Data Loss and Reconstruction in Sensor Networks

Linghe Kong*, Mingyuan Xia†, Xiao-Yang Liu*, Min-You Wu*, Xue Liu†
*Shanghai Jiao Tong University, †McGill University
*{linghe.kong, yanglet, mwu}@sjtu.edu.cn †{mingyuan.xia@mail, xueliu@cs}.mcgill.ca

***Abstract*—Reconstructing the environment in cyber space by sensory data is a fundamental operation for understanding the physical world in depth. A lot of basic scientific work (*e.g.*, nature discovery, organic evolution) heavily relies on the accuracy of environment reconstruction. However, data loss in wireless sensor networks is common and has its special patterns due to noise, collision, unreliable link, and unexpected damage, which greatly reduces the accuracy of reconstruction. Existing interpolation methods do not consider these patterns and thus fail to provide a satisfactory accuracy when missing data become large. To address this problem, this paper proposes a novel approach based on compressive sensing to reconstruct the massive missing data. Firstly, we analyze the real sensory data from Intel Indoor, GreenOrbs, and Ocean Sense projects. They all exhibit the features of spatial correlation, temporal stability and low-rank structure. Motivated by these observations, we then develop an *environmental space time improved compressive sensing* (ESTI-CS) algorithm to optimize the missing data estimation. Finally, the extensive experiments with real-world sensory data shows that the proposed approach significantly outperforms existing solutions in terms of reconstruction accuracy. Typically, ESTI-CS can successfully reconstruct the environment with less than 20% error in face of 90% missing data.**

I. Introduction

For the sake of discovering the physical world, people keep observing the environment. Recently, wireless sensor networks (WSNs) [1] are widely adopted to gather various environmental information and then reconstruct them in the cyber worlds [9]. There are plenty of real environment monitoring applications under the water [24], in the forest [17], and on the volcano [22]. *Environment Matrix (EM)* is a common way to represent a dynamic environment. EM is typically an $n \times t$ matrix that records data from n sensors over t time intervals. Environment reconstruction [12] attempts to obtain the full and accurate EM from raw sensory data, which is the essential step that precedes any further analysis.

Motivation: A great deal of basic scientific work heavily depends on the accuracy of environment reconstruction. For example, scientists reveal the nature of ocean currents from accurate underwater temperature data [24], understand the demand for plant evolution based on the light condition in the forest [17], discover the eruption omen by monitoring the shake of the volcano [22].

However, since data gathering is largely affected by hardware and wireless conditions, raw EMs usually have notable missing data. Furthermore, missing data becomes larger as deployed WSNs grow in scale [3]. Consequently, data loss in WSN becomes the key challenge against accurate environment reconstruction. Therefore, it is urgent and important to design effective methods to recover incomplete EMs.

Existing approaches and limitations: Missing value problem is fundamental in datasets. Lots of work has contributed in this field such as local interpolation method K-Nearest Neighbors (KNN) [6], global refinement method Delaunay Triangulation (DT) [12], and principal component analysis method Multi-channel Singular Spectrum Analysis (MSSA) [26]. These methods are often used when there are only a few missing values, but cannot scale when the missing data grow.

Compressive Sensing (CS) [5], [7] is a powerful and generic technique for estimating missing data. CS can recover an entire dataset from only a small fraction of data as long as these data contain certain structures or features. So far, CS has been applied to reconstruct network traffic [25], refine localization [18] and improve urban traffic sensing [14]. However, since WSN has unique data loss patterns, CS cannot be directly applied to gain notable accuracy improvement for EM interpolation.

Our contributions: In this paper, our work is threefold.

Firstly, we analyze real-world environmental data from Intel Indoor [10], GreenOrbs [17], and Ocean Sense [24] projects. We confirm the massive data loss in general applications and mine the specific data loss patterns in WSN. And then we reveal three features in real environmental datasets: 1) *Time stability*. The sensory values of one certain node are usually similar at adjacent time slots. 2) *Space correlation*. The sensory values of neighbor nodes are similar for a particular time instant. 3) *Low-rank structure*. The major energy concentrates on just a few principle data in EM, which underpins the applicability of CS.

Then, motivated by these three observations, we design a novel *environmental space time improved compressive sensing* (ESTI-CS) algorithm for estimating the missing data. ESTI-CS embeds customized features into baseline CS to deal with the specific data loss patterns, which computes the minimal low-rank approximations of the incomplete EM and refines the interpolation with spatio-temporal features.

Finally, we evaluate the effectiveness of our approach based on trace-driven simulation. We demonstrate that ESTI-CS can outperform existing approaches such as CS, KNN, DT, and MSSA when the raw data contain diverse real loss patterns. Typically, ESTI-CS can achieve an effective environment reconstruction with less than 20% error when there are 90% missing data in the collected data.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first work to study the data loss and reconstruction in WSN. We model the environment with EM and discover four data loss

978-1-4673-5946-7/13/$31.00 ©2013 IEEE

patterns.
- We mine several large WSN datasets and reveal the time stability, space correlation, and low-rank features in real environment.
- Based on the observed features, we design the ESTI-CS algorithm to accurately estimate the missing data in highly incomplete EM.
- The proposed ESTI-CS is simulated based on real data. The evaluation shows the ESTI-CS is effective for massively data loss against existing interpolation methods.

Paper organization: Section II presents the related work. Section III models the problem. Section IV analyzes the data loss. Section V mines the environmental features. Section VI proposes the ESTI-CS algorithm. Section VII evaluates the proposed approach. Section VIII concludes the paper.

II. Related Work

Missing value problem is common in datasets [3]. A great deal of existing work has been devoted to interpolate missing data. K-Nearest-Neighbor (KNN) [6] is a classical local interpolation method. KNN simply utilizes the values of the nearest K neighbors to estimate the missing one. It is frequently used in many low-fidelity estimation cases. Delaunay Triangulation (DT) [12] is a typical global refinement method, which treats the gathered data as vertices. DT takes advantage of these vertices and their global errors to rebuild virtual triangles for data interpolation. It is widely adopted in computer vision for surface rendering. Multi-channel Singular Spectrum Analysis (MSSA) [26] is a data adaptive and nonparametric method based on the embedded lag-covariance matrix. MSSA is often used in geographic data and meteorological data recovery.

Despite much progress in the area of data interpolation, existing methods are suitable for interpolation with only few missing values, but perform poorly when data loss rate is high, *e.g.*, for environment reconstruction in WSN.

Compressive sensing (CS) is a generic method to recover whole condition with just a few sampled data [4], [7]. Its fundamental theory has been utilized in plenty of fields such as statistics, image processing, signal recovery, and machine learning. As for missing value estimation, CS-based interpolation methods have been developed for network traffic estimation [25], road traffic interpolation [14], and localization in mobile networks [18]. CS also witnesses wide application in WSN, *e.g.*, recovering signal under noise [2], balancing load via compressive data gathering [16]. However, the study of CS for environment reconstruction in WSN is still vacant.

Existing CS-based interpolation methods cannot be directly applied for accurate environment reconstruction for two reasons 1) CS-based methods require the dataset to have inherent structure and redundancy. Existing features extracted from network trace [25] or road traffic [14] are not valid for WSN sensory data. 2) CS theory performs well when the missing values follow the Gaussian or pure random distribution [15], [21]. However, as shown in Section IV-C, the loss patterns of WSNs do not satisfy these prerequisites.

To address the above challenges, effective environment reconstruction of WSN data require considering massive data loss as well as studying WSN-specific loss patterns.

III. Problem Formulation

A. Environmental Data Reconstruction

Rebuilding the virtual environment (such as the dynamic temperature, light, humidity, gas concentration, or magnetic strength in real world) in cyber space based on the sensory data is called environment reconstruction.

In environment reconstruction systems, sensor nodes are scattered in the given area. They sense and report data to the sink periodically over a given time span. Suppose totally n sensor nodes are deployed. The monitoring period includes t time slots. Each sensor node is required to report its sensory data once per time slot through wireless transmission. $x(i,j)$ denotes the sensory data of node i at time slot j, where $i = 1, 2 \cdots n$ and $j = 1, 2 \cdots t$.

Definition 1 Environment Matrix (EM): is a mathematical method to describe the dynamic environment. EM is defined by $X = (x(i,j))_{n \times t}$.

Thereby, EM is a matrix constituting of n rows and t columns. A complete EM represents that every data points in the matrix are validly collected data, *i.e.*, no missing data points.

Definition 2 Binary Index Matrix (BIM): is an $n \times t$ matrix, which indicates if a data point in an EM is missing. BIM is defined by:

$$B = (b(i,j))_{n \times t} = \begin{cases} 0 & \text{if } x(i,j) \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases} \quad (1)$$

Definition 3 Sensory Matrix (SM): is an $n \times t$ matrix, which records the raw data collected from WSN. Due to the presence of missing data, elements of SM are either $x(i,j)$ gathered by WSN or zero (data loss).

Thereby, SM is an incomplete EM. SM is denoted by S and can be presented by the element-wise production of X and B,

$$S = X \cdot B, \quad (2)$$

B. Problem Statement

Data reconstruction is to rebuild the real environment (EM) based on the gathered sensory data (SM).

Definition 4 Reconstructed Matrix (RM): is generated by interpolating the missing values in an SM to approximate EM. RM is denoted by $\hat{X} = (\hat{x}(i,j))_{n \times t}$.

Problem: Environment Reconstruction in Sensor Network (ERSN): Given an SM S, the ERSN problem is to find an optimal RM $\hat{X}$ that approximates the original EM X as closely as possible. *i.e.*,

$$\begin{array}{ll} \text{Objective:} & min||X - \hat{X}||_F, \\ \text{Subject to:} & S, \end{array} \quad (3)$$

where $||\cdot||_F$ is the Frobenius norm used to measure the error between matrix X and $\hat{X}$. For calculating, take EM X as an example, $||X||_F = \sqrt{\sum_{i,j}(x(i,j))^2}$.

TABLE I
THE RATIO OF DATA LOSS IN REAL DATASETS

Data Name	Nodes	Time Interval	Data Loss Ratio
Intel Indoor	54	30 seconds	23%
GreenOrbs	450	10 minutes	35%
Ocean Sense	20	2 minutes	64%

In the ERSN problem, the objective is to minimize the absolute error. In order to measure the error of reconstruction in different scenarios among different methods, we further define the following metric.

Definition 5 Error Ratio (ER): is a metric for measuring the reconstruction error after interpolation:

$$\epsilon = \frac{\sqrt{\sum_{i,j:b(i,j)=0}(x(i,j) - \hat{x}(i,j))^2}}{\sqrt{\sum_{i,j:b(i,j)=0}(x(i,j))^2}}. \quad (4)$$

Note that the condition $b(i,j) = 0$ in Eqn. (4) indicates that only errors on the missing data are counted.

IV. DATA LOSS IN SENSOR NETWORKS

A. Environmental Datasets

In this section, we analyze the data loss in real WSN datasets. The analysis is based on three datasets gathered by Intel indoor experiment, GreenOrbs, and OceanSense projects.

The data of Intel indoor experiment [10] are gathered by Intel Berkeley Research lab from February 28th to April 5th, 2004. There are 54 Mica2Dot nodes placed in a 40m×30m room. Every node reports once every 30 seconds. Sensory data include temperature, light, and humidity.

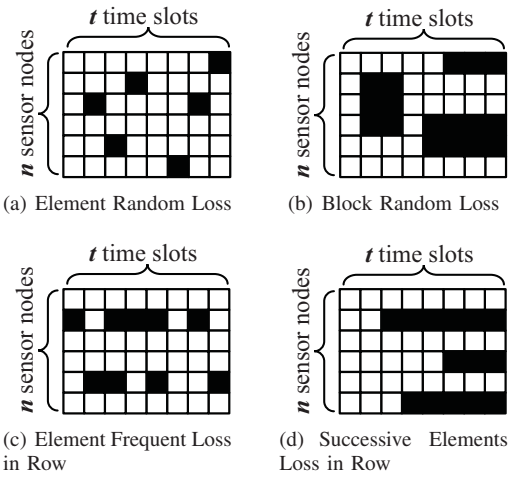
GreenOrbs project [17] is a real WSN application for forest surveillance from 2008 to present. More than 450 TelosB nodes are scattered on the Tianmu Mountain, China and gather temperature, light, and humidity once every 10 minutes.

Ocean Sense project [24] carried out by Ocean University of China. This dataset contains 20 TelosB nodes deployed in the sea of Taipingjiao, China from 2007 to present, monitoring an area of 300m×100m. Each sensing node reports temperature and light data every 2 minutes.

B. Massive Data Loss

Through statistics analysis, we verify that the significant data loss exists in all of these original datasets.

We investigate totally 54 nodes and 84600 time slots (one month) data from Intel Indoor dataset. 23% data points are missing. The GreenOrbs dataset also observes 35% data loss. And this loss is even larger in OceanSense, which is about 64% for 20 nodes and 5040 time slots (one week). The basic information of three datasets and their data loss ratios are listed in Table I. We find that the data loss is common and significant in real WSNs.



(a) Element Random Loss (b) Block Random Loss

(c) Element Frequent Loss in Row (d) Successive Elements Loss in Row

Fig. 1. Data loss patterns in WSN. The tessellations illustrate the sensory matrix. The black cells represent the elements of missing data.

C. Data Loss Pattern

Traditional work usually assumes that the data loss follows a random distribution [14], [26]. However, this claim does not apply to the WSN situation. According to the nature of WSN, we synthesize several typical data loss patterns.

Pattern 1 Element Random Loss (ERL): This is the simplest loss pattern. Data elements in the matrix are dropped independently and randomly. As shown in Fig. 1(a), the missing data for ERL are randomly distributed in the SM. The noise and collision [11] in WSN are the root cause of random element loss.

Pattern 2 Block Random Loss (BRL): Data from adjacent nodes in adjacent time slots are dropped independently and randomly. In WSN, congestion [8] always causes data loss on high-density sensor nodes during a period of time. Fig. 1(b) visualizes this scenario.

Pattern 3 Element Frequent Loss in Row (EFLR): Unreliable links [23] are common phenomenon in real wireless scenarios. When the quality of link state is not good, sensory data are prone to loss due to the intermittent transmission. As shown in Fig. 1(c), in EFLR, elements in some particular rows have a higher missing probability.

Pattern 4 Successive Elements Loss in Row (SELR): This pattern models that a given node starts losing from a particular time slot. This type of loss occurs when some sensor nodes are damaged or run out of energy [20], which is made visible by Fig. 1(d).

Pattern 5 Combinational Loss (CL): In real world, data loss always happens as a combination of some loss patterns above.

V. ENVIRONMENTAL DATA MINING

In order to discover the environmental features, the complete datasets are always desired. However, we cannot directly derive EMs from the three original datasets since they all observe considerate data loss. To generate EMs, we perform

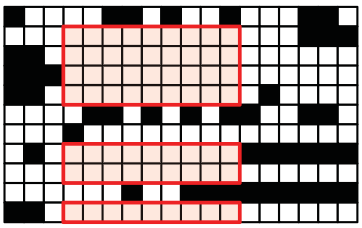


Fig. 2. Selecting the red parts from the original dataset to consist a small but completed dataset for environmental features analysis.

TABLE II
SELECTED DATASETS FOR ENVIRONMENTAL FEATURES ANALYSIS

Data Name	Matrix Size	Time Interval
Intel Indoor	49 nodes × 149 intervals	90 seconds
GreenOrbs	281 nodes × 170 intervals	10 minutes
Ocean Sense	10 nodes × 42 intervals	30 minutes

preprocessing (as demonstrated in Fig. 2) that selects complete subset from these three datasets. As a result, six EMs are generated from preprocessing: indoor temperature, indoor light, forest temperature, forest light, ocean temperature, and ocean light.

A. Low-Rank Structure Discovery

Environmental data of different locations over different times are not independent. There exists inherent structure or redundancy. We mine these features in above real dataset by using Principal Component Analysis (PCA), which is an effective non-parametric technique for revealing sometimes hidden, simplified structure that often underlies a dataset [13].

Any $n \times t$ matrix X can be decomposed into three matrices according to Singular Value Decomposition (SVD):

$$X = U\Sigma V^T = \sum_{i=1}^{min(n,t)} \sigma_i u_i v_i^T, \tag{5}$$

where V^T is the transpose of V, U is an $n \times n$ unitary matrix (*i.e.*, $UU^T = U^TU = I_{n\times n}$), V is a $t \times t$ unitary matrix (*i.e.*, $VV^T = V^TV = I_{t\times t}$), and Σ is an $n \times t$ diagonal matrix constraining the singular values σ_i of X. Typically, the singular values in Σ are sorted, *i.e.*, $\sigma_i \geq \sigma_{i+1}, i = 1, 2, \cdots min(n,t)$, where $min(n,t)$ is the number of singular values. The rank of a matrix, denoted by r, is equal to the number of its non-zero singular values. If $r << min(n,t)$, the matrix is low-rank.

In Eqn. (5) the singular value σ_i also indicates the energy of the i-th principal component. The total energy is equal to the sum of all singular value $\sum_{i=1}^{min(n,t)} \sigma_i$. According to PCA, a low-rank matrix [25] exhibits that its first r singular values occupy the total or near-total energy $\sum_{i=1}^{r} \sigma_i \approx \sum_{i=1}^{min(n,t)} \sigma_i$.

In Fig. 3(a), we illustrate the distribution of singular values in 6 EMs. The X-axis presents the i-th singular values. Since the scales of 6 EMs are different, we normalize the X-axis. So $min(n,t)$ of every EM is normalized to 100%. The Y-axis presents the values of i-th singular value. Due to the same reason of X-axis, the Y-axis is also normalized. *i.e.*, $\max(\sigma_i)$ of every EM is normalized to 1. This figure suggests that the energy is always contributed by the top several singular values in real environments. For example, the top 5% singular values contribute all energy in Indoor-Temp; the top 12% σ_i include all energy in Forest-Temp; and even in the worst case of Ocean-Light, the top 25% singular values contribute the most of energy. The universal existence of $\sum_{i=1}^{r} \sigma_i \approx \sum_{i=1}^{min(n,t)} \sigma_i$ and $r << min(n,t)$ reveals that EMs exhibit obvious low-rank structures. Low-rank features [14] serve for the prerequisite for using compressive sensing.

B. Temporal Stability Feature

In real world, most of measured data (e.g., temperature) usually change stably, *i.e.*, there is little mutation on environmental value between adjacent time slots. On the basis of this natural phenomenon, we analyze the datasets in time dimension to reveal temporal features.

We measure the temporal stability at node i and time slot j by computing the normalized difference values between adjacent time slots $\Delta x(i,j)$:

$$\Delta x(i,j) = \frac{|x(i,j) - x(i,j-1)|}{max(|x(I,J) - x(I,J-1)|)}, \tag{6}$$

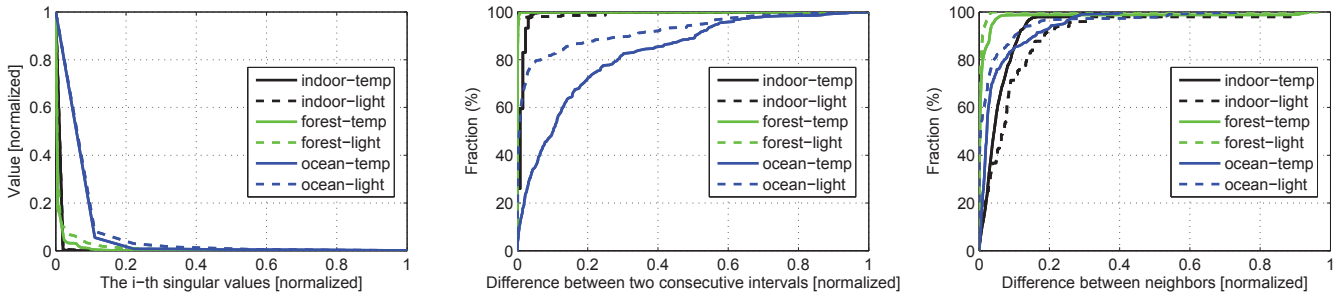
where I varies from 1 to n, J varies from 1 to t, and $max(|x(I,J) - x(I,J-1)|)$ is the maximal difference between any two consecutive time slots in the EM.

The CDF of $\Delta x(i,j)$ is plotted in Fig. 3(b). The X-axis presents the normalized difference values between two consecutive time slots, *i.e.*, $\Delta x(i,j)$. The Y-axis presents the cumulative probability. We observe that $> 80\%$ in the Forest datasets, $> 60\%$ in the Indoor datasets, and $> 50\%$ in the Ocean-Temp, the value of $\Delta x(i,j)$ is 0. *i.e.*, the environmental value is not changed between two consecutive time slots. In addition, near all ($> 95\%$) $\Delta x(i,j)$ are very small (< 0.05) in Forest and Indoor datasets. Even in the worst case, the ocean-light values between two consecutive time slots mostly ($> 80\%$) change only a little (< 0.3). These results indicate that temporal stability exists in real environments. Based on this discovery, we can adopt the time feature to optimize the compressive sensing technique for missing data estimation.

C. Spatial Correlation Feature

We also consider the difference value from the space dimension. We know that environments are often smooth in a small area, *i.e.*, at the same time, environmental values are similar at nearby locations.

In real WSN applications, the locations of nodes can either be known [17] or unknown [24]. Generally, it is not easy to know the actual distance between nodes from a WSN without GPS information. Although physical distance may not be available, the network topology is always easy to obtain. The topology can be known from the routing information when the sink gathers sensory data from nodes. Constrained by the wireless power, sensors are usually located near their one-hop neighbors. First, Topology Matrix for One-Hop (TM-1H) nodes H is defined as:



(a) Low-rank feature (b) Temporal stability feature (c) Spatial correlation feature

Fig. 3. Environmental features mining from the selected datasets.

$$H = (h(y,z))_{n\times n} = \begin{cases} 1 & \text{if } y \text{ and } z \text{ are 1-hop neighbors;} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $y = 1, 2, \cdots, n$, $z = 1, 2, \cdots, n$. Both rows and columns in a TM-1H represent sensor nodes, and $h(y,z)$ represents whether the node y and node z are one-hop neighbor or not. The TM-1H demonstrates the binary relationship between nodes, so H is an $n \times n$ symmetry matrix.

Then, the spatial correlation at node i and time slot j is measured by computing the normalized difference between the value of a node and the average value of its all one-hop neighbors $\blacktriangledown x(i,j)$:

$$\blacktriangledown x(i,j) = \frac{x(i,j) - (H_{(i)}X^{(j)} / \sum H_{(i)})}{max(x(I,J)) - min(x(I,J))}, \tag{8}$$

where $H_{(i)}$ is the i-th row of matrix H, $X^{(j)}$ is the j-th column of matrix X. $H_{(i)}X^{(j)}$ depicts the sum values of all one-hop neighbors of node i at time slot j. $\sum H_{(i)}$ represents the number of one-hop neighbors of node i. $max(x(I,J))$ and $min(x(I,J))$ are the maximum and minimum environmental value in the EM, and $max(x(I,J)) - min(x(I,J))$ stands for the maximal difference value.

The CDF of $\blacktriangledown x(i,j)$ is plotted in Fig. 3(c). The X-axis presents the normalized difference between value of one node and the average value of its all one-hop neighbors, *i.e.*, $\blacktriangledown x(i,j)$. The Y-axis presents the cumulative probability. We find that no matter in which dataset, $> 95\%$ the value of $\blacktriangledown x(i,j)$ is < 0.3. These results imply that real environments also have the feature of spatial correlation, *i.e.*, the value of a node is similar to the value of its neighbors. Thus, the compressive sensing based estimation approach can be also optimized by space feature.

VI. Environmental Space Time Improved Compressing Sensing (ESTI-CS) Approach

We propose a novel missing data estimation approach to address ERSN problem. The proposed algorithm, namely *environmental space time improved compressive sensing* (ESTI-CS), takes into consideration the spatio-temporal features to optimize the estimation accuracy.

A. Compressive Sensing Based Approach Design

Compressive sensing, which can tolerate high data loss, is a potential approach for ERSN. Mathematically, CS based approach can only be applied to sparse matrices. Furthermore, a low-rank matrix can be well approximated by a sparse matrix. Since we have revealed the low-rank structure in most real environment datasets, we propose to use CS method to estimate missing data from SM.

The goal of solving ERSN problem is to estimate $\hat{X}$. According to Eqn. (5), any matrix can be decomposed by SVD into $\sum_{i=1}^{min(n,t)} \sigma_i u_i v_i^T$. Through the inverse process, we can also create a r-rank approximation $\hat{X}$ by using only the r largest singular values and abandoning the others:

$$\sum_{i=1}^{r} \sigma_i u_i v_i^T = \hat{X}. \tag{9}$$

This $\hat{X}$ is known as the best r-rank approximation that minimizes the error measured by Frobenius norm. Nevertheless, the optimal $\hat{X}$ cannot be obtained directly by this way as we do not know matrix X and the proper rank in advance.

Thus we propose to find $\hat{X}$ as follows:

$$\begin{aligned} &\text{Objective:} \quad min(rank(\hat{X})), \\ &\text{Subject to:} \quad B \cdot \hat{X} = S. \end{aligned} \tag{10}$$

We make this assumption according to two reasons. On the one hand, since RM is generated from SM, it is reasonable to be as close as SM. On the other hand, like EM, RM should also have a low-rank structure. Given this, it is still difficult to solve this minimization problem because it is non-convex. To bypass this difficulty, we take advantage of the SVD-like factorization, which re-writes Eqn. (5) as:

$$\hat{X} = U\Sigma V^T = \mathscr{L}\mathscr{R}^T, \tag{11}$$

where $\mathscr{L} = U\Sigma^{1/2}$ and $\mathscr{R} = V\Sigma^{1/2}$. Substituting Eqn. (11) to Eqn. (10), we can solve the minimization problem according to the compressive sensing theory in [5], [7]. Specifically, if the restricted isometry property holds [19], minimizing the nuclear norm can result to rank minimization exactly for a low-rank matrix. Hereby, we just need to find matrix $\mathscr{L}$ and

$\mathscr{R}$ that minimize the summation of their Frobenius norms:

$$\begin{aligned} &\text{Objective:} \quad min(||\mathscr{L}||_F^2 + ||\mathscr{R}^T||_F^2) \\ &\text{Subject to:} \quad B \cdot (\mathscr{L}\mathscr{R}^T) = S, \end{aligned} \tag{12}$$

Looking for $\mathscr{L}$ and $\mathscr{R}$ that strictly satisfy Eqn. (12) is likely to fail due to two reasons. First, real EMs usually approximate low-rank but not exact low-rank. Second, noises in sensory data may lead to the over-fitting problem if strict satisfaction is required. Thus, instead of solving Eqn. (12) directly, we solve the following equation using the Lagrange multiplier method:

$$min(||B \cdot (\mathscr{L}\mathscr{R}^T) - S||_F^2 + \lambda(||\mathscr{L}||_F^2 + ||\mathscr{R}^T||_F^2)), \tag{13}$$

where the Lagrange multiplier λ allows a tunable tradeoff between rank minimization and accuracy fitness. This solution provides the low-rank approximation but not strict satisfaction.

In Eqn. (13), 1) B and S are known, 2) any $||\cdot||_F^2$ is non-negative, 3) the optimal values approximate 0 by minimizing all non-negative parts. Hence, $\mathscr{L}$ and $\mathscr{R}$ can be estimated in this optimization problem under the tuning of λ.

B. Environmental Spatio-Temporal Improvement

ESTI-CS includes two key components: 1) compressive sensing based method for estimating massive missing values and 2) environmental spatio-temporal improvement for increasing the accuracy against diverse loss patterns. On the one hand, the compressive sensing method relies on the low-rank structure. On the other hand, after exploiting the temporal stability and spatial correlation features, we complete ESTI-CS approach by developing Eqn. (13) as following:

$$\begin{aligned} min(||B \cdot (\mathscr{L}\mathscr{R}^T) - S||_F^2 + \lambda(||\mathscr{L}||_F^2 + ||\mathscr{R}^T||_F^2) \\ +||\mathbb{H}\mathscr{L}\mathscr{R}^T||_F^2 + ||\mathscr{L}\mathscr{R}^T\mathbb{T}||_F^2), \end{aligned} \tag{14}$$

where $\mathbb{H}$ and $\mathbb{T}$ are the spatial and temporal constraint matrices respectively. We set $||\mathbb{H}\mathscr{L}\mathscr{R}^T||_F^2$, $||\mathscr{L}\mathscr{R}^T\mathbb{T}^T||_F^2$, and $||B \cdot (\mathscr{L}\mathscr{R}^T) - S||_F^2$ to be equal in the similar order of magnitude, otherwise, they may overshadow the others when solving Eqn. (14).

Temporal stability improvement: The temporal constraint matrix $\mathbb{T}$ captures the temporal stability feature, which outlines that the change between two consecutive time slots is small. Hence, we set $\mathbb{T} = Toeplitz(0, 1, -1)_{t\times t}$. The Toeplitz matrix is defined with central diagonal given by 1, and the first upper diagonal given by -1, and the others given by 0. i.e.,

$$\mathbb{T} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ 0 & 0 & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{t\times t}. \tag{15}$$

This Toeplitz matrix adds the temporal constraint into the estimation. Importing $||\mathscr{L}\mathscr{R}^T\mathbb{T}||_F^2$ into Eqn. (14) is equal to induct an additional constraint into the original optimization problem. Since the temporal constraint is an inherent feature of environment, this additional constraint can filter more noises and errors in $\mathscr{L}\mathscr{R}^T$ estimation.

Spatial stability improvement: The spatial constraint matrix $\mathbb{H}$ captures the spatial correlation feature, which reveals that values among one-hop neighbors nodes are usually similar. Hence, we set $\mathbb{H}$ to be a row-normalized H^*, where $H^* = H + D$. The matrix H is a TM-1H, i.e., the one-hop topology matrix mentioned before. And D is an $n \times n$ diagonal matrix, which is defined with central diagonal given by $diag(d_1, d_2, \ldots d_n)$, and the others given by 0. In D, $d_i = -\sum H_{(i)}$. *e.g.*, if there is a TM-1H:

$$H = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \tag{16}$$

then,

$$H^* = H + D = \begin{bmatrix} -2 & 1 & 0 & 1 \\ 1 & -3 & 1 & 1 \\ 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & -2 \end{bmatrix}, \tag{17}$$

thus, the corresponding spatial constraint matrix is:

$$\mathbb{H} = \begin{bmatrix} 1 & -1/2 & 0 & -1/2 \\ -1/3 & 1 & -1/3 & -1/3 \\ 0 & 1 & -1 & 0 \\ -1/2 & -1/2 & 0 & 1 \end{bmatrix}. \tag{18}$$

The spatial correlation constraint is added by the matrix $\mathbb{H}$. Computing the result of $\mathbb{H}X$ is to get the matrix of differences between the elements and the average value of their one-hop neighbors in X. As the same purpose of time improvement part, we introduce the part of minimizing $||\mathbb{H}\mathscr{L}\mathscr{R}^T||_F^2$ into Eqn. (14). It also takes advantage of the inherent environment feature as an additional constraint in optimization problem, which leads to a more accurate estimation of $\mathscr{L}\mathscr{R}^T$, *i.e.*, $\hat{X}$.

C. ESTI-CS Algorithm

We propose an efficient ESTI-CS algorithm to solve the estimation in the optimization problem Eqn. (14). The detail pseudo code of this algorithm is shown in Algorithm 1.

First, we scale the $\mathbb{T}$ and $\mathbb{H}$ as all $||\cdot||_{F2}$ in Eqn. (14) have the similar order of magnitude. Hence, they will not overshadow each other when optimizing. The scaling method is similar to [25]. Then ESTI-CS algorithm solves the optimization in an iterative manner. $\mathscr{L}$ is initialized randomly, so $\mathscr{R}$ can be computed by solving the following contradictory equation:

$$\begin{bmatrix} B \cdot (\mathscr{L}\mathscr{R}^T) \\ \sqrt{\lambda}\mathscr{R}^T \end{bmatrix} = \begin{bmatrix} S \\ 0 \end{bmatrix}. \tag{19}$$

This equation can be rewritten as:

$$\begin{bmatrix} Diag(B_{(i)}) * \mathscr{L}\mathscr{R}_{(i)}^T \\ \sqrt{\lambda}\mathscr{R}_{(i)}^T \end{bmatrix} = \begin{bmatrix} S_{(i)} \\ 0 \end{bmatrix}, \tag{20}$$

where i ranges from 1 to t. This is a combination of multiple standard linear least squares problems. We then have

Algorithm 1 ESTI-CS algorithm

Input:
$S_{n\times t}$: sensory matrix
$B_{n\times t}$: binary index matrix
r: rank bound
λ: tradeoff coefficient
$MaxIter$: iteration times

Output:
$\hat{X}_{n\times t}$: estimated environment matrix

Main procedure:
$\mathscr{L} \leftarrow random\_matrix(n, r)$;
for 1 to MaxIter **do**
  $\mathscr{R} \leftarrow myInverse(B, \mathscr{L}, \lambda, r, S)$
  $\mathscr{L} \leftarrow myInverse(B^T, \mathscr{R}^T, \lambda, r, S^T)$
  $v \leftarrow ||B \cdot (\mathscr{L}\mathscr{R}^T) - S||_F^2 + \lambda(||\mathscr{L}||_F^2 + ||\mathscr{R}^T||_F^2) + ||\mathbb{H}\mathscr{L}\mathscr{R}^T||_F^2 + ||\mathscr{L}\mathscr{R}^T\mathbb{T}^T||_F^2$
  if $v < \hat{v}$ **then**
    $\hat{\mathscr{L}} \leftarrow \mathscr{L}; \hat{\mathscr{R}} \leftarrow \mathscr{R}; \hat{v} \leftarrow v$;
  end if
end for
$\hat{X} \leftarrow \hat{\mathscr{L}}\hat{\mathscr{R}}^T$;
return $\hat{X}$;

// return solution to contradictory equation
Procedure Y=myInverse(B,L,λ,r,S):
for i=1 to t **do**
  $P_i \leftarrow [Diag(B(:,i)) * \mathscr{L}; \sqrt{\lambda} * I_r]$
  $Q_i \leftarrow [S(:,i); \mathbf{0}_r]$
  $Y(:,i) = (P_i^T * P_i) \backslash (P_i^T * Q_i)$
end for
return Y;

$R_{(i)}^T = (P_i^T P_i) \backslash (P_i^T Q_i)$, where $P_i = [Diag(B_{(i)}) * \mathscr{L}; \sqrt{\lambda} I_r]$ and $Q_i = [S_{(i)}; \mathbf{0}_r]$. This procedure is reflected by the subfunction `myInverse` in the pseudo code. Similarly, $\mathscr{L}$ can be computed by fixing $\mathscr{R}_{\mathscr{T}}$. This process repeats until the optimal value is reached.

We analyze the complexity of the ESTI-CS algorithm. The key operation is the procedure for computing the inverse matrix, which provides the best approximate solution to the contradictory equation. This procedure is completed by a matrix multiplication. Thus, its time complexity is $O(rnt)$. Since ESTI-CS repeats the procedure for k times, the total complexity is $O(rntk)$. From our evaluation experience in Section VII, $\mathscr{L}$ and $\mathscr{R}^T$ converge after 5 iterations.

D. Design Optimization

There are two parameters in ESTI-CS algorithm, i.e., rank bound r and tradeoff coefficient λ. These two parameters influence the quality of $\hat{X}$ estimation. The genetic algorithm in [14] is adopted to derive the optimal rank bound r and tradeoff coefficient λ. ER is served as fitness in that algorithm. The optimal parameters will be obtained when the fitness is stalled after several generations.

VII. Performance Evaluation

A. Methodology

The proposed ESTI-CS approach is compared with existing algorithms for missing data interpolation for environmental reconstruction in WSN.

Since the performance evaluation needs complete EMs X to compute ER, in our experiment, we use the pre-processed datasets as shown in Table II. Six EMs are adopted: indoor-temp, indoor-light, forest-temp, forest-light, ocean-temp and ocean-light.

To verify the effectiveness ESTI-CS, we choose other classic interpolation methods for comparison. They are compressive sensing (CS) [14], Delaunay Triangulation (DT) [12], Multi-channel Singular Spectrum Analysis (MSSA) [26], and K-Nearest Neighbor (KNN) [6]. The parameter K in KNN is set to be $\sum_{i=1}^{n} H_{(i)}/n$. The parameter M in MSSA is set to 32 as suggested by [26].

The procedure of simulation is: 1) Generate BIM B according to four loss patterns. 2) Compute SM S according to Eqn. (2) $S = X \cdot B$. 3) All interpolation algorithms being tested take SMs as input and generate RMs. 4) The accuracy metric ER is computed for all these algorithms and datasets. And finally, these errors are compared for performance evaluation.

Two series of experiments are evaluated. The basic experiment measures the performance of different algorithms against typical random loss probability. And the second experiment evaluates the performance in diverse loss patterns.

B. Performance Analysis: Basic Comparison

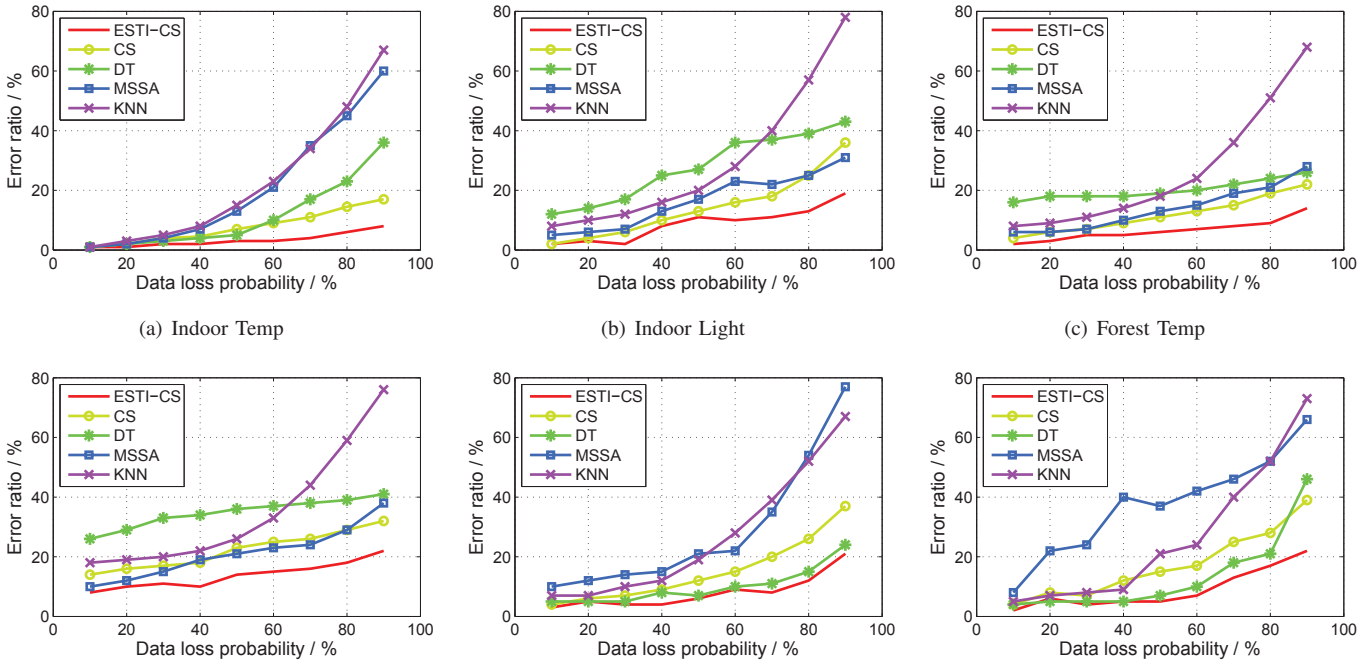
In the basic comparison, we test the error ratios among algorithms on the element random loss (ERL) pattern. The data loss rate p_{ERL} ranges from 10% to 90%. If the loss rate is 0, *i.e.*, the dataset is complete, it is unnecessary to be interpolated. If the loss rate is raised to 100%, *i.e.*, all data are lost, no methods can work.

Fig. 4 shows basic comparison results. The X-axis presents the data loss probability, and the Y-axis is the value of ER. In general, ER increases with the data loss rate.

In the indoor-temp, ESTI-CS shows the best performance. Even 90% data have been lost, ESTI-CS still can reconstruct the environment with ER$\leq 10\%$. While ER of CS is about 19%, DT is close to 38%, and ER of KNN and MSSA are more than 60%. ESTI-CS is much better than other algorithms in this scenario. In the indoor-light, ESTI-CS still outperforms the others, but the advantage is less significant than that in indoor-temp. The reason is that the indoor temperature change has strong spatio-temporal feature. However, the change of indoor light is largely influenced by the light switch. So the indoor light dataset observes more artificial changes than spatio-temporal stability.

The performance of Forest-Temp and the Forst-Light are similar. ESTI-CS achieves the best environment reconstruction among the five algorithms. CS, MSSA and DT fall behind ESTI-CS a little. KNN is not bad when $p_{ERL} < 50\%$, but when $p_{ERL} > 50\%$, ER of KNN increases quickly.

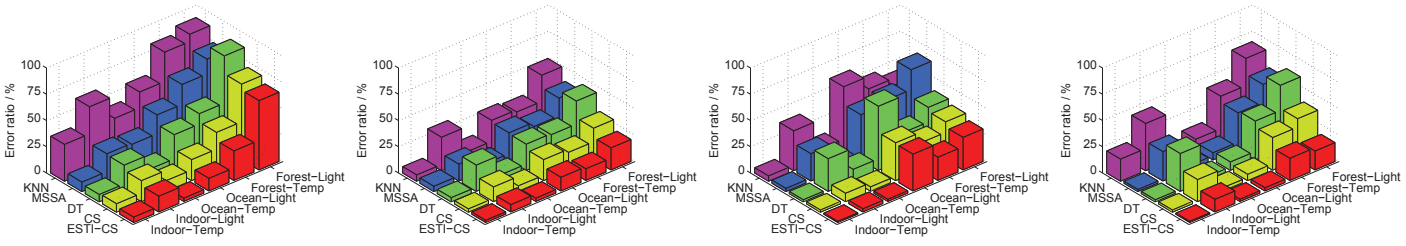
In the ocean-temp, ESTI-CS and DT produce the similar performance. When the data loss is 90%, they achieve ER$< 30\%$. Meanwhile, the ERs of CS, KNN and MSSA are bigger. In the ocean-light, the performance of ESTI-CS and DT are similar with the range of loss rate from 10% to 80%. When the loss rate increases to 90%, ER of DT also increases rapidly,



(a) Indoor Temp (b) Indoor Light (c) Forest Temp

(d) Forest Light (e) Ocean Temp (f) Ocean Light

Fig. 4. Error ratio performance of five algorithms in the basic data loss pattern: element random loss.



(a) Block Random Loss (b) Element Frequent Loss in Row (c) Successive Element Loss in Row (d) Combinational Loss

Fig. 5. Error ratio performance in different loss patterns.

and ER of ESTI-CS still keeps within 22%. These two figures indicate that ESTI-CS and DT perform better than CS, KNN and MSSA in this outdoor and small-scale WSN scenario.

Overall, ESTI-CS obtain lower interpolation error, which can be used in almost all tested datasets with different loss ratios. KNN and DT produce similar but the poor ER performance, because both of them interpolate with only the space relation among nodes but no time relation consideration. CS and MSSA are better than KNN and DT, but still worse than ESTI-CS. Especially, at the high data loss cases (data loss $\geq 80\%$), ESTI-CS exhibits an evident advantage over other algorithms. In all dataset, ESTI-CS can successfully achieve an environment reconstruction with 20% error when there are 90% data are missing.

C. Performance Analysis: Data Loss Patterns Comparison

In Fig. 5, we plot the comparison histograms of five algorithms for reconstructing the environment with four different data loss patterns.

In the simulation for BRL pattern, each of the six EMs is set to lose data with the block pattern as shown in Fig. 1(b). The scale and the number of the blocks are random, but the amount of total data loss is 50% in this simulation. In Fig. 5(a), most algorithms in most EMs perform not well. For example, in forest-light, ER of all algorithms are bigger than 60%. The reasons are 1) in the forest, many shadows disturb the spatio-temporal stability. 2) if large blocks of data lose, spatio-temporal optimized estimation is helpless either. These two reasons lead to the result. However, in indoor-temp, ocean-temp, and ocean-light, the environment changes are smoothly, ER of ESTI-CS are less than 5% despite 50% BRL data loss. Even indoor-light, forest environments, ESTI-CS is still a bit better than the others. In addition, we find that KNN is in big trouble for estimating the missing data in BRL.

In the simulation of EFLR pattern, the rows are randomly selected, the loss frequency in these rows is set $> 75\%$, and the

totally lose data in matrix is set 50%. We find that the results in Fig. 5(b) are close to the basic comparison, because the data loss in EFLR is similar to ERL pattern. In EFLR, the temporal optimization can contribute a partial effect, but the space optimization still works. So our ESTI-CS still outperforms CS, KNN, DT, and MSSA.

In the simulation of SELR pattern, the starting points are randomly selected, and then all elements after the starting points in those rows are lost as shown in Fig. 1(d). The total amount of data loss is set to be 50%. The results of all algorithms are between those in EFLR and BRL. For ESTI-CS, these results are reasonable. The reason is that ESTI-CS can only use space optimization, but the time optimization has no effect due to elements lost in all time of a node. Note that all algorithm plays not well for ocean-light in this simulation. Since the scale of ocean-light is small, after some additional rows are lost, it becomes smaller, which is hard to be estimated.

In the simulation of Combinational Loss pattern, we set $20\%\text{ERL} + 10\%\text{BRL} + 10\%\text{EFLR} + 10\%\text{SELR}$. The results of five algorithms are shown in Fig. 5(d). The ER of ESTI-CS is $\leq 20\%$ in any dataset in the combinational loss pattern.

In summary, ESTI-CS outperforms CS, KNN, DT and MSSA in any data loss pattern.

VIII. Conclusion

In this paper, we studied the environmental data loss and reconstruction problem in WSN. We verified the massive data loss in real datasets and modeled the special data loss patterns of WSN. Then, we mined the spatial, temporal, and low-rank features from WSN datasets. By drawing on these observations, we designed the ESTI-CS algorithm to estimate the missing data. The proposed algorithm combines the benefits of compressive sensing and environmental space-time features. Trace-driven experiments illustrated that ESTI-CS outperforms existing interpolation methods. Notably, ESTI-CS can achieve an effective reconstruction with 20% error in face of 90% missing values in the original data.

There are three avenues for our future work. First, exploit the correlations between multiple environmental factors to further improve the accuracy of estimation. *e.g.*, light and temperature have compact correlation in many scenarios. Second, study the tradeoff between the computation time and accuracy in environment reconstruction. Third, data interpolation in mobile sensor networks is also interesting and challenging.

Acknowledgment

This research was supported by NSF of China under grant No. 61073158, No. 61100210, STCSM Project No. 12dz1507400, Doctoral Program Foundation of Institutions of Higher Education under grant No. 20110073120021. This work was also supported in part by the FQRNT grant 131844.

References

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless Sensor Networks: A Survey," Elsevier Computer Networks, vol. 38(4), pp. 393-422, March 2002.

[2] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," In Proc. of ACM IPSN, Nashville, Tennessee, April, 2006.

[3] M. Balazinska, A. Deshpande, M. J. Franklin, P. B. Gibbons, J. Gray, M. Hansen, M. Liebhold, S. Nath, A. Szalay, and V. Tao, "Data Management in the Worldwide Sensor Web," IEEE Pervasive Computing, vol. 6(2), pp. 30-40, April-June 2007.

[4] E. Candes, and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?" IEEE Trans. on Information Theory, vol. 52(12), pp. 5406 - 5425, December 2006.

[5] E. J. Candes, J. Romberg, and T. Tao, "Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information," IEEE Trans. on Information Theory, vol. 52(2), pp. 489 - 509, February 2006.

[6] T. Cover, and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. on Information Theory,vol. 13(1), pp. 21-27, January 1967.

[7] D. Donoho, "Compressed sensing," IEEE Trans. on Information Theory, vol. 52(4), pp. 1289 - 1306, April 2006.

[8] S. Floyd, and V. Jacobson, "Random Early Detection Gateways for Congestion Avoidance," IEEE/ACM Trans. on Networking, vol. 1(4), pp. 397-413, August 1993.

[9] T. He, S. Krishnamurthy, J. A. Stankovic, T. F. Abdelzaher, L. Luo, R. Stoleru, T. Yan, L. Gu, J. Hui, B. H. Krogh, "Energy-Efficient Surveillance System Using Wireless Sensor Networks," In Proc. of ACM MOBISYS, Boston, MA, USA, 2004.

[10] Intel Indoor Test Data: http://www.select.cs.cmu.edu/data/labapp3/index.html.

[11] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, "Impact of Interference on Multi-Hop Wireless Network Performance," In Proc. of ACM MOBICOM, San Diego, CA, USA, 2003.

[12] L. Kong, D. Jiang, and M.-Y. Wu, "Optimizing the Spatio-Temporal Distribution of Cyber-Physical Systems for Environment Abstraction," In Proc. of IEEE ICDCS, Genoa, Italy, 2010.

[13] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, andN. Taft, "Structural Analysis of Network Traffic Flows," In Proc. of ACM SIGMETRICS, New York, NY, USA, 2004.

[14] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive Sensing Approach to Urban Traffic Sensing," In Proc. of IEEE ICDCS, Minneapolis, MN, USA, 2011.

[15] P. Li, T. J. Hastie, and K. W. Church, "Very Sparse Random Projections," In Proc. of ACM KDD, Philadelphia, PA, USA, 2006.

[16] C. Luo, F. Wu, J. Sun, C. Chen, "Compressive Data Gathering for Large-Scale Wireless Sensor Networks," In Proc. of ACM MOBICOM, 2009.

[17] L. Mo, Y. He, Y. Liu, J. Zhao, S. Tang, X. Li and G. Dai, "Canopy Closure Estimates with GreenOrbs: Sustainable Sensing in the Forest," In Proc. of ACM SENSYS, Berkeley, CA, USA, 2009.

[18] S. Rallapalli, L. Qiu, Y. Zhang, and Y. Chen, "Exploiting Temporal Stability and Low-Rank Structure for Localization in Mobile Networks," In Proc. of ACM MOBICOM, Chicago, IL, USA, 2010.

[19] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization," SIAM Review, 2007.

[20] C. Shen, C. Srisathapornphat, and C. Jaikaeo, "Sensor Information Networking Architecture and Application," IEEE Personal Communications, vol. 8(4), pp. 52-59, August 2001.

[21] W. Wang, M. Garofalakis, and K. Ramchandran, "Distributed sparse random projections for refinable approximation," In Proc. of ACM IPSN, Cambridge, MA, USA, 2007.

[22] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees and M. Welsh, "Fidelity and Yield in a Volcano Monitoring Sensor Network," In Proc. of USENIX OSDI, Seattle, WA, USA, 2006.

[23] A. Woo, and D. Culler, "A Transmission Control Scheme for Media Access InSensor Networks," In Proc. of ACM MOBICOM, Rome, Italy, 2001.

[24] Z. Yang, M. Li, and Y. Liu, "Sea Depth Measurement with Restricted Floating Sensors," In Proc. of IEEE RTSS, Tucson, AR, USA, 2007.

[25] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu, "Spatio-Temporal Compressive Sensing and Internet Traffic Matrices," In Proc. of ACM SIGCOMM, Barcelona, Spain, 2009.

[26] H. Zhu, Y. Zhu, M. Li, and L. Ni, "Seer: Metropolitan-Scale Traffic Perception Based on Lossy Sensory Data," In Proc. of IEEE INFOCOM, Rio de Janeiro, Brazil, 2009.