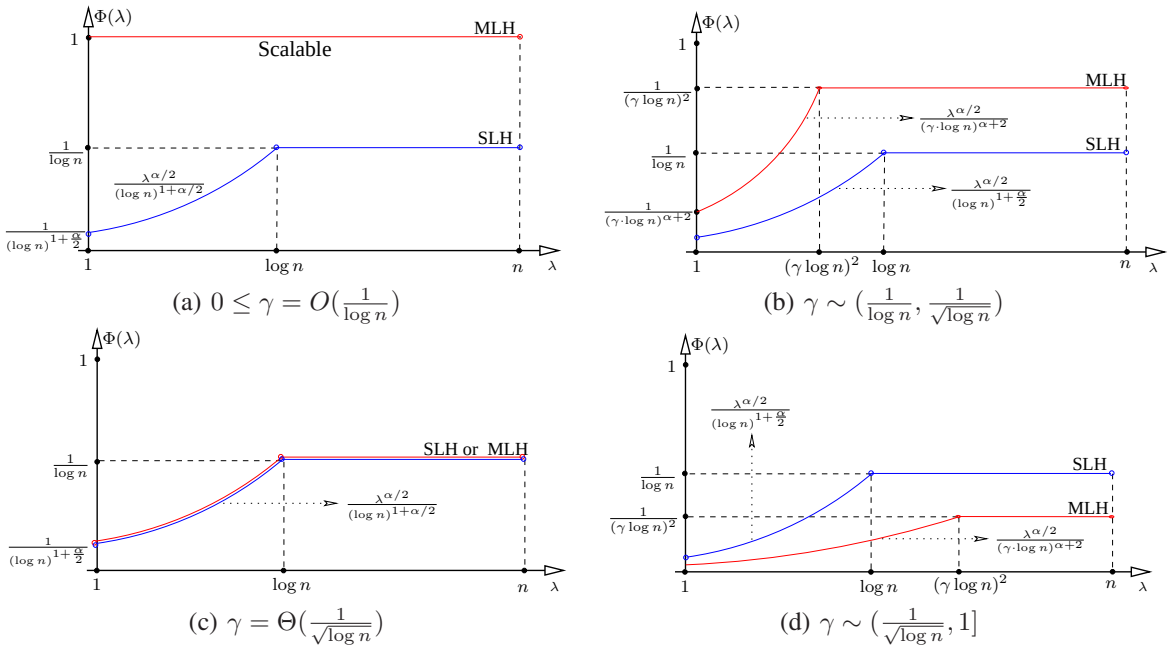


Fig. 1. Tradeoffs $\Phi(\lambda, n)$ between aggregation throughput $\Lambda(\lambda, n)$ and gathering efficiency $\Psi(n)$ for ordinary perfectly compressible functions under the schemes $\mathscr{A}_1(n, L, \mathcal{S}(\Psi n))$ and $\mathscr{A}_2(n, L, \mathcal{S}(\Psi n))$, denoted by **SLH** and **MLH**, respectively, where $\gamma = \max\left\{\frac{\Psi(n)}{1-\varepsilon_9} - \frac{\varpi}{\kappa},\ 0\right\}$, and the constant parameters ε_9, ϖ and κ are defined in Table.II.

B. Multiple-Length-Hop (MLH) Aggregation Scheme

We design another aggregation scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi \cdot n))$ based on the scheme lattice $\mathbb{L}_2 = \mathbb{L}(\sqrt{\frac{n}{\lambda}}, \frac{c}{\sqrt{\lambda}}, \frac{\pi}{4}, \lambda)$, where $c > 0$ is a constant and the specific value is determined in Table.II. Choose randomly a sensor from each nonempty cell, called *aggregation station*, then, we can build the *aggregation backbones* using the method in [6]. Please see the illustration in Fig.3. The *backbone stations*, i.e., the stations on the aggregation backbones, are connected by only *short* links, whereas every *peripheral station*, i.e., the stations other than aggregation stations, can access a specific aggregation station node in one-hop transmission.

For a given constant $\kappa > 0$, partition the scheme lattice $\mathbb{L}_2$ into horizontal (vertical) rectangle slabs with the horizontal (vertical) width of $\sqrt{\frac{n}{\lambda}}$ and the vertical (horizontal) width of

$$w_{\mathrm{R}} = (\kappa \ln m) \cdot c\sqrt{2/\lambda}, \tag{3}$$

where $m = \frac{\sqrt{n}}{\sqrt{2}c}$. We assume that $\frac{m}{\kappa \ln m}$, denoting the number of rectangle slabs, is an integer. Then, according to Theorem 5 of [6], we have

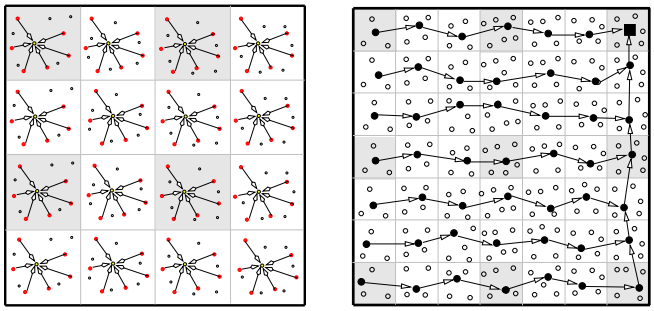
Lemma 1: For any constants c, κ satisfying $0 < \frac{2}{c^2 - \ln 6} < \kappa < \infty$, there exists a constant ϖ depending on κ and c such that for all horizontal (or vertical) slabs, there are w.h.p. at least $\varpi \cdot \ln m$ horizontal (or vertical) aggregation backbones, where $0 < \varpi < \frac{\kappa \cdot (c^2 - \ln 6) - 2}{\ln(1 - e^{-c^2}) + c^2}$.

When the aggregation backbones are built, the cells in each slab can be assigned averagely to $\varpi \cdot \ln n$ aggregation backbones. For instance, each slab is further divided into $\varpi \cdot \ln n$ slices, and each slice is mapped to a specific backbone. Anyway, the distance between a peripheral station to the corresponding backbone station is within $(0, w_{\mathrm{R}}]$. Now, we give the aggregation scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi \cdot n))$. The involved constants are all defined in Table.II.

- **Selection**: Choose a subset of aggregation stations, denoted by $\mathbb{C}(\Psi)$, that are at distance of at most $\gamma \cdot w_{\mathrm{R}}$ to the corresponding aggregation backbones, where

$$\gamma = \max\left\{\frac{\Psi}{1-\varepsilon_9} - \frac{\varpi}{\kappa},\ 0\right\}. \tag{4}$$

- **Local Aggregation**: In each cell in $\mathbb{C}(\Psi)$, choose randomly at most $\mathfrak{c} = \lceil (1+\varepsilon_6) \cdot c^2 \rceil$ sensors, if applicable, where $\varepsilon_6 > 0$ is defined in Table.II by letting $\sigma\lambda = c^2$; L rounds of measurements from those chosen sensors to be aggregated to the aggregation station by a single hop; all transmissions are scheduled by a 4-TDMA scheme based on $\mathbb{L}_2$.
- **Draining Aggregation**: All peripheral stations in $\mathbb{C}(\Psi)$ drain the L rounds of data into the corresponding backbone stations by a single hop of distance at most $\gamma \cdot w_{\mathrm{R}}$; all transmissions are scheduled by a K^2-TDMA scheme based on $\mathbb{L}_2$, where $K = 2 \cdot (\left\lceil \frac{\gamma \cdot w_{\mathrm{R}}}{c/\sqrt{\lambda}} \right\rceil + 1)$.
- **Horizontal Backbone Aggregation**: L rounds of data held by each backbone station are horizontally aggregated to the adjacent backbone stations in the order from left to right in a similar pipelined fashion to Algorithm 1, until the data are aggregated into the backbone stations on the backbones passed through by the sink node s_0, denoted by $\mathfrak{b}_{s_0}$; all transmissions are scheduled by a 9-TDMA scheme, as illustrated in Fig.2(b).
- **Vertical Backbone Aggregation**: L rounds of data held by each backbone station in the backbone $\mathfrak{b}_{s_0}$ are aggregated to the adjacent aggregation stations in the order from bottom to top in a similar pipelined fashion to Algorithm 1; all transmissions are scheduled by a 3-TDMA scheme.



(a) Selection and Aggregation (b) Aggregation Backbones

Fig. 2. Aggregation Scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi \cdot n))$. The shaded cells are simultaneously scheduled. (a)In each cell, $\beta \cdot \ln n$ sensors are selected, if applicable. (b) The black square is the sink node.

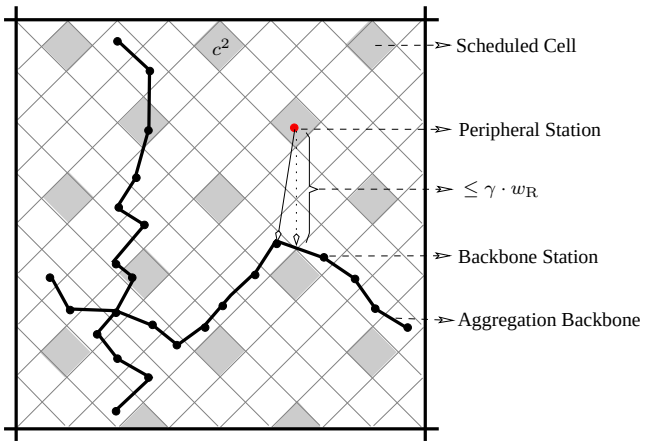


Fig. 3. Aggregation Backbone under Scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi \cdot n))$.

IV. Ψ-achievable Aggregation Throughput

A. Ψ-achievable Throughput under SLH Scheme

First, according to Step 1 of $\mathscr{A}_1(n, L, \mathcal{S}(\Psi \cdot n))$, the number of sensors is at least $\Psi \cdot n$, then, it is easy to get that

Lemma 2: Under the scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi \cdot n))$, the derived throughput is Ψ-achievable..

Theorem 1: Under the scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi \cdot n))$ with $L = \Omega(\frac{\sqrt{n}}{\sqrt{\log n}})$, the achievable throughput for perfectly compressible aggregation functions is of order

$$\Lambda_1(\lambda, n) = \begin{cases} \Omega(\frac{1}{\beta \cdot \log n}) & \text{when} \quad \lambda \sim [\log n, n] \\ \Omega(\frac{\lambda^{\frac{\alpha}{2}}}{\beta \cdot (\log n)^{1+\frac{\alpha}{2}}}) & \text{when} \quad \lambda \sim [1, \log n] \end{cases}$$

where β is defined in Equation (2).

As mentioned in Section II-B2, the ordinary achievable throughput [2], [5], is indeed the 1-achievable throughput under Definition 1. Giridhar and Kumar [5] designed an aggregation scheme under the *fixed-range protocol model* [3] by which the achievable throughput of a special case of random dense scaling WSNs, i.e., $\mathcal{N}(n, n)$, is of $\Omega(\frac{1}{\log n})$.

B. Ψ-achievable Throughput under MLH Scheme

First, we have the following result.

Lemma 3: Under the scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi \cdot n))$, the derived throughput is Ψ-achievable.

Theorem 2: Under the scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi \cdot n))$ with $L = \Omega(\sqrt{n})$, the Ψ-achievable throughput for perfectly compressible aggregation functions is of order

$$\Lambda_2(\lambda, n) = \begin{cases} \Omega(\frac{1}{(\gamma \cdot \log n)^2 + 1}) & \text{when} \quad \lambda \sim [(\gamma \log n)^2, n] \\ \Omega(\frac{\lambda^{\frac{\alpha}{2}}}{(\gamma \cdot \log n)^{\alpha+2} + \lambda^{\frac{\alpha}{2}}}) & \text{when} \quad \lambda \sim [1, (\gamma \log n)^2] \end{cases}$$

where γ is defined in Equation (4).

C. Tradeoffs between Throughput and Gathering Efficiency

Based on Theorem 1 and Theorem 2, we get that

Theorem 3: Under the schemes $\mathscr{A}_1(n, L, \mathcal{S}(\Psi n))$ and $\mathscr{A}_2(n, L, \mathcal{S}(\Psi n))$, the detailed tradeoffs are presented in Fig.1.

From Theorem 3, we have the following observations:

1) The SLH scheme $\mathscr{A}_1(n, L, \mathcal{S}(\Psi n))$ is not scalable.
2) Under the MLH scheme $\mathscr{A}_2(n, L, \mathcal{S}(\Psi n))$, when $\Psi(n) = (1 - \varepsilon_9) \cdot (\frac{\varpi}{\kappa} + O(\frac{1}{\log n}))$, the Ψ-achievable throughput is of order $\Theta(1)$, which means that the MLH scheme is indeed *scalable*.
3) When $\frac{\Psi(n)}{1-\varepsilon_9} - \frac{\varpi}{\kappa} = \omega(\frac{1}{\sqrt{\log n}})$, the tradeoff under the $\mathscr{A}_1(n, L, \mathcal{S}(\Psi n))$ is better than that under the $\mathscr{A}_2(n, L, \mathcal{S}(\Psi n))$ (Fig.1(d)); otherwise, it is not better, as illustrated in Fig.1(a)-(c).

V. Conclusion

We study the data aggregation of perfectly compressible functions for random WSNs with general density. We design two protocols, called SLH and MLH schemes, to derive the optimal aggregation throughput depending on a given gathering efficiency, and provide the optimal tradeoffs. Particularly, we show that the MLH scheme is scalable.

Acknowledgments

The research of authors is partially supported by the National Basic Research Program of China (973 Program) under grants No. 2010CB328101, No. 2011CB302804, and No. 2010CB334707, the Program for Changjiang Scholars and Innovative Research Team in University, the Shanghai Key Basic Research Project under grant No. 10DJ1400300, the NSF CNS-0832120 and CNS-1035894, the National Natural Science Foundation of China under grant No. 60828003 and No. 61003277, the Program for Zhejiang Provincial Key Innovative Research Team, and the Program for Zhejiang Provincial Overseas High-Level Talents.

References

[1] K. Fan, S. Liu, and P. Sinha, "Scalable data aggregation for dynamic events in sensor networks," in *Proc. ACM SenSys 2006*.
[2] T. Moscibroda, "The worst-case capacity of wireless sensor networks," in *Proc. ACM/IEEE IPSN 2007*.
[3] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. on Info. Theory*, vol. 46, no. 2, pp. 388–404, 2000.
[4] D. Marco, E. Duarte-Melo, M. Liu, and D. Neuhoff, "On the many-to-one transport capacity of a dense wireless sensor network and the compressibility of its data," in *IPSN 2003*.
[5] A. Giridhar and P. Kumar, "Computing and communicating functions over sensor networks," *IEEE JSAC*, vol. 23, no. 4, pp. 755–764, 2005.
[6] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. on Info. Theory*, vol. 53, no. 3, pp. 1009–1018, 2007.