# *CDC*: Compressive Data Collection for Wireless Sensor Networks

Xiao-Yang Liu[*], Yanmin Zhu[*], Linghe Kong[†*], Cong Liu[‡], Yu Gu[†], Athanasios V. Vasilakos[§], Min-You Wu[*]

[*]Shanghai Jiao Tong University, China
[†]Singapore University of Technology and Design, Singapore
[‡]University of Texas at Dallas, USA
[§]University of Western Macedonia, Greece

*Abstract*—**Data collection is a crucial operation of wireless sensor networks. The design of data collection schemes is challenging due to the limited energy supply and the hot spot problem. Leveraging empirical observations that sensory data possesses strong spatiotemporal compressibility, this paper introduces a novel compressive data collection scheme for wireless sensor networks. We adopt a power-law decaying data model verified by real data sets and propose a random projection-based estimation algorithm for this data model. Our scheme requires fewer measurements, even fewer number of sensor readings in each measurement, thus greatly reduces the energy consumption without introducing much computation and control overheads. Analytically, we prove that it provides the same order of estimation error with the optimal approximation. Evaluations on real data sets (from the GreenOrbs, IntelLab and NBDC-CTD projects) shows that compared with existing approaches, this new scheme almost doubles the network lifetime for estimation error within $20\%$.**

## I. Introduction

A crucial operation of wireless sensor networks (WSNs) [1] is to perform data collection [2], where sensor readings are collected from sensor nodes. Various applications rely on efficient data collection, such as battlefield surveillance [3], habitat monitoring [4], infrastructure monitoring [5] [6], and environmental monitoring [7][8].

A primary challenge faced by the design of data collection is in prolonging the network lifetime. First of all, each sensor node, being a micro-electronic device, can only be equipped with a limited power source while in many applications recharging is impractical. Thus, a WSN can only support limited traffic load. Even worse, the data information that a WSN can effectively transport might be even less since the network capacity decreases as the number of nodes increases [9]. Moreover, the many-to-one traffic pattern called *convergecast* [10], of data collection induces load unbalancing. It leads to the *hot spot* problem, i.e., the sensor nodes closer to the sink node will soon run out of energy and the network lifetime of WSNs will be significantly shortened. Furthermore, the unreliability of low-power wireless communication and the limited computational ability of sensor nodes make the design of effective and efficient data collection schemes even more challenging. As WSNs adopt low-power wireless communication, packet loss is a common problem [11][12]. Thus, transporting the sensor readings from sensor nodes to the sink requires significant efforts. On the other hand, sensor nodes can only support simple computing tasks, therefore the preprocessing or in-network processing of data collection schemes should take this hardware constraint into consideration.

Existing solutions have limitations and thus are unsatisfactory. Generally, data collection in WSNs follows two approaches: raw-data collection and aggregated-data collection. As WSNs are typically composed of hundreds to thousands of sensor nodes, generating tremendous amount of sensory data, raw-data collection is usually rather inefficient. Aggregated-data collection takes advantage of the correlations (or compressibility) within sensory data to reduce the communication cost. More specifically, in-network data compression [13] is adopted to reduce global traffic, such as distributed source coding [14][15] or transform coding [16][17]. However, they may incur significant computation and control overheads that are not suitable for WSNs.

The compressive data gathering (CDG) [18] exploits compressive sensing to reduce global scale communication cost without introducing intensive computation or complicated transmission control overheads while also achieving load balancing. However, it assumes that the routing tree is fixed and perfectly reliable. Although in simulation it behaves well, the practical performance is unsatisfactory.

*Main Contributions*: In this paper, we propose a novel compressive data collection scheme for wireless sensor networks. Firstly, based on three data sets, i.e., GreenOrbs [19] (mountain data), IntelLab [20] (indoor data) and NBDC-CTD [21] (ocean data), we reveal that there exists strong compressibility in these data sets and identify that the "power-law decaying" data model fits well for sensor network data. Secondly, an opportunistic routing is adapted to compress the sensory data "on the fly", i.e., compressing the sensor reading of a newly encountered node while the packets are forwarded back to the sink. Thirdly, we model the opportunistic routing as a Markov chain and calculate the compression probability of each sensor node. Fourthly, by regarding the random linear compression as nonuniform sparse random projections (*NSRP*) [22][23], we prove that NSRP-based estimator guarantees optimal error bound for the power-law decaying data model. Finally, based on real sensor data sets, we evaluate our scheme which prolongs the network lifetime by $1.5\times$ to $2\times$ for estimation error within $20\%$, compared with the baseline scheme and the CDG [18] scheme.

The reminder of the paper is organized as follows. In Section II, we present our empirical observation on real sensory data sets. In Section III, network models, a compressible data model with "fitness" verification and the design overview are described. We introduce the major design of our scheme in Section IV. Evaluation results are presented in Section V. Related work is discussed in Section VI, while we conclude in Section VII.

## II. OBSERVATIONS

The spatiotemporal correlation among the sensory data is a significant and unique characteristic of WSNs which can be exploited to dramatically enhance the overall network performance. To the best of our knowledge, we are the first to quantitatively investigate this kind of "compressibility" using the concepts of spatiotemporal marginal and conditional entropy.

Entropy measures the information contained in the data sets. The difference between spatial marginal entropy and conditional entropy measures the compression space one can exploit by jointly compressing data across nearby sensors, while the temporal counterpart measures the compression space by jointly compressing data across sequential slots.

### A. Data Sets

We extract three matrix subsets from the GreenOrbs [19], IntelLab [20] and NBDC-CTD [21] projects which are deployed in mountain, indoor and ocean environment, respectively. Their properties are summarized in Table I.

The monitoring period is evenly divided into $T$ time slots, denoted as $\{0, 1, ..., t, ..., T-1\}$. A record at the sensor node includes the reading, node ID, position (longitude and latitude), and time stamp. The format of a record is:

| Record: | reading | ID | position | time stamp |

Let $U_{i,t}$ denote the reading of the $i$-th node at slot $t$, it may be temperature, humidity, illumination, etc. Then, a physical condition (e.g., temperature) can be represented by a data matrix:

$$\mathbf{U} = \begin{bmatrix} U_{0,0} & ... & U_{0,t} & ... & U_{0,T-1} \\ U_{1,0} & ... & U_{1,t} & ... & U_{1,T-1} \\ ... & ... & ... & ... & \\ U_{n-1,0} & ... & U_{n-1,t} & ... & U_{n-1,T-1} \end{bmatrix} \quad (1)$$

where the $i$-th row is the $i$-th node's reading sequence, and the $t$-th column is the whole network's readings at slot $t$.

### B. Marginal Entropy and Conditional Entropy

*Discretization*: Since the sensory readings have real values, we discretize them in the following way:
- Get the range $[U_{min}, U_{max}]$ of $\mathbf{U}$;
- Divide this range into $Q$ equal sections $\{s_1, ..., s_k, ..., s_Q\}$ with each being called a *sensing state*, where $Q$ is set by the user.
- Construct a state matrix $S_{n \times T}$ for $\mathbf{U}$.

*Temporal Marginal Entropy*: The *temporal marginal entropy* is the entropy of a node's state sequence. Let $\sigma_k$ denote the occurrence frequency of state $s_k$ in the $i$-th row $S_i$, we have the probability $P(s_k)$ of each state and then the marginal entropy $H(S_i)$ as:

$$P(s_k) = \lim_{T \to \infty} \frac{\sigma_k}{T}$$
$$H(S_i) = -\sum_{k=0}^{Q-1} P(s_k) \cdot log_2 P(s_k) \quad (2)$$

*Temporal Conditional Entropy*: The *temporal conditional entropy* is defined as the entropy of $S_{i,t}$ when its immediately previous state $S_{i,t-1}$ is known, as:

$$H(S_{i,t}|S_{i,t-1}) = H(S_{i,t}, S_{i,t-1}) - H(S_{i,t-1}). \quad (3)$$

where $H(S_{i,t}, S_{i,t-1})$ is the joint entropy of two consecutive sensing states $(S_{i,t}, S_{i,t-1})$. Therefore, we have $n$ temporal marginal and conditional entropy. Similarly, one can compute the $k$-th order temporal conditional entropy of $S_{i,t}$ given its previous $k$ states $S_{i,t-k}, ..., S_{i,t-2}, S_{i,t-1}$.

*Spatial Marginal and Conditional Entropy*: The *spatial marginal entropy* is the entropy of $n$ nodes' sensing states at time slot $t$. Let $\sigma_k$ denote the occurrence frequency of state $s_k$ in the $t$-th column $S_t$, then $P(s_k)$ and $H(S_t)$ can be computed similarly as in Eqn.(2).

Let the $j$-th node be the $i$-th node's nearest neighbor. The *spatial conditional entropy* is defined as the entropy of $S_{i,t}$ when the $j$th node's sensing state $S_{j,t}$ is known, computed as:

$$H(S_{i,t}|S_{j,t}) = H(S_{i,t}, S_{j,t}) - H(S_{j,t}). \quad (4)$$

where $H(S_{i,t}, S_{j,t})$ is the joint entropy of two nearby nodes' sensing states $(S_{i,t}, S_{j,t})$. Therefore, we can have $n$ spatial marginal and conditional entropy. Similarly, one can compute the $k$-th order spatial conditional entropy of $S_{i,t}$ given its $k$ nearest neighbors' states.

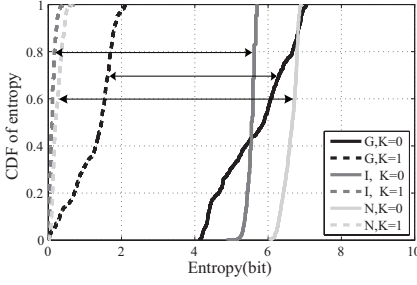### C. Spatiotemporal Compressibility

Here, we present the observations for temperature as a representative since all three data sets contain this physical condition. Similar results also hold for other conditions.

The cumulative distribution functions (CDFs) of temporal/spatial marginal entropy and conditional entropy are presented in Fig.1(a) and Fig.1(b). The marginal entropy of both cases is quite small, namely $5 \sim 8$ bits, indicating that it is not an efficient way to use the double type variable (64 bits) to store sensor readings as $\approx 90\%$ of them are wasted.
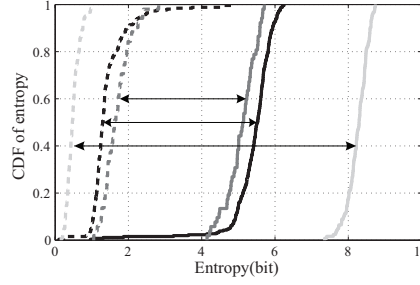
We use 1st order conditional entropy. For large order conditional entropy, the compression space will enlarge. For Fig.1(a), temporal conditional entropy is significantly smaller than the corresponding marginal entropy. Therefore the storage length can be greatly reduced by compressing a sensor's readings together with its previous readings. The temporal compressibility for GreenOrbs is weaker than that of the other two, possible reasons are: (1) the mountain region is much more variant; (2) the temperature in mountain region is more sensitive due to shielding effects caused by trees. From Fig.1(b), we know that the needed length can be greatly reduced by compressing a sensor's readings with its nearest

TABLE I
DATA SETS FOR THE COMPRESSIBILITY CHARACTERIZATION.

| Name | Environment | Time period | Physical conditions | Matrix subset size | Time interval |
|---|---|---|---|---|---|
| GreenOrbs [19] | Forest | Aug.03 ∼ 05, 2011 | Temperature, light, humidity | 326 nodes × 750 intervals | 5 minutes |
| IntelLab [20] | Indoor | Feb.28 ∼ Apr.5, 2004 | Temperature, light, humidity | 54 nodes × 500 intervals | 30 seconds |
| NBDC CTD [21] | Pacific Ocean | Oct.26 ∼ 28, 2012 | Temperature, salt, conductivity | 216 nodes × 300 intervals | 10 minutes |



(a) Temporal compressibility     (b) Spatial compressibility

Fig. 1. The spatiotemporal compressibility of temperature. $G, I, N$ stands for the GreenOrbs, IntelLab and NBDC-CTD data sets, $K = 0$ stands for marginal entropy while $K = 1$ for 1st-order conditional entropy.

Fig. 2. The *log-log* scaled graph for temperature in the GreenOrbs, IntelLab, NBDC-CTD data sets. $|\theta|_{(i)}$ denotes the $i$-th Fourier coefficient's magnitude, ordered in the decreasing manner.

neighbors' readings. Note that, the spatial regularity between nearby nodes is similar in indoor and mountain regions, while in ocean environment it is much stronger.

## III. SYSTEM MODEL AND DESIGN OVERVIEW

### A. Network Model

Consider a wireless sensor network consisted of $n$ sensor nodes and one sink. Sensor nodes are distributed in the target field to sense the physical conditions and report sensory data back to the sink through multi-hop transmissions. At the beginning, each sensor node generates a TOKEN with probability $p$. In expectation, there are $L = np$ TOKENs distributed across the WSN.

*Notations*: $U$ refers to either the data matrix or vector depending on the context, throughput the paper, $N = n \times T$, $U^{\mathcal{T}}$ is the transpose of matrix/vector $U$, $|\theta|$ denotes the magnitude of coefficient $\theta$, $\|U\|_2$ is the normal of vector $U$, $\Psi$ denotes the transform basis. A packet at the sink is called a measurement throughout the paper.

### B. Data Model

We consider a data vector $U \in \mathbb{R}^{N \times 1}$ (with $N = n \times T$), and fix an orthonormal transform $\Psi = [\Psi_1, \Psi_2, ..., \Psi_N] \in \mathbb{R}^{N \times N}$. $\Psi$ can be a wavelet or a Fourier transform basis. The coefficient vector $\theta = [U^{\mathcal{T}}\Psi_1, U^{\mathcal{T}}\Psi_2, ..., U^{\mathcal{T}}\Psi_N]^{\mathcal{T}}$ can be ordered decreasingly in terms of magnitude, such that $|\theta|_{(1)} \geq |\theta|_{(2)} \geq ... \geq |\theta|_{(N)}$.

*Power-law Decaying Data Model*: The coefficients' magnitude decays according to the power law [23][24][25], i.e., the $i$-th largest coefficient satisfies

$$|\theta|_{(i)} \leq Ci^{-1/\varpi}, \ i = 1, 2, ..., n \quad (5)$$

where $C$ is a constant and $-1/\varpi$ controls the compressibility of the data, i.e., larger $-1/\varpi$ implies faster decaying.
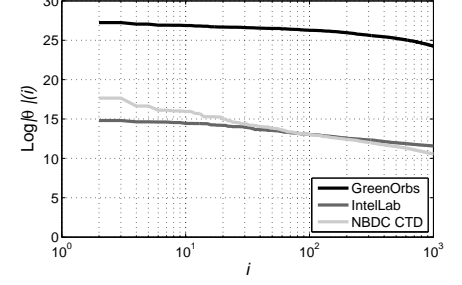
TABLE II
COMPRESSIBILITY PARAMETER $-1/\varpi$ OF OUR DATA SETS.

| Data name | $-1/\varpi$ | 95% confidence bound |
|---|---|---|
| IntelLab temperature | −0.3609 | [−0.3624, −0.3595] |
| IntelLab light | −0.7717 | [−0.7757, −0.7677] |
| IntelLab humidity. | −0.2326 | [−0.2342, −0.2310] |
| GreenOrbs temperature | −0.8218 | [−0.8265, −0.8171] |
| GreenOrbs light | −0.5545 | [−0.5565, −0.5525] |
| GreenOrbs humidity. | −0.6622 | [−0.6641, −0.6604] |
| NBDC-CTD temperature | −0.9797 | [−0.9830, −0.9763] |
| NBDC-CTD salt | −0.8363 | [−0.8380, −0.8346] |
| NBDC-CTD conductivity. | −0.8393 | [−0.8411, −0.8376] |

*Optimal Approximation*: The best $K$-term approximation is the optimal approximation for the power-law decaying data [23][24][25], i.e., keeping the largest $K$ coefficients and setting the others to be zero. The optimal estimation error bound is:

$$\|U - \hat{U}_{opt}\|_2^2 = \|\theta - \hat{\theta}_{opt}\|_2^2 = \eta_\varpi \|U\|_2^2 \quad (6)$$

where $\eta_\varpi$ is a constant that only depends on $-1/\varpi$.

We verify this data model by investigating the Fourier coefficients of these three data sets drawn in a *log-log* scaled graph, as shown in Fig.2. For coefficients decaying as in Eqn.(5), the magnitude and the rank is line with slop $-1/\varpi$. From Fig.2, we can see that the linear relation holds. Using the maximum likelihood estimation method [34], we measure the "compressibility" of each physical condition, as listed in Table II.

### C. Design Overview

The framework of our scheme is presented in Fig.3. It has two major components: an opportunistic routing and an estimator. The opportunistic routing is responsible for data compression and packet relaying. By modeling it as a Markov chain, the compression probability of each node can be estimated. A maximum likelihood estimation problem is set up
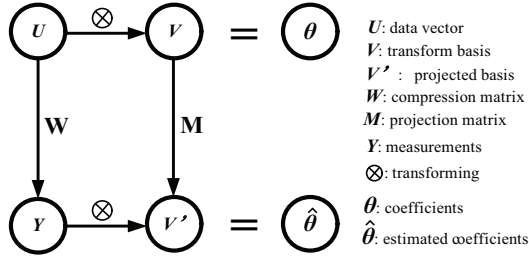
Fig. 3.   The framework of the compressive data collection scheme.



Fig. 4.   (a) Along $P_l$, the packet adds or subtracts the sensor reading of a newly encountered node. (b) The left process is modelled as: a sampling process $\mho(\cdot)$ randomly selects a subset of sensor readings, and a compression process sums them up with random coefficients chosen from the set $\{-1, +1\}$ to get one measurement.

for estimating the compression probability. Then, we prove that nonuniform sparse random projection (*NSRP*) preserves the inner product of two vectors and apply this property to design a simple but quite accurate estimator which guarantees the optimal error bound.

The compressive data collection scheme works in the following way:

- Data $U$ is stored in sensor nodes, with $\theta = U^{\mathcal{T}}V$.
- Each node with a TOKEN transmits one packet back to the sink.
- The packets are relayed according to the opportunistic routing. Compresion is performed at each new encountered node.
- At the sink, the compression process is modelled as $Y_{L \times 1} = W_{L \times N}U_{N \times 1}$, projecting the basis $V$ by $M$ to get $V'$. Then, $\hat{\theta} = Y^{\mathcal{T}}V'$ is an estimation of $\theta$.
- Finally, $\hat{U} = (\hat{\theta}V^{-1})^{\mathcal{T}}$, where $V^{-1}$ is the inverse of $V$.

***Key Issue***: The equation $\hat{\theta} = Y^{\mathcal{T}}V'$ is possible only when $L \geq N$, which is not energy efficient as it equals to transmit $N$ packets with each packet containing one reading. In the following we will show how to exploit the compressibility of the data vector $U$ by designing an estimation algorithm for the case $L \ll N$.

## IV. DESIGN

### A. Opportunistic Routing with Compression

The opportunistic routing has two parts: packet forwarding and data compression. We first define a *data collection path* $P_l$, and then the compression process of the $l$-th packet along this path.

***Packet Forwarding***: For node $s_i$, we define a *nearer neighbor set* $\mathcal{N}(i) = \{j|d(j, sink) \leq d(i, sink)\ \&\ d(i,j) \leq R_c\}$ where $R_c$ is the communication range. When a packet arrives at node $s_i$, $s_i$ compresses its sensory reading into the packet and then sends it out according to the opportunistic routing [26][27][28], i.e., passing on the packet to one of its nearer neighbors $s_j \in \mathcal{N}(i)$.

***Data Collection Path***: The trajectory of the $l$-th packet from a source node to the sink is called a *data collection path*, denoted as:

$$P_l = \langle p_0, p_1, ..., p_{\rho_l} \rangle \tag{7}$$

with $p_{\rho_l} = sink$, i.e., the packet travels across $\rho_l$ sensor nodes before it reaches the sink node.

Since opportunistic routing is adopted, the data collection paths are dynamic. It will bring about good features: energy
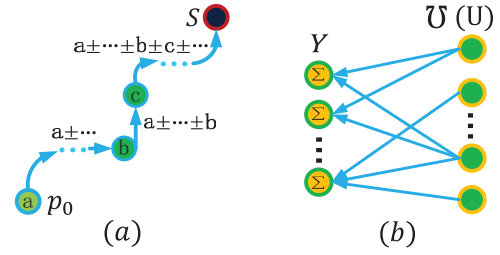
balancing and security, as this non-deterministic data collection path will mitigate the attacks on the routing path and balance the energy consumption.

***Data Compression***: As the packet travels towards the sink, the compression scheme linearly compresses the sensor reading of a newly encountered node with a random coefficient $r_i$, as shown in Fig.4(a). The format of the data packet is:

| Packet: | value | ID list | coefficient list | time stamp list |
|---|---|---|---|---|

Let $u_i$ ($i = 0, 1, ..., \rho_l - 1$) be the reading of the $i$-th node along path $P_l$. The data compression is performed as following:

***Step 1***: A node with the $l$-th TOKEN becomes node $p_0$ of $P_l$. It generates a packet containing data $y_0 = \pm u_0$, then transmits the packet to one of its nearer neighbors according to the opportunistic routing.

***Step 2***: The packet arrives at sensor $s_i$, sensor $s_i$ adds/substracts the data with probability $1/2$ as:

$$y_i = y_{i-1} + r_i u_i \tag{8}$$

Sensor ID, the coefficient and the current time slot is added to the packet's header. Then $s_i$ transmits it to one of its neighbors closer to the sink according to the opportunistic routing.

***Step 3***: The encoding process continues along path $P_l$ until the packet reaches the sink.

From the proof of Lemma. 1 (in Section IV. B), we know that the random coefficients $r_i$ can be real values in $[-1, 1]$ or chosen from $\{-1, +1\}$. We use the set $\{-1, +1\}$ because the nodes will only perform addition or substraction.

In the end of the data compression process, $L = np$ packets are collected by the sink. Next, we consider the sink's strategy to estimate the sensor readings using the collected packets.

### B. Problem Formulation for Estimation

***Traditional Compressive Sensing Approach***: In traditional compressive sensing approach [18][29][30], the sink establishes the following equations:

$$Y_{L \times 1} = A_{L \times N}U_{N \times 1} \tag{9}$$

where $A$ is an matrix with elements corresponding to $r_i$ in Eqn.(8). The sink can extract $A$ from the collected packets' headers.

Assuming that data $U$ is *sparse*. To be exact, there exists a transform basis $V_{N \times N}$ under which $U_{N \times 1}$ can be represented using $K \ll n$ nonzero coefficients. Compressive sensing (CS) [31][32] claims that, with probability at least $1 - N^{-\gamma}$ ($\gamma$ is set to be large), $U$ can be reconstructed exactly as the solution to the following $\ell_1$-minimization problem, i.e., $U = (\hat{\theta} V^{-1})^{\mathcal{T}}$ where

$$\hat{\theta} = \min_{\theta} \sum_{i=1}^{N} |\theta_i|, \quad s.t. \quad Y_{L \times 1} = A_{L \times N} U_{N \times 1} \tag{10}$$

$$with \ L = O(K \mu_{(A,V)}^2 log \frac{N}{K}), \ \mu_{(A,V)} = \max_{1 \le i, j \le N} |A_i^T V_j|$$

Compressive sensing suggests to compress WSNs' data as in [17][18]. Compressive sensing allows to use convex optimization to estimate sensory readings under the condition of the $RIP$ property [31][32], i.e., to decouple the matrix $A$ and the basis $V$ in order to have small value of $\mu$. However, the $RIP$ property does not hold for the above opportunistic routing, which postponed its utilization [30][33].

**Problem Formulation**: We take another approach by regarding the compression process as *nonuniform sparse random projections* (shown in Fig.4(b)), modeled as two mutually independent processes *nonuniform sampling process* $\mho(\cdot)$ and *linear encoding* $A$ as:

$$\mho(U_j) = \begin{cases} U_j, & prob. \ \pi_j \\ 0, & prob. \ 1 - \pi_j \end{cases}, \quad A_{ij} = \begin{cases} 1, & prob. \ \frac{1}{2}, \\ -1, & prob. \ \frac{1}{2}. \end{cases} \tag{11}$$

where $\pi_j \ne 0$, $j \in 1, 2, ..., N$ corresponds to the chance of $U_j$ being compressed in the collected packets. Thus, we model the compression scheme as:

$$Y_{L \times 1} = A_{L \times N} \mho(U_{N \times 1}) \tag{12}$$

Our problem becomes:

$$\hat{\theta} = \min_{\theta} \|U - \hat{U}\|_2^2$$
$$s.t. \quad Y_{L \times 1} = A_{L \times N} \mho(U_{N \times 1}) \tag{13}$$
$$where \ \hat{U} = (\hat{\theta} V^{-1})^{\mathcal{T}}.$$

The above opportunistic routing provides load balancing in the cost of nonuniform compression probability of each node. This leads to the failure of the traditional compressive sensing approach. We will introduce a new estimation algorithm in Section IV.E with the information of compression probability of each node. In the next subsection, we show how to estimate the compression probability.

### C. Compression Probability Estimation

**Modeling the Opportunistic Routing as a Markov Chain**: The packet forwarding according to the opportunistic routing can be modeled as a Markov Chain: the states are the nodes, the forwarding probability constitutes the transition probability matrix whose entry specifies the probability that a packet is transmitted from one node to one of its neighbors. First, we estimate the transition matrix $P$ based on the "incomplete observation" version of maximum likelihood estimation. Once the transition matrix $P$ is known, the compression probability $\pi$ can be derived.

**The Incomplete Observation Problem**: For every pair of nodes, assume that there is a *transition link* and this link has two states: *ON* and *OFF*. If we can obtain the complete observation of all links' state in every time slot, then *the estimation of the transition matrix is to maximize the log-likelihood of the posterior probability*, i.e., $log \ P((P_1, P_2, ...P_L)|P)$ with $(P_1, P_2, ...P_L)$ denotes the data collection paths of the $L$ collected packets. Maximum likelihood estimation (MLE) is the most used routine to solve this problem [34].

However, the observation of a transition link's state is to try a "packet-transmitting test", which is not practical. Furthermore, as our data collection aims to minimize the number of packets transmitted, the data collection paths recorded in headers of the collected packets are "incomplete" or under-sampled. Therefore, the traditional MLE scheme can not be used here.

**MLE with Incomplete Observation**: Here, we adopt an "incomplete observation" version of MLE proposed in [35]. Let $O_{ijt}$ denote the number of observed transition from node $s_i$ to node $s_j$ occurring over $t$ time slots and $(P^t)_{ij}$ the $ij$-th element of the matrix $P^t$ (the probability of a packet in node $s_i$ arrives at node $s_j$ after $t$ time slots), this new MLE is defined as:

$$\hat{P} = \max_{P} log \ P((P_1, P_2, ...P_L)|P)$$
$$= \max_{P} \sum_i \sum_j \sum_t O_{ijt} log(P^t)_{ij}. \tag{14}$$

A *Expectation Maximization Algorithm* is proposed to solve the above maximization problem. Refer to [35] for details and [36] for MATLAB codes.

**Estimation of the Compression Probability $\pi$**: The compression probability closely relates with the Markov chain-like occurrence, except that we should only count one time if the packet stay at a node waiting for transmission. The estimation algorithm is described as following:

**Step 1**: Set the initial probability $\pi_0 = \{L/n, L/n, ...., L/n\}$;
**Step 2**: Obtain $\bar{P}$ by setting the diagonal element of $P$ to zero;
**Step 3**: Calculate the expected occurrence frequency of nodes in data collection paths after $T$ time slots as:

$$O_i(T) = \sum_{i=0}^{T-1} \pi_0 \bar{P}^i; \tag{15}$$

**Step 4**: Average the occurrence frequency among the expected number of packets $\sum \pi_0$, (this can be regarded as a normalization), therefore we get the probability distribution

$$\pi_i = \frac{O_i(T)}{\sum \pi_0} = \frac{\sum\limits_{i=0}^{T-1} \pi_0 \bar{P}^i}{\sum \pi_0}, \tag{16}$$

where the summation and division above are a element-wise operation on row vectors.

### D. Nonuniform Sparse Random Projection

Eqn.(12) equals to the following linear equations:

$$Y_{L \times 1} = W_{L \times N} U_{N \times 1} \tag{17}$$

$$W_{ij} = \begin{cases} +1, & prob.\ \frac{1}{2}\pi_j \\ 0, & prob.\ 1-\pi_j \\ -1, & prob.\ \frac{1}{2}\pi_j \end{cases} \qquad (18)$$

*Sparse* and *nonuniform* raise from the fact that the opportunistic routing will neither pass all sensor nodes nor with equal probability, which is the case for most existing routings. *Sparse* allows the collected packets to compress sensor readings from a random selected subset, while traditional compressive sensing approaches require to compress all sensory readings together or the subset are randomly selected with equal probability [17][18].

Correspondingly, we construct a projection matrix $M \in \mathbb{R}^{L \times N}$ (where $L \ll N$) containing entries

$$M_{ij} = \frac{1}{\pi_j} \begin{cases} +1, & if\ W_{ij} = +1 \\ 0, & if\ W_{ij} = 0 \\ -1, & if\ W_{ij} = -1 \end{cases} \qquad (19)$$

The entries within each row are mutual-independent, while the entries across different rows are fully independent. In expectation, each row contains $\sum_{j=1}^{n} \pi_j$ nonzero elements, i.e., there are on average $\sum_{j=1}^{n} \pi_j$ sensor readings compressed in one collected packet.

Next, we prove that with high probability, nonuniform sparse random projections preserve inner products with predictable error in Lemma 1. Therefore, using only their random projections, we are able to estimate the inner product of two vectors. Please refer to the appendix for the detailed proof.

**Lemma 1** *For any vectors $U, V \in \mathbb{R}^{N \times 1}$, and $W, M \in \mathbb{R}^{L \times N}$ in Eqn.(18)(19). The random projections $Y = \frac{1}{\sqrt{L}} WU$, $V' = \frac{1}{\sqrt{L}} MV$, with the expectation and variance satisfying:*

$$E[Y^{\mathcal{T}}V'] = U^{\mathcal{T}}V \qquad (20)$$

$$Var(Y^{\mathcal{T}}V') \leq \frac{1}{L}((U^{\mathcal{T}}V)^2 + \xi\|U\|_2^2 \cdot \|V\|_2^2 + (\kappa - 2 - \xi)\sum_{j=1}^{N} U_j^2 V_j^2) \qquad (21)$$

*where $\xi = max(\frac{\pi_L}{\pi_m})$, $\kappa = \frac{1}{min(\pi)}$ denote the degree of nonuniform and the expected times to sample the "rarest" node, respectively.*

### E. NSRP-based Estimator

The intuition for our estimator design is that *nonuniform sparse random projections* preserve inner products within a small error. Hence we can use random linear measurements $Y = WU$ of the original data, and random linear projections $V' = MV$ of the orthonormal basis, to estimate the coefficients vector $\theta$.

The estimator works in the following way:
**Step 1**: Extract from the collected packets' headers to get $W_{L \times N}$ and $Y_{L \times 1}$, then construct the projection matrix $M$.
**Step 2**: Set $L_1 = C_1 \frac{1+\xi+\kappa H^2}{\epsilon^2}$ and $L_2 = C_2(1+\gamma)\log N$ such that $L = L_1 L_2$;
**Step 3**: Partition $Y_{L \times 1}$ into $L_2$ column vectors

$\{Y_1, Y_2, ..., Y_{L_2}\}$ with each of size $L_1 \times 1$, partition $M$ into $\{M_1, M_2, ..., M_{L_2}\}$ with each of size $L_1 \times N$, then project the basis $V$ to get $\{V'_1 = \frac{1}{\sqrt{L_1}} M_1 V, \cdots, V'_{L_2} = \frac{1}{\sqrt{L_1}} M_{L_2} V\}$;
**Step 4**: Compute $\zeta_l = Y_l^{\mathcal{T}} V'_l$, $l = 1, 2, \cdots, L_2$. Set each element of $\hat{\theta}$ as the median value of each column vector $\zeta_1, ..., \zeta_{L_2}$;
**Step 5**: Keep the $K$ largest coefficient in $\hat{\theta}$ and set the remaining to zero;
**Step 6**: Return $\hat{U} = (\hat{\theta}V^{-1})^{\mathcal{T}}$.

The following two theorems holds for the above estimator, please refer to the appendix for the detailed proof.

**Theorem 1** *For data vector $U \in \mathbb{R}^{N \times 1}$ satisfying*

$$\frac{\|U\|_\infty}{\|U\|_2} \leq H. \qquad (22)$$

*Let $V = \{V_1, ..., V_N\}$ be the transform basis with each vector $\mathbb{R}^{N \times 1}$, $W, M \in \mathbb{R}^{L \times N}$ as in Eqn.(18)(19) with the compression probability $\pi$, $L =$*

$$\begin{cases} O(\frac{1+\gamma}{\epsilon^2}(\xi + \kappa H^2)logn), & if\ (\xi + \kappa H^2) > \Omega(1) \\ O(\frac{1+\gamma}{\epsilon^2}logn), & if\ (\xi + \kappa H^2) \leq O(1) \end{cases} \qquad (23)$$

*Then, with probability at least $1 - N^{-\gamma}$, the random projections $Y = \frac{1}{\sqrt{L}} WU$ and $V'_i = \frac{1}{\sqrt{L}} MV_i$ can produce an estimate $\hat{\theta}_i$ for $U^{\mathcal{T}} V_i$ (Step 4 in the above estimator), satisfying*

$$|\hat{\theta}_i - U^{\mathcal{T}}V_i| \leq \epsilon\|U\|_2^2 \cdot \|V_i\|_2^2 \qquad (24)$$

*for all $i = 1, 2, ..., N$.*

**Theorem 2** *Suppose data $U \in \mathbb{R}^{N \times 1}$ satisfies condition (22), $W, M \in \mathbb{R}^{L \times N}$ in Eqn.(18)(19) with probability distribution $\pi$, and $L =$*

$$\begin{cases} O\left(\frac{1+\gamma}{\epsilon^2 \eta^2}(\xi + \kappa H^2)K^2 logn\right), & if\ (\xi + \kappa H^2) \geq \Omega(1) \\ O(\frac{1+\gamma}{\epsilon^2 \eta^2}K^2 logn), & if\ (\xi + \kappa H^2) \leq O(1). \end{cases} \qquad (25)$$

*Let $Y = \frac{1}{\sqrt{L}} \mathcal{W}U$, consider an orthonormal transform $\mathbf{\Psi} \in \mathbb{R}^{L \times N}$ and the corresponding transform coefficients $\theta = \mathbf{\Psi}U$. If the $K$ largest transform coefficients in magnitude gives an approximation with error $\|U - \hat{U}_{opt}\| \leq \eta\|U\|_2^2$, then given only $Y, W, M$ and $\mathbf{\Psi}$, one can produce an estimate $\hat{U}$ with error*

$$\|U - \hat{U}\| \leq (1+\epsilon)\eta\|U\|_2^2 \qquad (26)$$

*with probability at lest $1 - N^{-\gamma}$.*

For the above estimator, Theorem. 1 states that with high probability, the nonuniform sparse random projections of data vector and any projected basis vector can produce estimates of their inner products within a small error. Thus we can use the random projections of the data and the set of orthonormal basis to estimate the corresponding transform coefficients of the data. Theorem. 2 shows that with high probability, nonuniform sparse random projections can approximate compressible data with error comparable to the optimal approximation. Thus, with high probability, the above estimator produces an estimate of the original data within a small error.
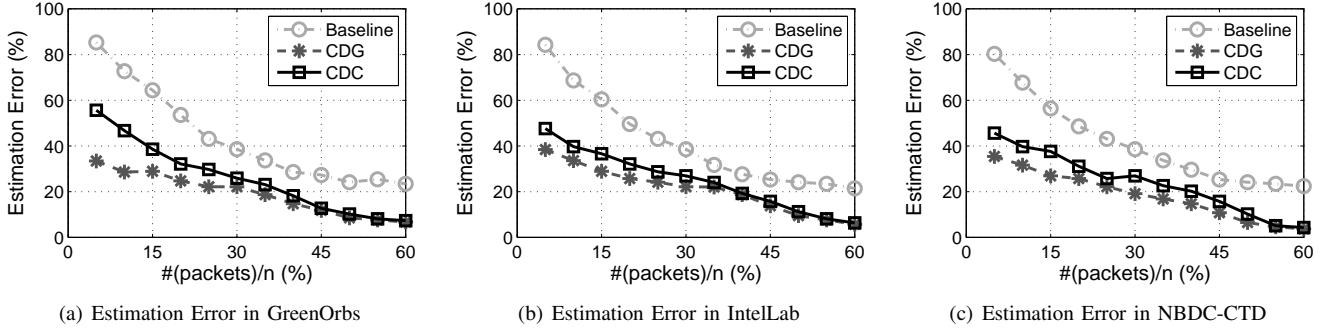
(a) Estimation Error in GreenOrbs

(b) Estimation Error in IntelLab

(c) Estimation Error in NBDC-CTD

Fig. 5. The estimation error for Basline, CDG, and CDC in the GreenOrbs, IntelLab and NBDC-CTD projects.
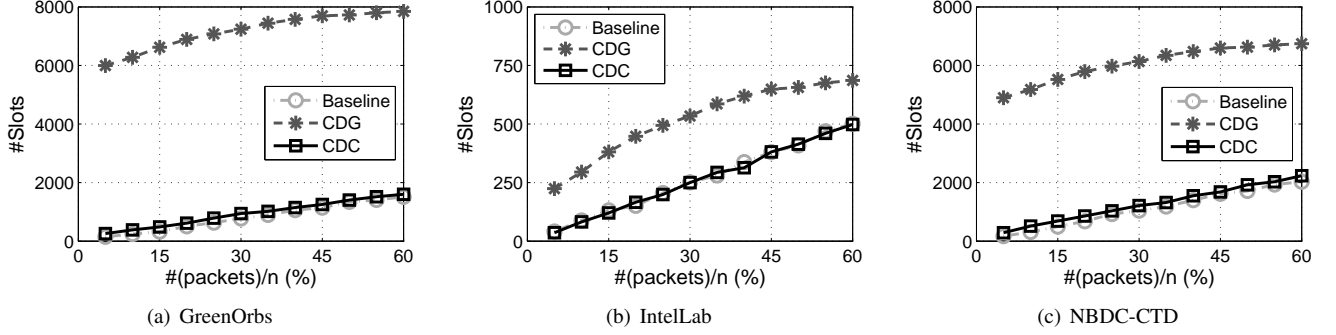


(a) GreenOrbs

(b) IntelLab

(c) NBDC-CTD

Fig. 6. The delay for Basline, CDG, and CDC in the GreenOrbs, IntelLab and NBDC-CTD projects.



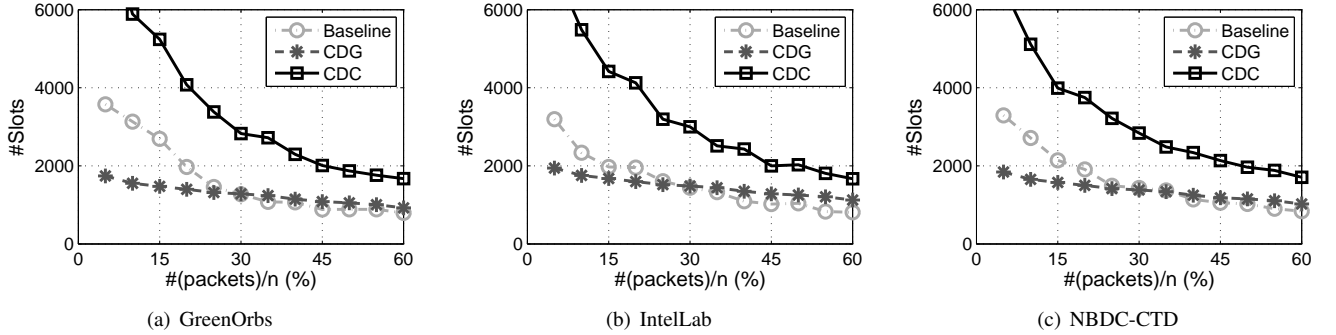(a) GreenOrbs

(b) IntelLab

(c) NBDC-CTD

Fig. 7. The network lifetime for Basline, CDG, and CDC in the GreenOrbs, IntelLab and NBDC-CTD projects.

## V. EVALUATION

### A. Experiment Settings

***Sensory Data***: The GreenOrbs [19], IntelLab [20] and NBDC-CTD [21] projects provide us the sensory data sets as described in Table. I. The WSNs continuously generate sensory readings from these data matrices in each slot.

***Network Topology***: Nodes' positions in these three WSNs are also provided. For GreenOrbs, the actual network topology can be reconstructed based on the neighbor set of each node. For IntelLab and NBDC-CTD, since such information is absent, we set the communication range to be $6.5m$ (indoor) and $180m$ (ocean) respectively, which ensures the network connectivity.

### B. Compared Algorithms

***Baseline***: Packets are transmitted back to the sink along the shortest path. Then the sink applies the $k$-Nearest Neighbors

(KNN) [38] method to estimate the readings, i.e., by averaging the $k$-nearest neighbors' values. Both the routing and estimation are the most basic ones, therefore we use it as the baseline algorithm.

***CDG*** (MobiCom'09): The CDG scheme compresses all sensor readings together in each collected packet. It uses the following tree-based routing: a node waits for all its children's packets, performs random liner compression, and then sends the packet to its parent node. The estimation uses the convex optimization method of traditional compressive sensing theory. It is a bit different from [18] as the link quality is not perfectly reliable, because we allow the transmission to fail and introduce the retransmission mechanism.

***CDC***: The CDC scheme is described in Section IV. More accurate link quality model will lead to more realistic simulation results. The *RSSI* value is the best indicator for link quality. However, this information may not be always known to

the network protocol designer, then the distance will serve an acceptable choice. We adopt the following link quality models of the opportunistic routing for simulation:

- The *RSSI* value between two neighbor nodes is given for the GreenOrbs project, we use a *RSSI-LinkQuality* model [26] to control the successful probability of transmitting a packet.
- For the IntelLab and NBDC-CTD projects such information is absent, we use a *Distance-LinkQuality* model [27] instead based the Euclidean distance to the sink.

*C. Metrics*

Based on the network topology and sensory data sets of these three wireless sensor networks, we run the above three schemes 10 times. For the baseline algorithm and the CDC scheme, we vary the probability $p$ of generating TOKENs to get different number of random measurements. For equality, the CDG scheme will collect accordingly the same number of random measurements.

***Estimation Error***: Each algorithm estimates a $\hat{U}$ for the original data $U$. The estimation error is defined as:

$$e = \frac{\|U - \hat{U}\|_2}{\|\hat{U}\|_2} \qquad (27)$$

***Delay***: The data collection delay is defined as the time when the last packet arrives at the sink, which is measured in terms of the number of slots.

***Network Lifetime***: The energy consumption is set according to the energy consumption model [39]. At the beginning, each node have initial energy of $1,000,000$ units which can support the sensor node to run about a month. The network lifetime is defined as the time when the first node runs out of energy, which is also measured in terms of the number of slots.

*D. Results*

From Fig. 5(a)(b)(c), CDG and CDC perform much better than the baseline algorithm and can reach error as low as $5\%$. This is because they both exploit the compressibility nature of the sensor readings and use random compression techniques. However, CDG behaves better in situations where less number of packets are collected. Possibly, less collected packets means: (1) stronger nonuniform nature of the compression probability, or (2) less observations of the routing process, then less accurate of probability estimation in Section IV.C.

From Fig. 6(a)(b)(c), it is quite unexpected that the delay of the CDG scheme is several or even hundreds of times longer than the other two. We analyze the possible reasons behind this case: CDG tries to encode every node's packets, and a parent node has to wait for all its children's packets before transmitting the compressed packet to its own parent node. Because the network size of the IntelLab project is smaller, the delay performance is closer for these three schemes. The baseline algorithm exhibits a good, stable, and moderate growth in delay since it used the shortest path routing. The CDC's routing strategy is quite similar with the baseline scheme, therefore it experiences quite similar performance.

From Fig. 7(a)(b)(c), we find that our scheme has the best performance. For estimation error within $20\%$, CDC prolongs the network lifetime by $1.5\times \sim 2\times$. This is because that CDC requires fewer measurements, even fewer number of sensory readings in each measurement, thus greatly reduces the energy consumption.

## VI. RELATED WORK

Energy conservation [40] is an import issue in wireless sensor networks. In data collection, in-network compression is a promising approach to reduce the amount of information to be transmitted by exploiting sensory data's redundancy. For detailed information of in-network compression techniques, please refer to [13].

According to how the sensory data is compressed, we classified existing data collection schemes into three categories: conventional compression, distributed source coding, and compressive sensing.

***Conventional compression***: Conventional compression techniques assume specific data structures and thus quire explicit data communication among sensor [5][40]. In joint entropy coding approach, nodes may use relayed data as side information to encode their readings. If the data are allowed to be communicated back and forth during encoding, sensor nodes may cooperatively perform perform transform to better utilize the correlation, such as the gossip-based technique use in [17]. There are two main problems with this approach. First, the route heavily influences the compression performance [13]. To achieve high compression ratio, data compression and packet routing are required to optimized jointly, which is proved to be NP-hard [41][42]. Second, structure-aware data compression induces computational and communication overheads [13] [29][30], rendering this kind of data collection schemes to be inefficient.

***Distributed source coding***: Distributed source coding intends to reduce complexity at sensor nodes and utilized correlation at the sink [14][15]. After encoding sensor readings independently, each node simply sends the compressed message along the shortest path to the sink [7]. Distributed source coding performs well for static correlation patterns. However, when correlation pattern changes or abnormal readings show up, the estimation accuracy will be greatly affected.

***Compressive sensing***: Recently, compressive sensing gains increasing attention in wireless sensor networks [16][17][18]. In both static and mobile sensor networks [29], the interplay of routing with compressive sensing is a key issue [33]. Some of them conclude that although sparsity exists in the environment, but the strict property is required by traditional compressive sensing decoder, hardly good approximation can be achieved as claimed by the theory. Then some proposed network-layer compression [23] to avoid this kind of problem. Our scheme adopt opportunistic routing with quite simple compression, therefore the data collection process is dynamic. This dynamic feature lead to energy balancing and finally benefits energy consumption.

## VII. CONCLUSION AND FUTURE WORK

We have proposed a novel compressive data collection scheme for wireless sensor networks. This scheme leverages the fact that raw sensory data have strong spatiotemporal compressibility. Our scheme consists of two parts: the opportunistic routing with compression, and the nonuniform random projection based estimation. The proposed scheme agrees with Braniuk's [2] suggestion that sensor data acquisition should be more efficient and new techniques that combine sensing and network communication together is a promising approach. We prove that this scheme can achieve optimal approximation error, and trace based evaluation show that its error is comparable with the existing method [18]. More important, our scheme exhibits good performance for energy-conservation.

The degree of nonuniform compression has direct relation with the approximation error. But we do not know its exact influence since in the formulas there are unknown constants governing the error bound. We would characterize this in the future.

## REFERENCES

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey", *Elsevier Computer Networks*, Vol. 38, No. 4, pp. 393-422, 2002.

[2] R.G. Baraniuk, "More is less: Signal processing and the data deluge", *Science*, Vol. 331, No. 6018, pp. 717-719, 2011.

[3] T. Bokareva, W. Hu, S. Kanhere, B. Ristic, N. Gordon, T. Bessell, M. Rutten, and S. Jha, "Wireless sensor networks for battlefield surveillance", *Proc. the land warfare conference*, 2006.

[4] R. Szewczyk, A. Mainwaring, J. Polastre, J. Anderson, and D. Culler, "An analysis of a large scale habitat monitoring application", *ACM Sensys*, pp. 214-226, 2004.

[5] N. Xu, S. Rangwala, K.K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A wireless sensor network for structural monitoring", *ACM Sensys*, pp. 13-24, 2004.

[6] S. Kim, S. Pakzad, D. Culler, J. Demmel, G. Fenves, S. Glaser, and M. Turon, "Health monitoring of civil infrastructures using wireless sensor networks", *ACM/IEEE IPSN*, pp. 254-263, 2007.

[7] L. Selavo, A. Wood, Q. Cao, T. Sookoor, H. Liu, A. Srinivasan, Y. Wu, W. Kang, J. Stankovic, D. Young, and J. Porter, "Luster: wireless sensor network for environmental research", *ACM Sensys*, pp. 103-116, 2007.

[8] L. Kong, D. Jiang, M.-Y. Wu, "Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction", *IEEE ICDCS*, pp. 179-188, 2010.

[9] P. Gupta and P.R. Kumar, "The capacity of wireless networks", *IEEE Transactions on Information Theory*, Vol. 46, No. 2, pp. 388-404, 2000.

[10] L. Fu, Y. Qin, X. Wang, and X. Liu, "Converge-cast with mimo", *IEEE INFOCOM*, pp. 649-657, 2011.

[11] L. Kong, M. Xia, X.-Y. Liu, M. -Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks", *IEEE INFOCOM*, pp. 1702-1710, 2013.

[12] F. Fazel, M. Fazel, and M. Stojanovic, "Random access compressed sensing for energy-efficient underwater sensor networks", *IEEE Journal Selected Areas in Communications (JSAC)*, 29(8), pp. 1660-1670, 2011.

[13] E. Fasolo, M. Rossi, J. Widmer, and M. Zorzi, "In-network aggregation techniques for wireless sensor networks: a survey", *IEEE Wireless Communications*, Vo. 14, No. 2, pp. 70-87, 2007.

[14] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "On network correlated data gathering", *IEEE INFOCOM*, pp. 2571-2582, 2004.

[15] J. Chou, D. Petrovic, and K. Ramachandran, "A distributed and adaptive signal processing approach to reducing energy consumption in sensor networks", *IEEE INFOCOM*, pp. 1054-1062, 2003.

[16] J. Haupt, W.U. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data", *IEEE Signal Processing Magazine*, Vol. 25, No. 2, pp. 92-101, 2008.

[17] H. Zheng, S. Xiao, X. Wang, and X. Tian, "Energy and latency analysis for in-network computation with compressive sensing in wireless sensor networks", *IEEE INFOCOM*, pp. 2811-2815, 2012.

[18] C. Luo, F. Wu, J. Sun, and C.W. Chen, "Compressive data gathering for large-scale wireless sensor networks", *ACM MobiCom*, pp. 145-156, 2009.

[19] Y. Liu, Y. He, M. Li, J. Wang, K. Liu, L. Mo, W. Dong, Z. Yang, M. Xi, J. Zhao, and X.-Y. Li, "Does wireless sensor network scale? a measurement study on greenorbs", *IEEE INFOCOM*, pp. 873-881, 2011.

[20] IntelLab data. http://www.select.cs.cmu.edu/data/labapp3/index.html.

[21] NBDC-CTD data. http://tao.noaa.gov/refreshed/ctd_delivery.php.

[22] P. Li, T.J. Hastie, and K.W. Church, "Very sparse random projections", *ACM SIGKDD*, pp. 287-296, 2006.

[23] W. Wang, M. Garofalakis, and KW Ramchandran, "Distributed sparse random projections for refinable approximation", *IEEE IPSN*, pp. 331-339, 2007.

[24] M.A. Davenport, M.F. Duarte, Y.C. Eldar, and G. Kutyniok, "Introduction to compressed sensing", *Preprint*, Vol. 93, 2011.

[25] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation", *J. Amer. Math. Soc*, 22(1), pp. 211-231, 2009.

[26] M.H. Lu, P. Steenkiste, and T. Chen, "Design, implementation and evaluation of an efficient opportunistic retransmission protocol", *ACM MobiCom*, pp. 73-84, 2009.

[27] S. Biswas and R. Morris, "Opportunistic routing in multi-hop wireless networks", *ACM SIGCOMM Computer Communication Review*, Vol. 34, No. 1, pp. 69-74, 2004.

[28] S. Chachulski, M. Jennings, S. Katti, and D. Katabi, "Trading structure for randomness in wireless opportunistic routing", *ACM SIGCOMM*, Vol. 37, No. 4, pp. 169-180, 2007.

[29] L. Guo, R. Beyah, and Y. Li, "Smite: a stochastic compressive data collection protocol for mobile wireless sensor networks", *IEEE INFOCOM*, pp. 1611-1619, 2011.

[30] W. Xu, E. Mallada, and A. Tang, "Compressive sensing over graphs", *IEEE INFOCOM*, pp. 2087-2095, 2011.

[31] E.J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information", *IEEE Transactions on Information Theory*, 52(2), pp. 489-509, 2006.

[32] E.J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, Vol. 52, No. 12, pp. 5406-5425, 2006.

[33] G. Quer, R. Masiero, D. Munaretto, M. Rossi, J. Widmer, and M. Zorzi, "On the interplay between routing and signal representation for compressive sensing in wireless sensor networks", *IEEE Information Theory and Applications Workshop*, pp. 206-215, 2009.

[34] S. Johansen and K. Juselius, "Maximum likelihood estimation and inference on cointegration‡with applications to the demand for money", *Oxford Bulletin of Economics and statistics*, 52(2), pp. 169-210, 1990.

[35] C. Sherlaw-Johnson, S. Gallivan, and J. Burridge, "Estimating a markov transition matrix from observational data", *Journal of the Operational Research Society*, pp. 405-410, 1995.

[36] B. Krishnamachar. *Matlab code*: http://http://ceng.usc.edu/~bkrishna/.

[37] T. Cover and P. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[38] X. Jiang. M.V. Ly, J. Taneja, P. Dutta, and D. Culler, "Experiences with a high-fidelity wireless building energy auditing network", *ACM Sensys*, pp. 113-126, 2009.

[39] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey", *Ad Hoc Networks*, Vol. 7. No. 3, pp. 537-568, 2009.

[40] R. Cristescu, B. Beferull-Lozano, and M. Vetterli, "Network correlated data gathering with explicit communication: NP-completeness and algorithms", *IEEE/ACM Transactions on Networking (ToN)*, Vol. 14, No. 1, pp. 41-54, 2006.

[41] X. Liu, J. Luo, A.V. Vasilakos, "Compressed data aggregation for energy efficient wireless sensor networks", *IEEE SECON*, pp. 46-54, 2011.

# APPENDIX A
## PROOFS

### A. Lemma 1

*Proof:* The pair of nonuniform sparse random projections $W, M \in \mathbb{R}^{L \times N}$ satisfies:

$$E[W_{ij}] = 0, E[\mathcal{M}_{ij}] = 0; E[W_{ij}M_{ij}] = 1,$$
$$E[W_{il}M_{im}] = E[W_{il}]E[M_{im}] = 0 \ \ if \ l \neq m,$$
$$E[W_{il}M_{il}W_{im}M_{im}] = E[W_{il}M_{il}]E[W_{im}M_{im}] = 1 \ \ if \ l \neq m,$$
$$E[W_{ij}^2] = \pi_j, E[M_{ij}^2] = \frac{1}{\pi_j}, E[W_{ij}^2 M_{ij}^2] = \frac{1}{\pi_j},$$
$$E[W_{il}^2 M_{im}^2] = E[W_{il}^2]E[M_{im}^2] = \frac{\pi_l}{\pi_m} \ \ if \ l \neq m.$$

The above results will be used in the following process without explicit mention.

Define the random variables

$$\omega_i = \left( \sum_{j=1}^{N} U_j W_{ij} \right) \left( \sum_{j=1}^{N} V_j M_{ij} \right)$$
$$\zeta = Y^{\mathcal{T}} V' = \frac{1}{L} \sum_{i=1}^{L} \omega_i \tag{28}$$

so that $\omega_1, \omega_2, ..., \omega_M$ are independent.

$$E(\omega_i) = E\left[ \sum_{j=1}^{N} U_j V_j W_{ij} M_{ij} + \sum_{l \neq m} U_l V_m W_{il} M_{im} \right]$$
$$= \sum_{j=1}^{n} U_j V_j E[W_{ij}M_{ij}] + \sum_{l \neq m} U_l V_m E[W_{il}M_{im}] \tag{29}$$
$$= U^{\mathcal{T}} V$$
$$E[\zeta] = U^{\mathcal{T}} V$$

Similarly, we can compute the second moments and variance as following:

$$E[\omega_i^2] = E\left[ \left( \sum_{j=1}^{n} U_j V_j W_{ij} M_{ij} \right)^2 + \left( \sum_{l \neq m} U_l V_m W_{il} M_{im} \right)^2 \right.$$
$$\left. + 2\left( \sum_{j=1}^{N} U_j V_j W_{ij} M_{ij} \right)\left( \sum_{l \neq m} U_l V_m W_{il} M_{im} \right) \right]$$
$$= \sum_{j=1}^{N} U_j^2 V_j^2 E[W_{ij}^2 M_{ij}^2] + 2 \sum_{l<m} U_l V_l U_m V_m E[W_{il}M_{il}W_{im}M_{im}]$$
$$+ \sum_{l \neq m} U_l^2 V_m^2 E[W_{il}^2 M_{im}^2] + 2 \sum_{l<m} U_l V_m U_m V_l E[W_{il}M_{il}W_{im}M_{im}]$$
$$= \sum_{j=1}^{N} U_j^2 V_j^2 \frac{1}{\pi_j} + 2 \sum_{l \neq m} U_l V_l U_m V_m + \sum_{l \neq m} U_l^2 V_m^2 \frac{\pi_l}{\pi_m}$$
$$(Let \ \xi = max(\frac{\pi_l}{\pi_m}), \kappa = \frac{1}{min(\pi)})$$
$$\leq 2 \left( \sum_{j=1}^{N} U_j^2 V_j^2 + 2 \sum_{l \neq m} U_l V_l U_m V_m \right)$$
$$+ \xi \left( \sum_{j=1}^{N} U_j V_j + \sum_{l \neq m} U_l^2 V_m^2 \right) + (\kappa - 2 - \xi) \sum_{j=1}^{N} U_j V_j$$
$$= 2(U'V)^2 + \xi \|U\|_2^2 \|V\|_2^2 + (\kappa - 2 - \xi) \sum_{j=1}^{N} U_j^2 V_j^2$$
$$Var(\omega_i) = E[\omega_i^2] - E[\omega_i]^2$$
$$\leq (U^{\mathcal{T}}V)^2 + \xi \|V\|_2^2 \|V\|_2^2 + (\kappa - 2 - \xi) \sum_{j=1}^{N} U_j^2 V_j^2 \tag{30}$$

$$Var(\zeta) = \frac{1}{L^2} \sum_{i=1}^{L} Var(\omega_i)$$
$$\leq \frac{1}{L} \left( (U^{\mathcal{T}}V)^2 + \xi \|U\|_2^2 \|V\|_2^2 + (\kappa - 2 - \xi) \sum_{j=1}^{N} U_j^2 V_j^2 \right) \tag{31}$$

$\blacksquare$

### B. Theorem 1

*Proof:* Fix any two vectors $U, V \in \mathbb{R}^{N \times 1}$, with $\|U\|_\infty / \|U\|_2 \leq H$. Set $L = L_1 L_2$, with $L_1$, $L_2$ are positive integers. Partition the $L \times N$ matrix $W$ and $M$ into $L_2$ matrices $\{W_1, W_2, ..., W_{L_2}\}$ and $\{M_1, M_2, ..., M_{L_2}\}$, each of size $L_1 \times N$. The corresponding random projections are $\{Y_1 = \frac{1}{\sqrt{L_1}} W_1 U, ..., Y_{L_2} = \frac{1}{\sqrt{L_1}} W_{L_2} U\}$, and $\{V_1' = \frac{1}{\sqrt{L_1}} W_1 V, ..., V_{L_2}' = \frac{1}{\sqrt{L_1}} W_{L_2} V\}$.

Define the independent random variables $\zeta_l = Y_l^{\mathcal{T}} V_l'$, $l = 1, 2, ..., L2$. Applying Lemma 1 to each $\zeta_l$, we derive that $E[\zeta_l] = U^{\mathcal{T}} V$ and

$$Var(\zeta) \leq \frac{1}{L_1} \left( (U'V)^2 + \xi \|U\|_2^2 \|V\|_2^2 + (\kappa - 2 - \xi) \sum_{j=1}^{N} U_j^2 V_j^2 \right) \tag{32}$$

By the Chebyshev inequality and

$$P(|\zeta_l - U^{\mathcal{T}}V| \geq \epsilon\|U\|_2\|V\|_2) \leq \frac{Var(\zeta_l)}{\epsilon^2\|U\|_2^2\|V\|_2^2}$$

$$\leq \frac{1}{\epsilon^2 L_1}\left(\frac{(U^{\mathcal{T}}V)^2}{\|U\|_2^2\|V\|_2^2} + \frac{\xi\|U\|_2^2\|V\|_2^2}{\|U\|_2^2\|V\|_2^2} + (\kappa - 2 - \xi)\frac{\sum_{j=1}^N U_j^2 V_j^2}{\|U\|_2^2\|V\|_2^2}\right)$$

(By the Cauchy − Schwarz inequality and Equ.(22))

$$\leq \frac{1}{\epsilon^2 L_1}\left(1 + \xi + \kappa\frac{H^2\|U\|_2^2\sum_{j=1}^N V_j^2}{\|U\|_2^2\|V\|_2^2}\right)$$

$$\leq \frac{1}{\epsilon^2 L_1}\left(1 + \xi + \kappa H^2\right) \triangleq p$$

(33)

Thus, we can obtain a constant probability $p$ by setting $L_1 = O(\frac{1+\xi+\kappa H^2}{\epsilon^2})$.

We define the estimate $\hat{a}$ as the median of the independent random variables $\zeta_1, ..., \zeta_{L_2}$, each of which lies outside of the tolerable approximation interval with probability $p$. Formally, let $I_l$ be the indicator random variable of the event that $\{|\zeta_l - U^{\mathcal{T}}V| \geq \epsilon\|U\|_2\|V\|_2\}$, which occurs with probability $p$.

Let $I = \sum_{l=1} L_2 I_l$ be the number of $\zeta_l$'s that lie outside of the tolerable interval, where $E[I] = L_2 p$. When the event that at least half of the $\zeta_l$'s are outside the tolerable interval occurs with arbitrarily small probability, then the median $\hat{a}$ is within the tolerable interval. So, if we set $p < 1/2$, say $p = 1/4$, and apply the Chernoff bound, we get:

$$P\left(I > (1+c)\frac{L_2}{4}\right) < e^{-c^2 L_2/12}$$

(34)

where $0 < c < 1$ is some constant.

Thus, for $U$ and $V_i \in V_1, ..., V_N \subset \mathbb{R}^{N\times 1}$, and the corresponding random projections $\alpha$ and $\beta$ produce an estimate $\hat{a}$ for $U^{\mathcal{T}}V$ that lies outside the tolerable approximation interval with probability at most $e^{-c^2 L_2/12}$. For the $N$ estimates $\hat{a}_i, i = 1, 2, ..., N$, the probability that at least one lies outside the tolerable interval is upper bounded by $P_e \leq Ne^{-c^2 L_2/12}$.

Setting $L_1 = O(\frac{1+\xi+\kappa H^2}{\epsilon^2})$ obtains $p = 1/4$, and setting $L_2 = O((1+\gamma)logN)$ obtains $P_e \leq N^{-\gamma}$ for some constant $\gamma > 0$. Therefore, with $L = L_1 L_2 = O(\frac{1+\gamma}{\epsilon^2}(1 + \xi + \kappa H^2)logN)$, the nonuniform random projection pair $W, M$ can produce inner products of vectors with probability at least $1 - N^{-\gamma}$. If $(\xi + \kappa H^2) > \Omega(1)$, then $L = O(\frac{1+\gamma}{\epsilon^2}(1 + \xi + \kappa H^2)logN)$. If $(\xi + \kappa H^2) \leq O(1)$, then $L = O(\frac{1+\gamma}{\epsilon^2}logN)$. ∎

### C. Theorem 2

For an orthonormal transform $\mathbf{\Psi} = \{\Psi_1, ..., \Psi_N\} \subset \mathbb{R}^{N\times 1}$. Denote the transform coefficients by $\theta = [u^{\mathcal{T}}\Psi_1, ..., U^{\mathcal{T}}\Psi_N]^{\mathcal{T}}$, we order them in decreasing magnitude as $|\theta|_{(1)} \geq |\theta|_{(2)} \geq ... \geq |\theta|_{(N)}$, then the approximation error by taking the largest $K$ coefficients and setting the remaining coefficients as zero, is $\|\theta - \hat{\theta}_{opt}\|_2^2 = \sum_{i=K+1}^N |\theta|_{(i)} \leq \eta\|\theta\|_2^2$.

Then by Theorem 1, the random projections $\frac{1}{\sqrt{L}}WU$ and $\{\frac{1}{\sqrt{L}}M\Psi_1, \frac{1}{\sqrt{L}}M\Psi_2, ...., \frac{1}{\sqrt{L}}M\Psi_N\}$ produces, with high

probability, estimates $\{\hat{\theta}_1, ..., \hat{\theta}_N\}$, each satisfying

$$|\hat{\theta}_i - \theta_i| \leq \delta\|\theta\|_2$$

(35)

where we plugged in $\|\Psi_i\|_2 = 1$ and $\|U\|_2 = \|\theta\|_2$ by orthonormality of $\Psi$. We get

$$|\theta_i| - \delta\|\theta\|_2 \leq |\hat{\theta}_i| \leq |\theta_i| + \delta\|\theta\|_2, \forall i.$$

(36)

Order the estimates $\breve{\theta}$ in decreasing magnitude $|\breve{\theta}|_{(1)} \geq |\breve{\theta}|_{(2)} \geq ... \geq |\breve{\theta}|_{(N)}$. We define our approximation $\hat{\theta}$ as keeping the $K$ largest components of $\breve{\theta}$ in magnitude, and setting the remaining components to zero. Let $\breve{\mathcal{L}}$ be the index set of the $K$ largest estimates $\breve{\theta}_i$'s which we keep (thus $\breve{\mathcal{L}}^C$ is the index set of the estimates we set to zero). Let $\mathcal{L}$ be the index set of the $K$ largest transform coefficients $\theta_i$'s

$$\|\theta - \breve{\theta}\|_2^2 = \sum_{i\in\breve{\mathcal{L}}}|\theta_i - \breve{\theta}_i|^2 + \sum_{i\in\breve{\mathcal{L}}^C}|\theta_i|^2$$

$$\leq K\delta^2\|\theta\|_2 + \sum_{i\in\breve{\mathcal{L}}^C}|\theta_i|^2$$

(37)

In the ideal case, $\breve{\mathcal{L}} = \mathcal{L}$, then $\sum_{i\in\breve{\mathcal{L}}^C}|\theta_i|^2 = \sum_{i\in\mathcal{L}^C}|\theta_i|^2$. If $\breve{\mathcal{L}} \neq \mathcal{L}$, then we have chosen to keep the estimates which did not belong to the $K$ largest, and consequently we have set to zero some coefficients which did belong to the $K$ largest.

Assume that there exists some $i \in \breve{\mathcal{L}}, i \notin \mathcal{L}, j \notin \breve{\mathcal{L}}, j \in \mathcal{L}$, that $|\hat{\theta}_i| > |\hat{\theta}_j|$ but $|\theta_i| < |\theta_j|$. Since the estimates are within a small interval $\pm\delta\|\theta\|_2$ around the transform coefficients. Thus, this kind of confusion can only happen if $|\theta_j| - |\theta_i| \leq 2\delta\|\theta\|$. Furthermore, $|\theta_i|^2 + |\theta_j|^2 \leq \|\theta\|_2^2$ implies that $|\theta_j| + |\theta_i| \leq \sqrt{3}\|\theta\|_2$. Thus, $|\theta_j|^2 - |\theta_i|^2 = (|\theta_j| - |\theta_i|)(|\theta_j| + |\theta_i| \leq 2\sqrt{3}\|\theta\|_2^2$.

For each time the above confusion happens, we get an additional error $+|\theta_j|^2 - |\theta_i|^2$, and this confusion can happen at most $K$ times. Therefore, we have:

$$\sum_{i\in\breve{\mathcal{L}}^C}|\theta_i|^2 \leq \sum_{i\in\mathcal{L}}^C |\theta_i|^2 + K(2\sqrt{3}\delta\|\theta\|_2^2)$$

(38)

$$\|\theta - \breve{\theta}\|_2^2 \leq K\delta^2\|\theta\|_2^2 + 2\sqrt{3}K\delta\|\theta\|_2^2 + \sum_{i\in\mathcal{L}^C}|\theta_i|^2$$

$$= K\delta^2\|\theta\|_2 + 2\sqrt{3}K\delta\|\theta\|_2^2 + \|\theta - \hat{\theta}_{opt}\|_2^2$$

$$\leq K\delta^2\|\theta\|_2 + 2\sqrt{3}K\delta\|\theta\|_2^2 + \eta\|\theta\|_2^2$$

(39)

Setting $K\delta^2\|\theta\|_2 + 2\sqrt{3}K\delta\|\theta\|_2^2 = \varrho\|\theta\|_2^2$ and solving for the positive root, we find that $\delta = -\sqrt{3} + \sqrt{3 + \varrho/K} = O(\varrho/K)$.

$$\|\theta - \breve{\theta}\|_2^2 \leq \varrho\|\theta\|_2^2 + \eta\|\theta\|_2^2$$

$$= \left(1 + \frac{\varrho}{\eta}\right)\eta\|\theta\|_2^2$$

(40)

Let $\epsilon = \frac{\varrho}{\eta}$, so that $\delta = O(\frac{\epsilon\eta}{K})$. Therefore, the number of random projections we need is $L = O\left(\frac{1+\gamma}{\delta^2}\kappa H^2 logN\right) = O\left(\frac{1+\gamma}{\epsilon^2\eta^2}\kappa H^2 K^2 logN\right)$ if $\kappa H^2 > \Omega(1)$, and $L = O\left(\frac{1+\gamma}{\epsilon^2\eta^2}K^2 logN\right)$ if $\kappa H^2 \leq O(1)$.