Recover Corrupted Data in Sensor Networks: A Matrix Completion Solution

Kun Xie, Xueping Ning, Xin Wang, *Member, IEEE*, Dongliang Xie, Jiannong Cao, *Fellow, IEEE*, Gaogang Xie, and Jigang Wen

Abstract—Affected by hardware and wireless conditions in WSNs, raw sensory data usually have notable data loss and corruption. Existing studies mainly consider the interpolation of random missing data in the absence of the data corruption. There is also no strategy to handle the successive missing data. To address these problems, this paper proposes a novel approach based on matrix completion (MC) to recover the successive missing and corrupted data. By analyzing a large set of weather data collected from 196 sensors in Zhu Zhou, China, we verify that weather data have the features of low-rank, temporal stability, and spatial correlation. Moreover, from simulations on the real weather data, we also discover that successive data corruption not only seriously affects the accuracy of missing and corrupted data recovery but even pollutes the normal data when applying the matrix completion in a traditional way. Motivated by these observations, we propose a novel Principal Component Analysis (PCA)-based scheme to efficiently identify the existence of data corruption. We further propose a two-phase MC-based data recovery scheme, named MC-Two-Phase, which applies the matrix completion technique to fully exploit the inherent features of environmental data to recover the data matrix due to either data missing or corruption. Finally, the extensive simulations with real-world sensory data demonstrate that the proposed MC-Two-Phase approach can achieve very high recovery accuracy in the presence of successively missing and corrupted data.

Index Terms—Corrupted data recovery, matrix completion, wireless sensor networks

1 Introduction

Wireless sensor networks (WSNs) are widely utilized to gather various environmental information, such as under water [1], in forests [2], along road [3], and on volcanoes [4]. In WSNs, the data collected from the monitoring of the dynamic environment can generally be represented by an $N \times T$ Environment Matrix (EM), which records data from N sensors over T time slots. Events occurred in the physical world, such as forest fire, earthquake or chemical spill, cannot be accurately detected using inaccurate and incomplete sensory data [5]. Thus, it is extremely important to obtain the full and accurate EM from raw sensory data before making any further analysis and decision.

- *K. Xie is with the College of Computer Science and Electronics Engineering, Hunan University, Changsha 410082, China, and the Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY 11794. E-mail: xiekun@hnu.edu.cn.*
- *X. Ning is with the College of Computer Science and Electronics Engineering, Hunan University, Changsha 410082, China. E-mail: ningxueping@gmail.com.*
- *X. Wang is with the Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY 11794. E-mail: xwang@ece.sunysb.edu.*
- *D. Xie is with the Department of Electrical and Computer Engineering, State University of New York, Stony Brook, NY 11794, and the State Key Lab of Networking & Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: xiedl@bupt.edu.cn.*
- *J. Cao is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. E-mail: csjcao@comp.polyu.edu.hk.*
- *G. Xie and J. Wen are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China. E-mail: {xie, wenjigang}@ict.ac.cn.*

Manuscript received 9 July 2015; revised 16 Mar. 2016; accepted 11 July 2016. Date of publication 29 July 2016; date of current version 31 Mar. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TMC.2016.2595569

Affected by hardware and severe wireless conditions [6], [7], [8] such as strong fading in WSNs, raw sensory data can have notable loss and corruption. Data generated by WSNs may also be unreliable and inaccurate as a result of the limitation in sensor resources such as energy, memory, computational capacity, and wireless bandwidth. Specially, when the battery power of a sensor is exhausted, the probability of generating erroneous data will grow rapidly [9]. In addition, in harsh and unattended environments, some sensor nodes may malfunction and result in noisy, faulty, missing and redundant data. Furthermore, sensor nodes are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping [10], in which data generation and processing will be manipulated by adversaries. The above internal and external factors make it difficult to obtain accurate EM data.

Several studies have been made to handle missing data, through methods such as local interpolation based on K-Nearest Neighbors (KNN) [11], global refinement through Delaunay Triangulation (DT) [12], Multi-channel Singular Spectrum Analysis (MSSA) based on principal component analysis [13]. However, the above data interpolation techniques may not well uncover spatio-temporal correlations in EM, and the interpolation quality is generally not high. Moreover, designed for missing data interpolation, these schemes cannot be applied to well handle the data corruption.

With the rapid progress of sparse representation, matrix completion (MC) [14], [15], [16], a remarkable new field, has emerged recently. According to the matrix completion theory, a matrix can be accurately recovered with a relatively small number of entries if the underlying matrix has a low-rank or approximately low-rank structure. Different from [11], [12], [13], matrix completion seeks to find a low-

rank matrix which agrees well with the observed entries of a matrix with incomplete data. Well exploiting spatio-temporal correlations, matrix completion can achieve good interpolation performance for random data missing.

Due to channel fading and sensor failures, successive data missing or corruption may occur along the column (temporal) and/or row (spatial) directions of EM, which makes it a big challenge to apply the matrix completion to accurately recover the whole matrix [17]. Although MC theory allows the recovering of a matrix with random missing entries, if a row or a column is completely lost, MC operation does not have an effect on these missing entries. Besides missing data, data corruption is also unavoidable in WSNs. Although the corruption of data and its impact on data recovery are discussed in [18], [19], they mainly consider the interpolation of missing data while not identifying and correcting the corrupted data.

This paper focuses on designing algorithms to detect the existence of erroneous data and more accurately recover the environmental data matrix in the presence of successive data missing or corruption. Our contributions can be summarized as follows:

- We first analyze large traces of real weather data, and our analysis verifies that weather data have the features of low-rank, temporal stability, and spatial correlation. Moreover, from simulations on the real weather data, we also discover that successive data corruption not only seriously impacts the accuracy in the recovery of missing and corrupted data but even pollutes the normal data when applying matrix completion in a traditional way.
- We propose a two-phase environmental data recovery scheme, named MC-Two-Phase, which takes advantage of the inherent features of the measurements to recover the monitoring data either due to data missing or corruption based on matrix completion. Specifically, in the first phase, our algorithm recovers the remaining data matrix by excluding the successively corrupted data to avoid their negative effect. In the second phase, we take the data obtained in the first phase and exploit matrix completion theory to take full advantage of *both the spatial and temporal stability* to recover the whole matrix.
- To accurately recover the EM data, in the MC-Two-Phase scheme, we propose three algorithms: a structure-fault detection algorithm based on Principal Component Analysis (PCA), a spatial pre-interpolation algorithm, and a temporal pre-interpolation algorithm. The results of these algorithms are fully integrated with the MC to more reliably recover an EM.
- Through comprehensive simulations based on real data traces, we show that our MC-Two-Phase scheme can accurately recover weather data even with a large amount of data missing or corruption. The error ratios on the missing data, the corrupted data, and the normal data under our MC-Two-Phase (under singular value thresholding (SVT)) are only 1, 2, and 1 percent of those under the conventional SVT.

To the best of our knowledge, this is the first work that exploits matrix completion to recover sensory matrix with successive data missing or corruption.

The rest of this paper is organized as follows. We introduce the related work in Section 2. The fundamentals of matrix completion and problem formulation are presented in Section 3. We introduce our empirical study with real weather data in Section 4. In Sections 5 and 6, we present our proposed PCA-based scheme for fault detection and our algorithm on corrupted data recovery, respectively. Finally, we evaluate the performance of the proposed MC-Two-Phase through extensive simulations in Section 7, and conclude the work in Section 8.

2 Related Work

Data missing and corruption are unavoidable during data gathering in WSNs.

A great deal of existing work has been devoted to interpolate missing data. Among these, K-Nearest-Neighbor (KNN) [11] simply utilizes the values of the nearest K neighbors to estimate a missing data value. As a classical local interpolation method, KNN is frequently applied in many low-fidelity estimation cases. Delaunay Triangulation (DT) method [12] treats the gathered data as vertices, and rebuilds virtual triangles for data interpolation by taking advantage of the vertices and their global errors. As a typical global refinement method, DT is widely adopted in computer vision for surface rendering. Multi-channel Singular Spectrum Analysis (MSSA) [13] is a data adaptive and nonparametric method based on the embedded lag-covariance matrix. MSSA is often applied in geographic data and meteorological data recovery. The above methods are only suitable for data interpolation with very few missing values. They perform poorly when the data missing rate is high.

Besides above techniques, with the rapid progress of sparse representation, matrix completion [14], [15], [16], a remarkable new field, has emerged recently. Matrix Completion is the procedure of filling in the missing entries of a partially observed matrix. Without any restrictions on the number of degree of freedom in the completed matrix, a matrix completion problem is under-determined since the hidden entries could be assigned arbitrary values. Thus matrix completion often seeks to find the low rank matrix that matches the known entries.

It is well-known that the trace-norm is a convex surrogate to the matrix rank, based on which, Candès et al. [14] proposes to reconstruct a matrix which agrees well with the observed entries while regulating the trace-norm. Candès et al. also show that most $n_1 \times n_2$ matrices of rank r ($r \ll \min\{n_1, n_2\}$) can be perfectly recovered with very high probability by solving a simple convex optimization program provided that the number of samples is sufficient. Our recent research results [20], [21] show that the recovery performance of matrix completion depends on the sampling ratio of the matrix. To further reduce the sample number and thus the communication and sensing cost in data gathering process, we propose a sampling stop condition in [20] and a sampling scheduling algorithm in [21].

Besides using the trace-norm to surrogate to the matrix rank, some matrix factorization approaches are proposed for the matrix completion problem. These approaches factorizes an incomplete matrix into two (low-rank) matrices, which are further multiplied to reconstruct the original matrix and infer the missing data. The typical matrix

factorization approaches include Environmental space time improved compressive sensing (ESTICS) [22], Sparsity Regularized SVD (SRSVD) [23], Sparsity Regularized Matrix Factorization (SRMF) [24], and Low-rank matrix fitting (LMaFit) [25].

Although matrix completion techniques can also be exploited for recovering random missing data due to the reasons such as unstable wireless transmissions, the matrix can be recovered only if there is no row or column to be completely empty. When there exist successive data missing along the column and/or the row due to reasons such as fading and sensor failures, current matrix completion technique does not work.

Besides missing data, data corruption is also unavoidable in WSNs. Very few studies [18], [19] in matrix completion consider large data corruption, although matrix completion is shown provably accurate when the few observed entries are added by Gaussian noise with a small variance [26]. The method in [26] can not deal with sparse random noise in the measurement. The authors in [18] propose a method to recover the missing data with adaptive outlier pursuit when part of the measurements are damaged by outliers, under the assumption that the number of corrupted data is known. The work in [19] attempts to recover a non-corrupted column in a low-rank matrix when some columns in the matrix are corrupted. Taking iterative procedures with each iterative step involving a matrix decomposition, the solutions in [18], [19] suffer from large computation cost, and are thus difficult to apply for EM reconstruction in large-scale WSNs. Although the impact of data corruption is discussed, these existing studies mainly consider the interpolation of missing data while not identifying and correcting corrupted data.

Moreover, existing matrix completion techniques seldom discuss the corrupted data pattern. In contrast, based on the simulation results presented in Section 4.4, we discover that successive data corruption seriously impacts the accuracy of missing data recovery and even pollutes the normal data when applying matrix completion to EM in a traditional way, which bring extremely high challenge to apply matrix completion for recovery of EM.

In summary, existing work generally consider the interpolation of random missing data. In this work, we propose to actively recover successive missing or corrupted data to obtain full and accurate EM. To the best of our knowledge, none of the existing studies consider the recovery of corrupted data. We propose a two-phase matrix completion scheme to solve the problem. Our algorithm eliminates the negative effect caused by corrupted data and takes advantage of spatial-temporal features of environmental data to accurately recover EM in WSNs.

3 Preliminary and Problem Formulation

In this section, we first introduce the fundamentals of matrix completion, then present our problem formulation.

3.1 Fundamentals of Matrix Completion

Matrix completion is a new technique which can be applied to recover a low-rank matrix from a subset of the matrix entries [14], [15], [16]. Specifically, given the incomplete data matrix $M \in R^{n_1 \times n_2}$ with rank $r \ll \min\{n_1, n_2\}$, the matrix completion problem can be formulated as follows:

$$\begin{aligned} \min_{X} \quad & rank(X) \\ \text{subject to} \quad & X_{ij} = M_{ij}, (i,j) \in \Omega, \end{aligned} \tag{1}$$

where Ω is the set of locations corresponding to the observed entries.

However, solving this rank minimization problem in (1) is often impractical because it is NP-hard. Then [14] proves that most matrices M of rank r can be perfectly recovered by solving the optimization problem

$$\begin{aligned} \min_{X} \quad & \|X\|_* \\ \text{subject to} \quad & X_{ij} = M_{ij}, (i,j) \in \Omega, \end{aligned} \tag{2}$$

provided that the number of samples m be sufficient and meet the following condition

$$m \geq Cn^{6/5} r \log\ n, \tag{3}$$

where C is a numerical constant and $n = \max\{n_1, n_2\}$.

In (2), $\|X\|_*$ is the nuclear norm (trace-norm) of the matrix X, which is the sum of its singular values. That is, $\|X\|_* = \sum_{i=1}^{\min\{n_1,n_2\}} \sigma_i$ and $\sigma_i \geqslant 0$ are the singular values of X.

Many matrix recovery approaches have been proposed to solve the convex optimization problem in (2), including iterative reweighted least squares algorithm (IRLS-M) [27], Spectral Matrix Completion[28], fixed point continuation algorithm [29], OptSpace [28], FixedPoint Continuation with Approximate SVD (FPCA) [30], and singular value thresholding [31].

Our MC-Two-Phase scheme does not depend on the underlying matrix recovery approach. We choose the singular value thresholding approach as an example to illustrate our MC-Two-Phase scheme in this paper.

3.2 Problem Formulation

We define a matrix $X_{N \times T}$ to hold the weather data, which records data from N sensors over T time slots. In the weather matrix, a row corresponds to a sensing location and a column corresponds to a time slot. An entry represents the weather data for a particular sensing location and time slot.

The observed sensory matrix is defined as $M_{N \times T}$. As mentioned in the introduction, because the data gathering process is largely affected by hardware and wireless conditions in WSNs, data measured and collected by WSNs are often unreliable. As a result, the sensory matrix $M_{N \times T}$ may have some data lost and corrupted.

Since weather data normally have strong correlation between neighboring locations and time slots, the weather matrix should have low rank. This is confirmed with our trace data in the next section. Low rank feature provides the possibility for us to apply matrix completion to recover the raw weather matrix $X_{N \times T}$ from the observed $M_{N \times T}$, which is our basic idea to solve the data recovery problem.

Despite the big progress in the area of matrix completion, existing methods are often applied when there are only random data missing. However, the failure of senor node, data tampering by attackers, and the severe communication condition may cause successive/mass data corruption in both

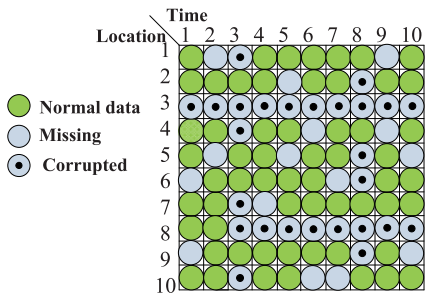


Fig. 1. Successive data corrupted with random data lost.

the matrix rows and columns, as shown in Fig. 1. Specifically, the successive data corruption in rows may be caused from the failure of the senor node or data tampering on the node. The mass data corruption in columns may result from strong channel fading or failure around the sink node or attacks to tamper data at multiple sensor nodes. In this paper, we call such successive or mass data corruption as a structure fault. The focus of this paper is to develop techniques to accurately interpolate missing data and recover corrupted data in the presence of structure fault.

Before we present our two-phase corrupted data recovery algorithm based on matrix completion in Section 6, we first analyze a large set of real weather data to better understand their structure and characteristics in the next section. We will show that the conventional matrix completion methods perform poorly because a structure fault can destroy the inherent feature in the weather matrix. It is thus very challenging to apply matrix completion theory in the practical weather gathering system.

4 Empirical Study with Traces of Weather Data

We have deployed 196 sensors to collect the weather data in Zhu Zhou, China. Fig. 2 shows the map of Zhu Zhou, where the red dot represents the location of a deployed sensor. Each sensor reports its data once an hour to the weather monitoring center via the cellular network. We have collected a large amount of weather trace data from Zhu Zhou. Each data element includes weather data of rain, temperature, and wind. Specially, we choose rain data to analyze because Zhu Zhou is in the area prone to flood. The trace data are collected in the duration of more than two years from 2011 to 2013. In our simulations, we set $N = 196$, $T = 168$. The trace data reveal the existence of some special structures.

4.1 Low-Rank Feature

Weather data collected over different locations and time slots are not independent. There exists inherent data redundancy. We first apply singular value decomposition (SVD) to examine whether the matrix has a good low-rank structure. A weather matrix $X_{N\times T}$ can be decomposed as

$$X = U\Sigma V^{tr}, \tag{4}$$

where U is an $N \times N$ unitary matrix, V is a $T \times T$ unitary matrix, V^{tr} is the transpose of V, and Σ is a $N \times T$ diagonal matrix with the diagonal elements (i.e., the singular values) organized in the decreasing order (i.e., $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_r, 0, \ldots, 0)$). The rank of a matrix X, denoted by r, is equal to the number of its non-zero singular values. A matrix is low-rank if its $r \ll \min\{N, T\}$.

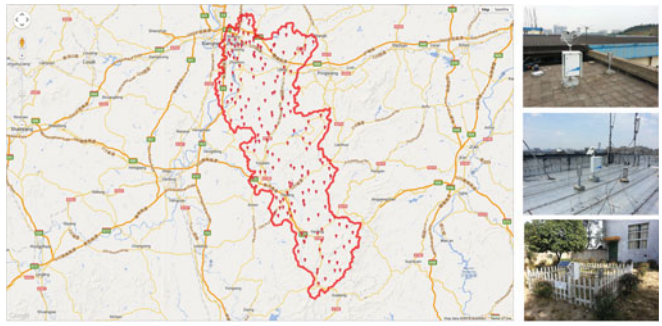


Fig. 2. Weather sensor deployment in Zhu Zhou, China.

If a matrix has low-rank, its top k singular values occupy the total or near-total energy $\sum_{i=1}^{k}\sigma_i^2 \approx \sum_{i=1}^{r}\sigma_i^2$. The metric we use is the fraction of the total variance captured by the top k singular values

$$g(k) = \sum_{i=1}^{k}\sigma_i^2 / \sum_{i=1}^{r}\sigma_i^2, \tag{5}$$

Fig. 3 plots the fraction of the total variance captured by the top k singular values for different weather trace data from different seasons. We find that the top 20 singular values capture 70-90 percent variance in the real traces. These results indicate that the data matrix X has a good low-rank approximation. The low-rank feature is the prerequisite for using matrix completion.

4.2 Temporal Stability

Weather data usually change slowly over time. To study the short-term stability of weather matrix, we calculate the gap between each pair of adjacent readings at a location. Specifically, the gap between each pair of adjacent readings captured in two consecutive time slots (j, and $j-1$) is equal to

$$T_{gap}(i,j) = \left|X_{ij} - X_{i,j-1}\right|, \tag{6}$$

where $1 \leqslant i \leqslant N$ and $2 \leqslant j \leqslant T$, X_{ij} represents the data generated at the location of sensor i at time slot j. Obviously, $T_{gap}(i,j) = 0$ if the weather data at location i is not changed from time slot $j-1$ to j. The smaller the $T_{gap}(i,j)$, the more stable the sensory readings for location i around the time slot j.

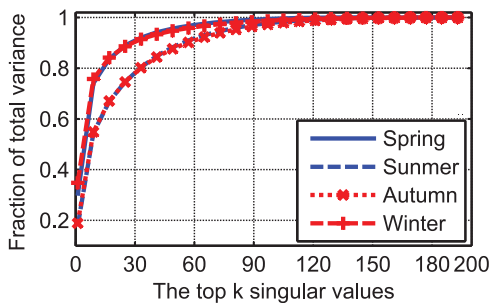


Fig. 3. Fraction captured by top k singular values.

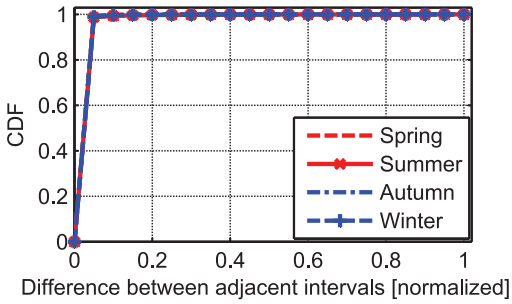


Fig. 4. Temporal stability feature.

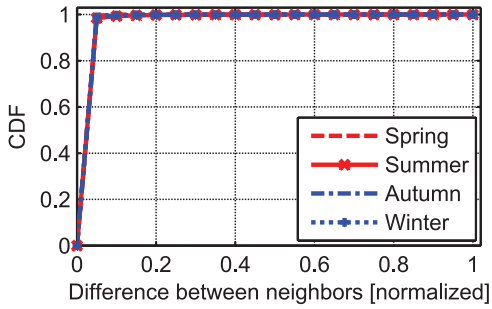


Fig. 5. Spatial correlation feature.

By computing the normalized difference values between adjacent time slots, we measure the temporal stability at node i and time slot j according to

$$\Delta T_{gap}(i,j) = \frac{\left|X_{ij} - X_{i,j-1}\right|}{\max_{1\leqslant i\leqslant N, 2\leqslant j\leqslant T}\left|X_{ij} - X_{i,j-1}\right|}, \tag{7}$$

where $\max_{1\leqslant i\leqslant N,2\leqslant j\leqslant T}\left|X_{ij} - X_{i,j-1}\right|$ is the maximal gap between any two consecutive time slots in the weather matrix.

We plot the cumulative distribution function (CDF) of $\Delta T_{gap}(i,j)$ in Fig. 4. The X-axis represents the normalized difference values between two consecutive time slots, i.e., $\Delta T_{gap}(i,j)$. The Y-axis represents the cumulative probability. We observe that more than 90 percent $\Delta T_{gap}(i,j)$ are very small (< 0.05). These results indicate that temporal stability exists in real environments. In Section 6.2, we design our temporal pre-interpolation algorithm based on this feature.

4.3 Spatial Correlation Feature

Weather data are often smooth in a small area, i.e., at a given time, the data recorded at nearby locations have similar values. The spatial correlation between a node i and its neighbors in a time slot j is measured by computing the difference between its data value and the average value of its one-hop neighbors

$$S_{gap}(i,j) = X_{ij} - \left(Y_{(i)}X^{(j)} / \sum Y_{(i)}\right), \tag{8}$$

where $Y_{(i)}$ is the ith row of matrix Y, $X^{(j)}$ is the jth column of matrix X. Y is the topology matrix and defined as

$$Y = \left(Y_{ij}\right)_{N\times N} = \begin{cases} 1 & \text{if i and j are 1-hop neighbors} \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

With the locations of all deployed weather sensors, the topology matrix is easily to obtain. Both rows and columns in a topology matrix Y represent sensor nodes, and Y_{ij} represents whether the node i and node j are one-hop neighbor. Y is an $N \times N$ symmetric matrix, which has binary values to capture the relationship between nodes.

In (8), the average data value of one-hop neighbors of node i is obtained by dividing the total values of neighbors, $Y_{(i)}X^{(j)}$, by the number of one-hop neighbors, $\sum Y_{(i)}$. The spatial correlation feature at node i and time slot j can be obtained by computing the normalized difference values between neighboring nodes

$$\Delta S_{gap}(i,j) = \frac{X_{ij} - \left(Y_{(i)}X^{(j)} / \sum Y_{(i)}\right)}{\max_{i,j}\left(X_{ij}\right) - \min_{i,j}\left(X_{ij}\right)}, \tag{10}$$

where $\max_{i,j}\left(X_{ij}\right)$ and $\min_{i,j}\left(X_{ij}\right)$ are the maximum and minimum data values in the weather matrix, and $\max_{i,j}\left(X_{ij}\right) - \min_{i,j}\left(X_{ij}\right)$ stands for the maximal difference value.

The CDF of $\Delta S_{gap}(i,j)$ is plotted in Fig. 5. The X-axis represents the normalized difference between value of one node and the average value of its one-hop neighbors, i.e., $\Delta S_{gap}(i,j)$, and the Y-axis represents the cumulative probability. No matter in which dataset, we can see that the probability of $\Delta S_{gap}(i,j) < 0.05$ is more than 95 percent, which indicates that real weather data have strong spatial correlation. In Section 6.1, we design our spatial pre-interpolation algorithm based on this feature.

4.4 Negative Effects When Structure Faults Happen

The structure faults can be classified into two categories, a row-structure fault if the successive data corruption (or successive data missing) are in a row due to the sensor error, and a column-structure fault if such data outliers are in a column due to communication faults. In order to evaluate the recovery performance for different methods and scenarios, we define the following three metrics.

Definition 1. *Error Ratio on Missing data (ERM): A metric for measuring the error in the recovery of the random missing entries in the matrix after interpolation*

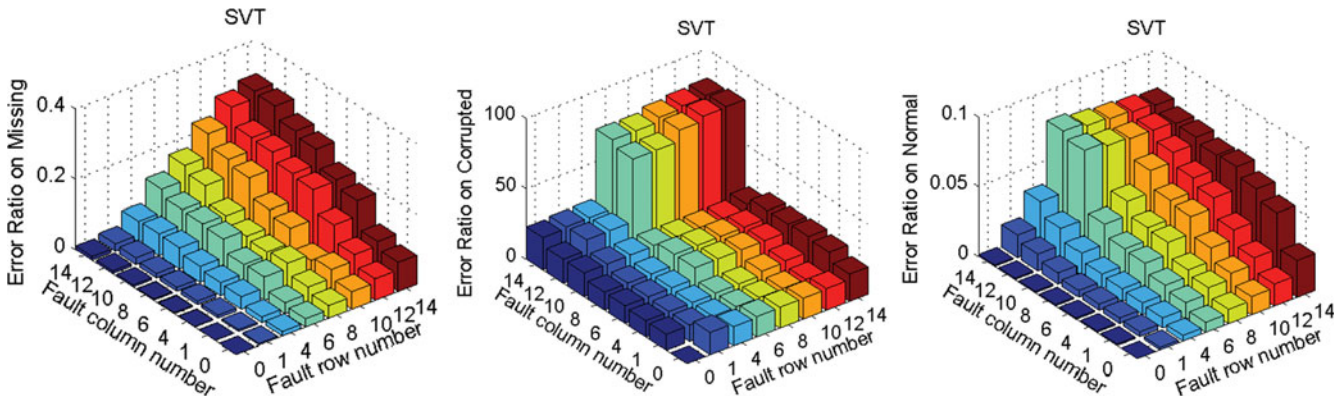
$$\varepsilon_{ERM} = \frac{\sqrt{\sum_{i,j\in\pi_l}\left(X_{ij} - \hat{X}_{ij}\right)^2}}{\sqrt{\sum_{i,j\in\pi_l} X_{ij}^2}}, \tag{11}$$

where π_l denotes missing data set in the sensory matrix, X_{ij} and $\hat{X}_{ij}$ denote the raw data and the recovered data at (i,j)th element of X. Note that the condition $i,j \in \pi_l$ in (11) indicates that only errors on missing entries are counted.

Definition 2. *Error Ratio on Corrupted data (ERC): A metric for measuring the recovery error of all corrupted entries in the matrix*

$$\varepsilon_{ERC} = \frac{\sqrt{\sum_{i,j\in\pi_c}\left(X_{ij} - \hat{X}_{ij}\right)^2}}{\sqrt{\sum_{i,j\in\pi_c} X_{ij}^2}}, \tag{12}$$

where π_c denotes corrupted data set in the sensory matrix.



(a) Error Ratio on Missing Data (b) Error Ratio on Corrupted Data (c) Error Ratio on Normal Data

Fig. 6. The recovery performance under conventional SVT when structure faults happen.

Definition 3. *Error Ratio on Normal data (ERN): A metric to measure the recovery error of all normal entries besides outliers (i.e., missing or corrupted data) in the matrix*

$$\varepsilon_{ERN} = \frac{\sqrt{\sum_{i,j \notin (\pi_l \cup \pi_c)} \left(X_{ij} - \hat{X}_{ij}\right)^2}}{\sqrt{\sum_{i,j \notin (\pi_l \cup \pi_c)} X_{ij}^2}}. \tag{13}$$

In order to investigate how structure faults impact the recovery performance, we generate data trace with random data missing as well as structure data corruption from the gathered weather data trace. Specifically, we choose the rain traces gathered from July 1 to July 7, 2012 as the raw data. We denote the raw trace data as $X_{N\times T}$. From the raw data, we generate the corrupted synthesized data, denoted as $D_{N\times T}$. The synthesized data D is generated through following steps.

Step 1. Among all the $N \times T$ entry locations in $X_{N\times T}$, 80 percent entries are randomly selected to form D. Denote the selected location set as Ω. After this step, the synthesized data D can be expressed as follows.

$$D_{ij} = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & otherwise, \end{cases} \tag{14}$$

Step 2. To generate structure data corruptions, we randomly select some rows and columns, from which 60percent successive entries on the rows and columns are set as corrupted by adding randomly generated noise. Denote the corrupted location set as Π:

$$D_{ij} = \begin{cases} D_{ij} + Z_{ij} & (i,j) \in \Pi \\ D_{ij} & otherwise, \end{cases} \tag{15}$$

where Z_{ij} is the generated noise at location (i,j) following a zero-mean normal distribution with variance δ^2, that is $Z_{ij} \sim N(0, \delta^2)$.

After the above two steps, the corrupted synthesized data matrix is obtained, and then the matrix completion is applied to the corrupted data matrix D to obtain the recovery data. Finally, we calculate the error ratio by comparing the recovered data with the raw data trace X.

Fig. 6 shows the error ratio under row-structure faults and column-structure faults. As expected, the Error Ratio increases as the number of structure faults becomes higher. As shown in Fig. 6a and 6c, structure faults have significant impact on the recovery performance of the random missing entries and even pollute normal entries. In Fig. 6b, the error ratio on the corrupted entities is very high. The error ratio even reaches 80 when the number of fault rows and the number of fault columns are 14. Even when the data matrix has only one corrupted row and column, the error ratio on the corrupted entities is still larger than 20. Therefore, we conclude that directly applying conventional matrix completion can not recover the structure corrupted data.

Moreover, these results demonstrate that structure faults seriously destroy the inherent feature of the weather matrix, which makes it difficult to directly apply conventional matrix completion to interpolate weather data. It is important and challenging to design a technique to accurately interpolate the missing data as well as recover the corrupted data in the presence of structure faults.

5 PCA-Based Structure Fault Detection

Without knowing the actual data value, it is very hard to determine if any data are corrupted. Although some efforts have been made in the literature to identify single faults, our preliminary studies indicate that it is very difficult if not completely impossible to find a structure fault which results in successively corrupted data on a row or column. In this section, we propose a PCA-based scheme to effectively detect structure fault occurred on rows or columns in the matrix.

PCA is a well-known technique for dimensionality reduction, with which original data can be projected into a lower dimensional linear space with orthogonal components, namely, principal component subspace. After the PCA process, a set of correlated variables can be represented by a set of uncorrelated variables, called principal components (PCs). By carrying out PCA, a few PCs can represent most of the information in the original data. Thus, dimensionality can be reduced with almost no loss of information.

Obviously, PCA-based approach can take effect only in the case that the original data are correlated and can be transformed and represented by principal components. From our empirical study in Section 4, we have known that the weather matrix has low-rank feature (thus the original data are correlated), which makes it possible for us to design

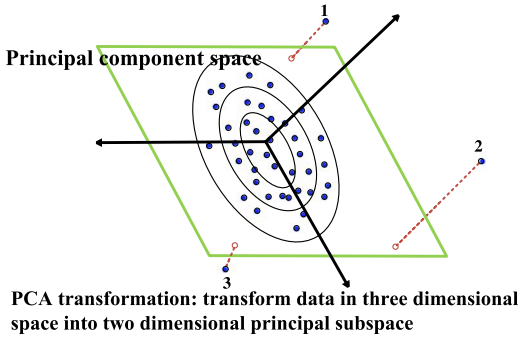


Fig. 7. Schematic representation of PCA-based structure fault detection.

row-structure/column-structure fault detection algorithms based on Principal Component Analysis (PCA) [32].

Different from the traditional fault detection algorithm, a PCA-based fault detection algorithm can effectively detect faults in the lower dimension principal component subspace instead of the original high dimension data space. In the principal component subspace, the first few PCs can contribute most to the data variance, we call these few PCs major PCs. As shown in Fig. 7, the PCA-based fault detection is designed based on the following principle: the faults are extreme data that cause a notable increase in variance and covariance in the original variables which can be detected through the examination of the major PCs only. As shown in Fig. 7, we can detect points 1, 2 and 3 to be faulty because their projected points are located far away from the points projected from normal data.

As PCA-based fault detection approaches do not need any assumption of the distribution of the original data and are more scalable to various data sets [33], [34], [35], [36], [37], we design our structure fault detection algorithm based on PCA.

As shown in Algorithm 1, suppose there is a sensory matrix $M_{N\times T}$, we design our row-structure fault detection algorithm as follows. On line 1, we first organize the sensory matrix into row-style $M_{N\times T} = [M_1, M_2, \ldots, M_N]^{tr}$ with each M_i $(1 \le i \le N)$ being an T-dimensional vector carrying the data of sensor i. On lines 2-4, an $T \times N'$ matrix $W_{T\times N'}$ is found to transform the original data matrix M to the data matrix P in the principal component subspace. On line 5, the transformed matrix is calculated through $P = M \times W$, where $P = [P_1, \ldots, P_{N'}]$ with $P_1, \ldots, P_{N'}$ being the PCs to represent most of the information in the original data. From our empirical study in Section 4, we observe that the top 20 singular values capture 70-90 percent rates of the accumulative contribution in the real data traces. The relationship between the singular value σ_i and the eigenvalue λ_i is $\lambda_i = \sigma_i^2$. Therefore, in our performance studies of this paper, we choose $N' = 20$.

In general, an observation is considered as an outlier if it is different from the majority of the data or has an unlikely value under the assumed probability model of the data. Compared with euclidean distance, Mahalanobis distance [38], [39] takes into account the correlation between observations in the distance calculation and thus is a more robust test statistic for the outlier detection. In this paper, instead of using the euclidean distance, we apply the Mahalanobis distance to determine the difference in the principal component subspace. Geometrically, P_{ik} in $P_{N\times N'}$ can be interpreted as the projection of the original data M_i $(1 \le i \le N)$ onto the principal component PC_k $(1 \le k \le N')$, and P_{ik} is called a principal component score. According to [40], the sum of the squares of the standardized principal component scores, that is, $d_i^2 = \sum_{k=1}^{N'} \frac{P_{ik}^2}{\lambda_k}$, is equivalent to the square of the Mahalanobis distance from M_i to the center of the sensory data where $1 \le i \le N$, λ_k is the kth eigenvalue of the sensory covariance matrix of M. Therefore, to detect the row structure fault, on line 6, we calculate the square of the Mahalanobis distance of data M_i.

Algorithm 1. Row-Structure Fault Detection Algorithm

Input: a sensory matrix $M_{N\times T}$

1: Organize the sensory matrix into row-style $M_{N\times T} = [M_1, M_2, \ldots, M_N]^{tr}$ with each M_i $(1 \le i \le N)$ being an T-dimensional vector carrying the data of sensor i.

2: Normalize the original sensory matrix M, denoted by M_0. Compute the covariance matrix R of M_0.

3: Find the eigenvalues and eigenvectors of matrix R, sort the eigenvalues in the descending order, $\lambda_1, \lambda_2, \ldots, \lambda_T$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_T$, and select the first N' eigenvectors by their rates of accumulative contributions, denoted by $\frac{\sum_{i=1}^{N'} \lambda_i}{\sum_{i=1}^{T} \lambda_i}$.

4: Matrix W is composed by the eigenvectors whose corresponding eigenvalues are selected, and the kth column in W is the eigenvector corresponding to the kth largest eigenvalue λ_k.

5: Calculate $P = M \times W$ where $P = [P_1, \ldots, P_{N'}]$ is the data matrix in the principal component subspace and $P_1, \ldots, P_{N'}$ are the PCs which can be applied to represent most of the information in the original data.

6: In the principal component subspace, calculate the square of the Mahalanobis distance from data M_i to the center of the sensory data, that is,

$$d_i^2 = \sum_{k=1}^{N'} \frac{P_{ik}^2}{\lambda_k}, \tag{16}$$

where $1 \le i \le N$, λ_k is the kth eigenvalue of the sensory covariance matrix of M.

7: Data M_i is abnormal if

$$d_i^2 > \chi_{N',\alpha}^2, \tag{17}$$

where $\chi_{N',\alpha}^2$ is the upper α percentage point of the chi-square distribution with the degree of freedom N'.

Since the principal components are assumed to be uncorrelated, the distribution of the Mahalanobis distance follows a chi-square distribution with N' degrees of freedom (N' is the number of principal components). Therefore, our criterion for row-structure fault detection is given on line 7: a row-structure fault is considered to happen if the square of the Mahalanobis distance of M_i is $d_i^2 > \chi_{N',\alpha}^2$ where $\chi_{N',\alpha}^2$ is the upper α percentage point of the chi-square distribution with the degree N'. In this paper, α is set as 0.975.

Above is our proposed row-structure fault detection algorithm. We can also organize the sensory matrix M into a column-style with each column being an N-dimensional vector. Our column-structure fault detection algorithm can be

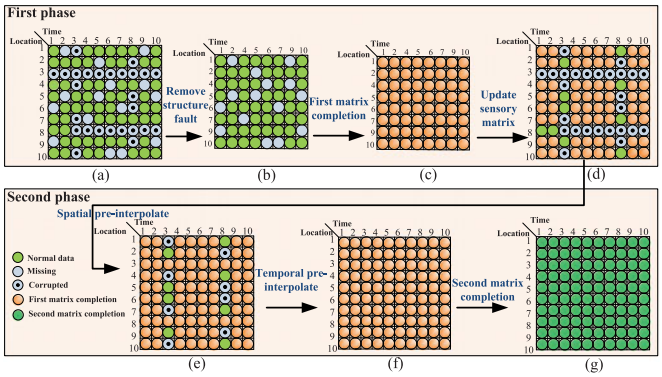


Fig. 8. Two-phase corrupted data recovery scheme.

designed in the similar way with row-structure fault detection algorithm. Due to the space limitation, we do not describe our PCA-based column-structure fault detection algorithm.

After a sensory matrix is processed to detect the row-structure faults and the column-structure faults, we will remove these faults to reduce their negative effects, and then apply the matrix completion to the remaining data in the first phase of our MC-Two-Phase scheme as introduced in Section 6.

6 Corrupted Data Recovery

To avoid the negative effects brought by structure faults, we design an innovative corrupted data recovery scheme (MC-Two-Phase) based on matrix completion, which recovers the data by taking advantage of the low-rank, temporal stability, and spatial correlation features of the data matrix. Our algorithm MC-Two-Phase has two phases with steps shown in Fig. 8 as follows:

Phase 1: *Fault data removing.* To eliminate the negative effects brought by structure faults, we first preprocess the matrix. This phase includes three steps: 1) The rows and columns detected with structure faults are removed from the sensory matrix. If the number of fault rows and columns are n_r and n_c, respectively, we denote the remaining data matrix as $M_{(N-n_r)\times(T-n_c)}$, as shown in Fig. 8b. 2) The matrix completion technique is applied to $M_{(N-n_r)\times(T-n_c)}$ to obtain the complete matrix by filling the items experiencing random missing data, denoted as $R_{(N-n_r)\times(T-n_c)}$, as shown in Fig. 8c. 3) Update $M_{N\times T}$ by replacing the parts without structure faults with $R_{(N-n_r)\times(T-n_c)}$, as shown in Fig. 8d.

Phase 2: *Recovery of data with structure faults.* The original data matrix is recovered based on the pre-processed matrix. This phase also has three steps: 1) Spatial pre-interpolation, where the row data with structure faults are replaced with the date from neighboring sensors, taking advantage of spatial correlation, as shown in Fig. 8e. 2) Temporal pre-interpolation, where the column data experiencing structure faults are replaced with the data from adjacent time slots, taking advantage of temporal stability, as shown in Fig. 8f. 3) Matrix re-processing, where matrix completion is applied to recover the original data matrix $X_{N\times T}$ from the matrix obtained following above two procedures, exploiting both spatial correlation and temporal stability, as shown in Fig. 8g.

In the next two sections, we will present in details how these two interpolation steps in phase 2 work.

6.1 Spatial Pre-Interpolation

With the existence of spatial correlation, data values captured by sensors within one-hop distance are similar. To ensure more accurate recovery of data, one key issue is to avoid using faulty data in the interpolation process. We construct a *spatial constraint matrix* H following four steps below to facilitate the spacial interpolation:

- **Step 1.** For a WSN consisting of N sensors, $H_{N\times N}$ is initialized by a $diag(d_1, d_2, \ldots, d_N)$ with the central diagonal elements of the matrix set to 1, that is, $d_1 = d_2 = \cdots = d_N = 1$.
- **Step 2.** Replace the row-structure fault data with their neighboring data. For a row i which experiences the structure fault, in the data matrix ($1 \le i \le N$), set $H_{(i)} = Y_{(i)}$, where $H_{(i)}$ and $Y_{(i)}$ are the ith row of matrix H and matrix Y, respectively. Matrix Y is the one-hop topology matrix introduced in Section 4.3.
- **Step 3.** To avoid utilizing a faulty value to interpolate other faulty data when there are multiple row-structure faults, for each fault at ith row of the matrix M ($1 \le i \le N$), set the ith column of the spatial constraint matrix to 0, that is $H^{(i)} = 0$ (where $H^{(i)}$ denotes the ith column of matrix H).
- **Step 4.** Normalize the rows of H so that all elements of a row add up to 1.

After obtaining the spatial constraint matrix, for a sensor i with fault, we can apply the average data value of its one-hop neighbors to pre-interpolate the corresponding faulty row data through $H \times M$, where M is the data matrix obtained in the first phase of our MC-Two-Phase scheme.

Example 1. For a WSN consisting of five sensor nodes, the topology matrix Y and the sensory matrix M are expressed as

$$Y = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{18}$$

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} \\ M_{31} & M_{32} & M_{33} & M_{34} & M_{35} \\ M_{41} & M_{42} & M_{43} & M_{44} & M_{45} \\ M_{51} & M_{52} & M_{53} & M_{54} & M_{55} \end{bmatrix}. \tag{19}$$

Assume that sensors 2 and 4 are running out of the energy and the row-structure faults are detected at rows 2 and 4 in the sensory matrix. From the topology matrix, we know sensors 1, 3, and 4 are the one-hop neighbors of sensor 2, while sensors 1, 2, and 5 are the one-hop neighbors of sensor 4. According to the steps introduced above, the spatial constraint matrix H is built as follows:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{20}$$

where the rows 2 and 4 of the matrix H is replaced with the corresponding rows of Y.

Sensors 2 and 4 are one-hop neighbors. To avoid utilizing error data at row 2 and row 4 in matrix M to interpolate other data, we let columns $H^{(2)}$ and $H^{(4)}$ to be 0, and we obtain

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{21}$$

Then we normalize the rows of H

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{22}$$

Finally, through $H \times M$, we can obtain the spatial pre-interpolated matrix

$$H \times M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ \frac{M_{11}+M_{31}}{2} & \frac{M_{12}+M_{32}}{2} & \frac{M_{13}+M_{33}}{2} & \frac{M_{14}+M_{34}}{2} & \frac{M_{15}+M_{35}}{2} \\ M_{31} & M_{32} & M_{33} & M_{34} & M_{35} \\ \frac{M_{11}+M_{51}}{2} & \frac{M_{12}+M_{52}}{2} & \frac{M_{13}+M_{53}}{2} & \frac{M_{14}+M_{54}}{2} & \frac{M_{15}+M_{55}}{2} \\ M_{51} & M_{52} & M_{53} & M_{54} & M_{55} \end{bmatrix}. \tag{23}$$

6.2 Temporal Pre-Interpolation

Similar to spatial pre-interpolation in Section 6.1, we construct a temporal constraint matrix G following four steps below to capture the temporal stability feature:

- **Step 1.** Initialize the matrix $G_{T\times T}$ with its diagonal elements set to 1.
- **Step 2.** Replace the column-structure fault data utilizing those from neighboring time slots. For each column-structure fault at time slot i ($1 \le i \le T$), we first set the i-th column of G to 0, i.e., $G^{(i)} = 0$, and then set entries $G_{j,i} = w_{ji}$ where a slot j is the one that has strong relationship with slot i and w_{ji} is the weight that reflects how strong the relationship is.
- **Step 3.** To avoid interpolating data with other faulty data when there are multiple column-structure faults, for each fault at ith column in matrix M ($1 \le i \le T$), set the ith row in temporal constraint matrix to 0, that is $G_{(i)} = 0$.
- **Step 4.** Normalize columns of G so that all elements of a column add up to 1.

After obtaining the temporal constraint matrix G, we can apply $M \times G$ to obtain the temporal pre-interpolated matrix, where M is the data matrix with corrupted rows spatially interpolated.

Exponentially Weighted Moving Average (EWMV) is considered to be effective in estimating data in time series model [41], [42]. We design our temporal pre-interpolation algorithm based on EWMA. Taking EWMA as an example, we show a temporal pre-interpolation algorithm utilizing the adjacent four time slots (i.e., two slots before and two slots after the fault) to replace the data corrupted by the column-structure fault.

Example 2. for a WSN consisting of 4 sensor nodes, its sensory matrix M with 5 time slots is

$$M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} \\ M_{31} & M_{32} & M_{33} & M_{34} & M_{35} \\ M_{41} & M_{42} & M_{43} & M_{44} & M_{45} \end{bmatrix}. \tag{24}$$

Assuming that the time slot 3 is detected to have column-structure fault, following steps 1-3 above, we first set the 3rd column of G to 0, that is, $G^{(3)} = 0$, and then set entries $G_{2,3} = \alpha$, $G_{1,3} = 1-\alpha$, $G_{4,3} = \alpha$, $G_{5,3} = 1-\alpha$, where $0 \le \alpha \le 1$. The temporal constraint matrix G is built as

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow G = \begin{bmatrix} 1 & 0 & (1-\alpha) & 0 & 0 \\ 0 & 1 & \alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha & 1 & 0 \\ 0 & 0 & (1-\alpha) & 0 & 1 \end{bmatrix}. \tag{25}$$

Normalize columns of G, we have

$$G = \begin{bmatrix} 1 & 0 & (1-\alpha)/2 & 0 & 0 \\ 0 & 1 & \alpha/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \alpha/2 & 1 & 0 \\ 0 & 0 & (1-\alpha)/2 & 0 & 1 \end{bmatrix}. \tag{26}$$

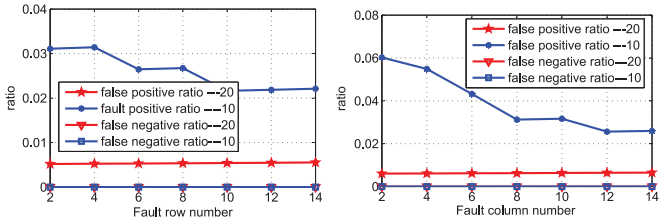
Finally, with $M \times G$, we can obtain the temporally pre-interpolated matrix with the 3rd column $M^{(3)}$ as

$$M^{(3)} = \alpha\left(\frac{M^{(4)} + M^{(2)}}{2}\right) + (1-\alpha)\left(\frac{M^{(5)} + M^{(1)}}{2}\right), 0 \le \alpha \le 1. \tag{27}$$

Obviously, $M^{(3)}$ in (27) is in the EWMA style, where $0 \le \alpha \le 1$ is the smoothing constant (also referred to as the discount factor). $M^{(1)}$, $M^{(2)}$, $M^{(4)}$, and $M^{(5)}$ are the sensory data in the time slot 1, 2, 4, and 5, respectively. In this paper, we set $\alpha = 0.8$.

Our temporal pre-interpolation algorithm can be easily extended to support other types of interpolation schemes and with different time slots and weight setting.

As shown in Fig. 8, our MC-Two-Phase consists of two phases which includes several main techniques: PCA-based structure fault detection, matrix recovery algorithm, spatial pre-interpolation, and temporal pre-interpolation. For the PCA-based structure fault detection, PCA transformation is the main step for fault detection with its complexity generally being $O(p^2n + p^3)$ for a data set of size n with p features. Our MC-Two-Phase does not depend on the underlying matrix recovery algorithms, while different matrix recovery algorithms have different computation complexity. In practice, we will choose approximate matrix recovery algorithms according to the application requirements. In MC-Two-Phase, the proposed spatial pre-interpolation and temporal pre-interpolation are based on simple matrix



(a) Row-structure fault detection (b) Column-structure fault detection

Fig. 9. Data matrix includes only row or column structure faults.

multiplication with neighbor information, thus the complexity is very low.

7 Performance Evaluations

To evaluate the performance of our MC-Two-Phase scheme, we have performed extensive simulations driven by real weather traces collected by our deployed 196 sensors.

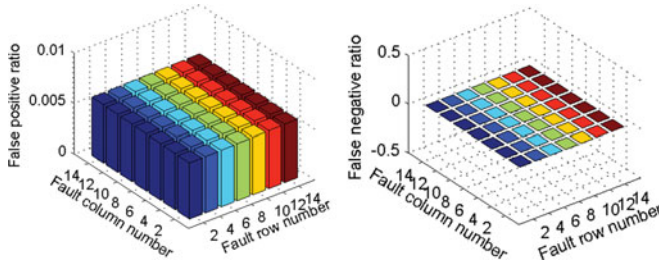
Three series of simulations are conducted. We evaluate the performance of our PCA-based structure fault detection algorithm in the first simulation, and then evaluate our MC-Two-Phase scheme on handling the structure corruption in the second simulation. The corrupted data set utilized in the first two series of simulations is generated from the raw trace data following the steps described in Section 4.4. In the third simulation, to evaluate the performance of our MC-Two-Phase scheme on handling the missing of whole rows or columns, we generate the data set from the raw trace data by randomly letting some rows and columns be empty.

7.1 Evaluation of PCA-Based Structure Fault Detection

We use two performance metrics to evaluate the PCA-based Structure Fault Detection algorithm: false positive ratio and false negative ratio. False positive ratio is the proportion of normal rows/columns that are erroneously reported as being corrupted. False negative ratio is the proportion of corrupted rows/columns that are erroneously reported as normal. Following the procedure of Section 4.4, we generate corrupted data matrix in two different ways: 1) Only row or column structure faults are generated in the data matrix; and 2) The corrupted data matrix includes both row and column structure faults. To investigate how the noise level impacts the performance of structure fault detection algorithms, two different noise levels $\delta^2 = 20$ and $\delta^2 = 10$ are simulated.

Figs. 9a and 9b show the performance of the proposed row-structure fault detection algorithm and column-structure fault detection algorithm when the data matrix includes only row or column structure faults, respectively. We can see that the false negative ratios of our algorithms are zero under both noise levels. It is easily observed that the larger the noise level is, the smaller the false positive ratio thus the larger detection accuracy. Even $\delta^2 = 10$, the false positive ratio can be controlled to be very low (i.e., less than 0.03 for row-structure fault detection, and 0.06 for column-structure fault detection).

Figs. 10 and 11 show the false positive ratio and false negative ratio when both row and column structure faults exist in the data matrix. Our PCA-based fault detection approach can achieve false positive ratio less than 0.04 ($\delta^2 = 10$), 0.006 ($\delta^2 = 20$) in all cases and zero false negative ratio.



(a) Fault positive ratio $\delta^2 = 20$ (b) False negative ratio $\delta^2 = 20$

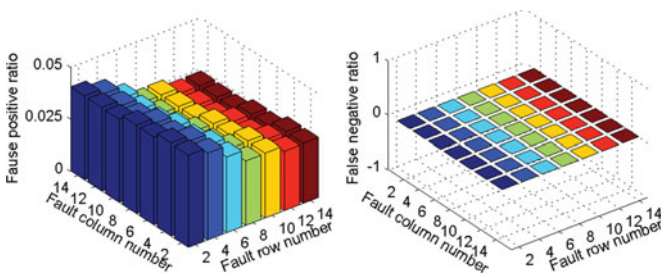
Fig. 10. Data matrix includes both row and column structure faults with the noise level $\delta^2 = 20$.

Therefore, our PCA-based fault detection approach can detect various structure faults, which may be caused by failures of senor nodes or data tampering from attackers. In following simulations, we set $\delta^2 = 20$.

7.2 Performance Comparison

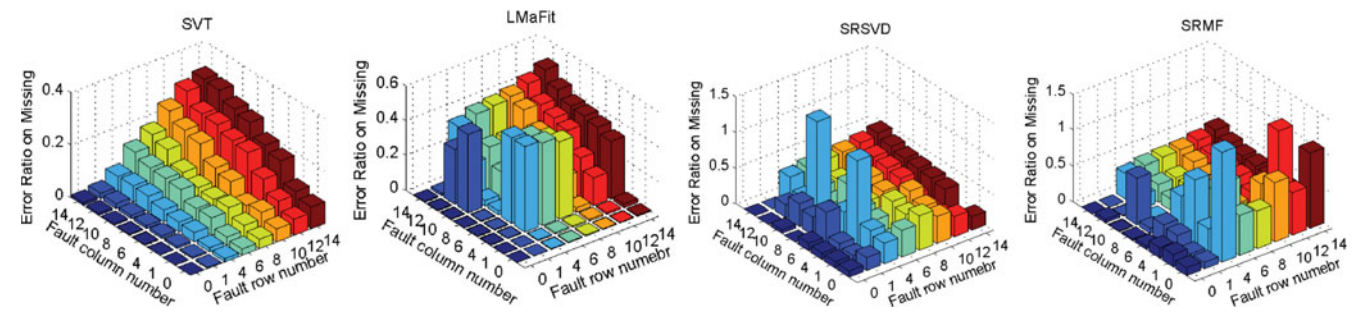
As our MC-Two-Phase is designed to not depend on the underlying matrix recovery approach, to evaluate the performance of the proposed MC-Two-Phase, four different matrix recovery approaches are implemented under our MC-Two-Phase scheme.

- *SVT* [31]. SVT approximates the matrix with the minimum nuclear norm obeying a set of convex constraints. SVT has two remarkable features: applying the soft-thresholding operation is to a sparse matrix, and the rank of the matrix obtained in iterations is empirically non decreasing.
- *SRSVD* [23]. SRSVD derives two decomposed matrix L and R using an alternating least-square procedure. It solves the interpolation problem by fixing one of the decomposed matrices, L or R, and taking the other as the optimization variable. Then the roles of the two matrices are swapped to continue alternating towards a solution till the convergence. The recovered matrix can be calculated from LR^T at last.
- *SRMF* [24]. Similar to SRSVD, SRMF solves the interpolation problem using alternating least squares. Different from SRSVD, SRMF exploits spatio-temporal properties in the matrix decomposition.
- *LMaFit* [25]. LMaFit is an alternating minimization scheme which can be started from a rough overestimate of the true matrix rank for completion, and updates each of the three variables X, Y or Z (Z is



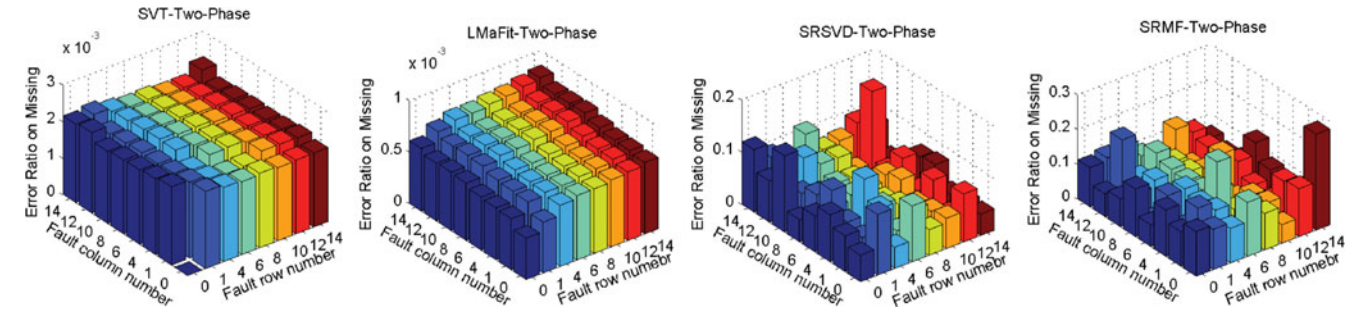
(a) Fault positive ratio $\delta^2 = 10$ (b) False negative ratio $\delta^2 = 10$

Fig. 11. Data matrix includes both row and column structure faults with the noise level $\delta^2 = 10$.



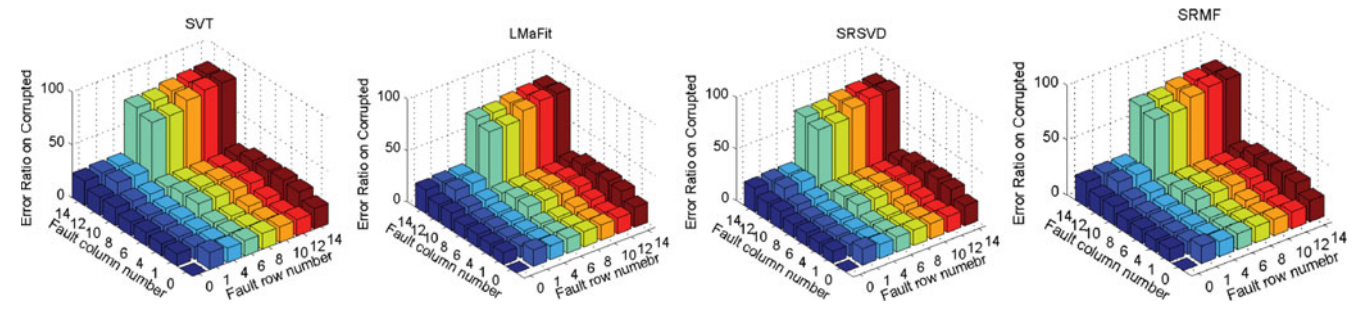
(a) Error Ratio Under SVT (b) Error Ratio Under LMaFit (c) Error Ratio Under SRSVD (d) Error Ratio Under SRMF

Fig. 12. The recovery performance on missing data when structure corrupted faults exist under conventional method.



(a) Error Ratio Under SVT-Two-Phase (b) Error Ratio Under LMaFit-Two-Phase (c) Error Ratio Under SRSVD-Two-Phase (d) Error Ratio Under SRMF-Two-Phase

Fig. 13. The recovery performance on missing data when structure corrupted faults exist under Two-Phase scheme.



(a) Error Ratio Under SVT (b) Error Ratio Under LMaFit (c) Error Ratio Under SRSVD (d) Error Ratio Under SRMF

Fig. 14. The recovery performance on corrupted data when structure corrupted faults exist under conventional method.

the estimated matrix and X, Y are decomposed matrices) efficiently while fixing the other two.

The implemented MC-Two-Phase schemes under different matrix recovery approaches are denoted as SVT-Two-Phase, SRSVD-Two-Phase, SRMF-Two-Phase, and LMaFit-Two-Phase, respectively. Moreover, for performance comparison, the above four matrix recovery approaches are also directly implemented on the corrupted synthesized data to recover the sensory matrix. In our performance studies, the directly implemented schemes are denoted as SVT, SRSVD, SRMF, and LMaFit, respectively.

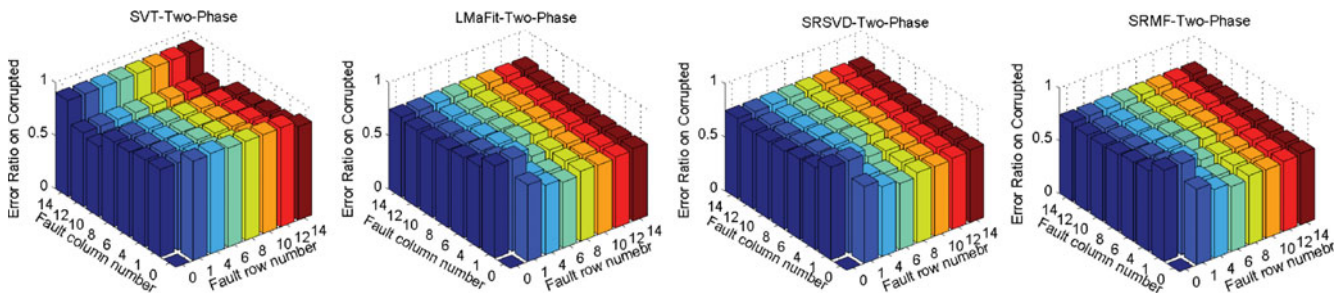
7.2.1 Structure Corrupted Fault

Although the performance under the conventional SVT is shown in Fig. 6, for the convenience of comparison, we still draw the performance results of conventional SVT in Figs. 12a, 14a, and 16a.

In Figs. 12, 13, 14, 15, 16, and 17, our MC-Two-Phase scheme is shown to be able to control the error ratios on the missing data, the corrupted data, and the normal data at a very low level, while the error ratios under conventional SVT, LMaFit, SRSVD, SRMF are much higher. Taking SVT as an example, although SVT-Two-Phase and SVT adopt the same singular value thresholding approach to recover the matrix, the error ratios on the missing data, the corrupted data, and the normal data under SVT-Two-Phase are only 1, 2, and 1 percent of those under the conventional SVT.

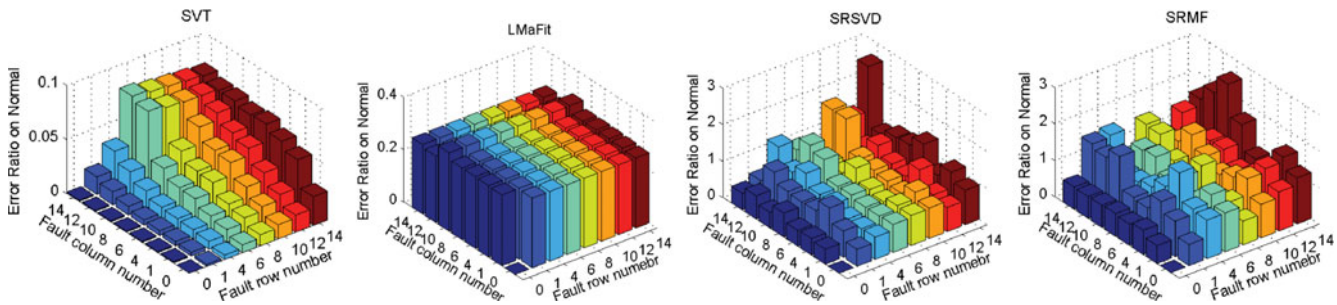
It is worth noticing that, in Figs. 14a, 14b, 14c, and 14d, the recovery error ratios for corrupted rows and columns are high under the conventional SVT, LMaFit, SRSVD and SRMF. Even when the data matrix has only one corrupted row or column, the error ratio for the corrupted entries is a big value larger than 20. In contrast, the error ratios under SVT-Two-Phase, LMaFit-Two-Phase, SRSVD-Two-Phase, and SRMF-Two-Phase are less than 0.8 in all scenarios studied (Fig. 15).

When there exist corrupted rows or columns, the recovery error ratios on normal data and missing data are also higher under conventional SVT, LMaFit, SRSVD and SRMF,



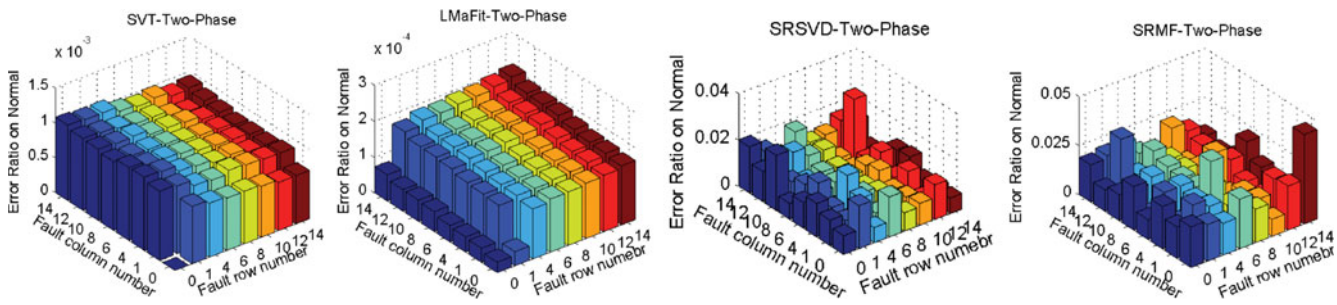
(a) Error Ratio Under SVT-Two-Phase (b) Error Ratio Under LMaFit-Two-Phase (c) Error Ratio Under SRSVD-Two-Phase (d) Error Ratio Under SRMF-Two-Phase

Fig. 15. The recovery performance on corrupted data when structure corrupted faults exist under Two-Phase scheme.



(a) Error Ratio Under SVT (b) Error Ratio Under LMaFit (c) Error Ratio Under SRSVD (d) Error Ratio Under SRMF

Fig. 16. The recovery performance on normal data when structure corrupted faults exist under conventional method.



(a) Error Ratio Under SVT-Two-Phase (b) Error Ratio Under LMaFit-Two-Phase (c) Error Ratio Under SRSVD-Two-Phase (d) Error Ratio Under SRMF-Two-Phase

Fig. 17. The recovery performance on normal data when structure corrupted faults exist under Two-Phase scheme.

as shown in Figs. 16 and 12. These results also demonstrate that, regardless of the underlying matrix recovery approaches, successive data corruption seriously impacts the accuracy of missing data recovery and even pollutes the normal data.

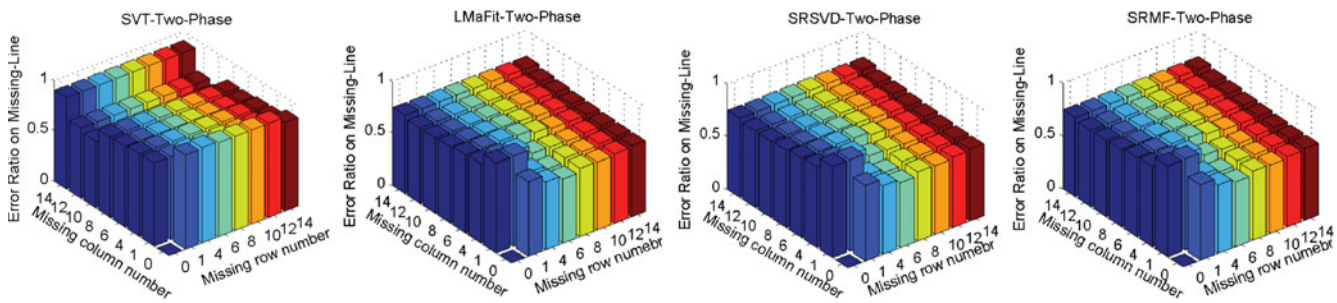
From Figs. 16c and 16d, we can see that, the error ratios on normal data under conventional SRSVD and SRMF can even reach three due to the existence of corrupted entries. In contrast, as shown in Figs. 17c and 17d, the error ratios under SRSVD-Two-Phase and SRMF-Two-Phase are much smaller and in the range of [0.01, 0.03] and [0.01, 0.04] respectively. Though the error ratio on normal data under conventional SVT (Fig. 16a) is under 0.08, it is still higher than that of SVT-Two-Phase (Fig. 17a).

All these simulation results demonstrate that our MC-Two-Phase scheme is very effective in handling structure faults and recovering the corrupted matrix data. Moreover, the four MC-Two-Phase implementation with four different matrix recovery techniques also demonstrate that our MC-Two-Phase is a general matrix completion scheme and dose not depend on the underlying matrix recovery approaches.

7.2.2 *Row and Column Missing*

Fig. 18 shows the recovery performance when some rows and columns are missing in the data matrix through our MC-Two-Phase scheme. As shown in Fig. 18, MC-Two-Phase schemes achieve the very low error ratios (which are within the range of [0.5, 0.8]) under all matrix recovery approaches.

From the literature work, we know that conventional matrix completion approaches can only recover data if there is no row or column to be completely empty. If a row or a column is missing, these schemes do not have effect on these missing entries. Different from conventional matrix completion approaches, in our MC-Two-Phase scheme, we utilize our proposed spatial pre-interpolation, temporal pre-interpolation algorithms to fill in the empty rows and columns first, and then apply the matrix completion to smooth the data. Therefore, our MC-Two-Phase scheme can



(a) Error Ratio Under SVT-Two-Phase (b) Error Ratio Under LMaFit-Two-Phase (c) Error Ratio Under SRSVD-Two-Phase (d) Error Ratio Under SRMF-Two-Phase

Fig. 18. The recovery performance when whole rows and columns are missing under two-phase scheme.

break the restriction in the conventional matrix completion techniques to correctly recover the whole row and column of missing data.

The low-rank feature is the prerequisite for matrix completion and thus our proposed techniques in this paper. As almost all physical conditions monitored are continuous without sudden changes, sensory data generally exhibit strong spatio-temporal correlation [43]. Thus the sensory data matrix has a low-rank feature. Some previous studies [22], [44], [45], [46] also show that sensory matrices of temperature, humidity, light, and PM2.5 are low-rank and have high spatiotemporal correlations. Although this paper utilizes weather data collected from 196 sensor notes as a case to verify the effectiveness of the proposed MC-Two-Phase scheme, we expect that our scheme can also work well to recover other sensory matrices. In our future work, we will evaluate the performance of our MC-Two-Phase on other sensory matrices.

8 Conclusion

This paper proposes a two-phase matrix completion scheme to recover successively missing or corrupted data, named MC-Two-Phase. The scheme applies matrix completion to fully exploit the inherent features of environmental data to perform data recovery. MC-Two-Phase scheme includes three algorithms: structure fault detection based on Principal Component Analysis (PCA), spatial pre-interpolation, and temporal pre-interpolation. Most importantly, our scheme exploits matrix completion to fully integrate results from the three algorithms for more efficient and reliable data recovery.

We have performed extensive simulations with real-world sensory data. The simulation results demonstrate that our MC-Two-Phase can achieve very good recovery performance when successive data corruption exists. Specifically, the error ratios on the missing data, the corrupted data, and the normal data under our SVT-Two-Phase are only 1, 2, and 1 percent of those under the conventional SVT. Moreover, our MC-Two-Phase scheme can break the restriction in conventional matrix completion techniques to correctly recover the data matrix even when some rows or columns are completely empty.

Acknowledgments

The work is supported by the National High Technology Research and Development Program of China (863 Program) under Grant No. 2015AA01A705, the Prospective Research Project on Future Networks (Jiangsu Future Networks Innovation Institute) under Grant No. BY2013095-4-06, the National Natural Science Foundation of China under Grant Nos. 61572184, 61300219, 61472283, 61271185, and 61472131, US National Science Foundation CNS 1526843.

References

[1] J. Shen, H.-W. Tan, J. Wang, J.-W. Wang, and S.-Y. Lee, "A novel routing protocol providing good transmission reliability in underwater sensor networks," *J. Internet Technol.*, vol. 16, no. 1, pp. 171–178, 2015.

[2] L. Mo, et al., "Canopy closure estimates with GreenOrbs: Sustainable sensing in the forest," in *Proc. 7th ACM Conf. Embedded Networked Sensor Syst.*, 2009, pp. 99–112.

[3] K. Xie, et al., "Decentralized context sharing in vehicular delay tolerant networks with compressive sensing," in *Proc. 36th IEEE Int. Conf. Distrib. Comput. Syst.*, 2016, pp. 169–178.

[4] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh, "Fidelity and yield in a volcano monitoring sensor network," in *Proc. 7th Symp. Operating Syst. Design Implementation*, 2006, pp. 381–396.

[5] F. Martincic and L. Schwiebert, "Distributed event detection in sensor networks," in *Proc. Int. Conf. Syst. Netw. Commun.*, 2006, Art. no. 43.

[6] K. Xie, X. Wang, X. Liu, J. Wen, and J. Cao, "Interference-aware cooperative communication in multi-radio multi-channel wireless networks," *IEEE Trans. Comput.*, vol. 65, no. 5, pp. 1528–1542, May 2016.

[7] K. Xie, X. Wang, J. Wen, and J. Cao, "Cooperative routing with relay assignment in multiradio multihop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 24, no. 2, pp. 859–872, Apr. 2016.

[8] K. Xie, J. Cao, X. Wang, and J. Wen, "Optimal resource allocation for reliable and energy efficient cooperative communications," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 4994–5007, Oct. 2013.

[9] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos, "Online outlier detection in sensor data using non-parametric models," in *Proc. 32nd Int. Conf. Very Large Data Bases*, 2006, pp. 187–198.

[10] X. Du and H.-H. Chen, "Security in wireless sensor networks," *IEEE Wireless Commun.*, vol. 15, no. 4, pp. 60–66, Aug. 2008.

[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. TIT-13, no. 1, pp. 21–27, Jan. 1967.

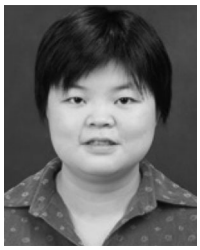
[12] L. Kong, D. Jiang, and M.-Y. Wu, "Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction," in *Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst.*, 2010, pp. 179–188.

[13] H. Zhu, Y. Zhu, M. Li, and L. M. Ni, "SEER: Metropolitan-scale traffic perception based on lossy sensory data," in *Proc. IEEE INFOCOM*, 2009, pp. 217–225.

[14] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations Comput. Math.*, vol. 9, no. 6, pp. 717–772, 2009.

[15] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learning Res.*, vol. 99, pp. 2057–2078, 2010.

[16] A. Eriksson and A. Van Den Hengel, "Efficient computation of robust low-rank matrix approximations in the presence of missing data using the l_1 norm," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2010, pp. 771–778.
[17] R. Ma, N. Barzigar, A. Roozgard, and S. Cheng, "Decomposition approach for low-rank matrix completion and its applications," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1671–1683, Apr. 2014.
[18] M. Yan, Y. Yang, and S. Osher, "Exact low-rank matrix completion from sparsely corrupted entries via adaptive outlier pursuit," *J. Scientific Comput.*, vol. 56, no. 3, pp. 433–449, 2013.
[19] Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi, "Matrix completion with column manipulation: Near-optimal sample-robustness-rank tradeoffs," *IEEE Trans. Inform. Theory*, vol. 62, no. 1, pp. 503–526, Jan. 2016.
[20] K. Xie, et al., "Sequential and adaptive sampling for matrix completion in network monitoring systems," in *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 2443–2451.
[21] K. Xie, L. Wang, X. Wang, J. Wen, and G. Xie, "Learning from the past: Intelligent on-line weather monitoring based on matrix completion," in *Proc. IEEE 34th Int. Conf. Distrib. Comput. Syst.*, 2014, pp. 176–185.
[22] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *Proc. IEEE INFOCOM*, 2013, pp. 1654–1662.
[23] W. W. Yin Zhang and M. Roughan, "Spatio-temporal compressive sensing and internet traffic matrices," in *Proc. ACM SIGCOMM Conf. Data Commun.*, 2009, pp. 267–278.
[24] M. Roughan, Z. Yin, W. Willinger, and Q. Lili, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 662–676, Jun. 2012.
[25] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Math. Programming Comput.*, vol. 4, no. 4, pp. 333–361, 2012.
[26] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.
[27] M. Fornasier, H. Rauhut, and R. Ward, "Low-rank matrix recovery via iteratively reweighted least squares minimization," *SIAM J. Optimization*, vol. 21, no. 4, pp. 1614–1640, 2011.
[28] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.
[29] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Foundations Comput. Math.*, vol. 11, no. 2, pp. 183–210, 2011.
[30] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Math. Programming*, vol. 128, no. 1/2, pp. 321–353, 2011.
[31] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
[32] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, "ROBPCA: A new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.
[33] K. S. Daniela Brauckhoff, "Applying PCA for traffic anomaly detection: Problems and solutions," in *Proc. IEEE INFOCOM*, 2009, pp. 2866–2870.
[34] S. C. Meiling Shyu, "A novel anomaly detection scheme based on principal component classifier," in *Proc. IEEE Foundations New Directions Data Mining Workshop, Conjunction 3rd IEEE Int. Conf. Data Mining*, 2003, pp. 171–179.
[35] K. Xu, "A first step toward understanding inter-domain routing dynamics," in*ACM SIGCOMM Workshop Mining Netw. Data*, 2005, pp. 207–212.
[36] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proc. 5th ACM SIGCOMM Conf. Internet Measurement*, 2005, pp. 30–30.
[37] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," in *Proc. ACM SIGMETRICS Int. Conf. Measurement Modeling Comput. Syst.*, 2007, pp. 109–120.
[38] H. Braun, M. Banavar, and A. Spanias, *Signal Processing for Solar Array Monitoring, Fault Detection, and Optimization*. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.
[39] G. Shmueli, N. R. Patel, and P. C. Bruce, *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. New York, NY, USA: Wiley, 2011.
[40] J. Jobson, *Applied Multivariate Data Analysis: Volume II: Categorical and Multivariate Methods*. Berlin, Germany: Springer, 2012.
[41] L. C. Alwan and H. V. Roberts, "Time-series modeling for statistical process control," *J. Business Economic Statistics*, vol. 6, no. 1, pp. 87–95, 1988.
[42] D. C. Montgomery, "The use of statistical process control and design of experiments in product and process improvement," *IIE Trans.*, vol. 24, no. 5, pp. 4–17, 1992.
[43] M. C. Vuran, Ö. B. Akan, and I. F. Akyildiz, "Spatio-temporal correlation: Theory and applications for wireless sensor networks," *Comput. Netw.*, vol. 45, no. 3, pp. 245–259, 2004.
[44] G. Chen, et al., "Multiple attributes-based data recovery in wireless sensor networks," in *Proc. IEEE Global Commun. Conf.*, 2013, pp. 103–108.
[45] J. Cheng, Q. Ye, H. Jiang, D. Wang, and C. Wang, "STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp. 850–861, Feb. 2013.
[46] Q. Yuan, Z. Liu, J. Li, S. Yang, and F. Yang, "An adaptive and compressive data gathering scheme in vehicular sensor networks," in *Proc. IEEE 21st Int. Conf. Parallel Distrib. Syst.*, 2015, pp. 207–215.



Kun Xie received the PhD degree in computer application from Hunan University, Changsha, China, in 2007. She worked as a postdoctoral fellow in the Department of Computing, Hong Kong Polytechnic University from Dec. 2007 to Feb. 2010. She worked as a visiting researcher in the Department of Electrical and Computer Engineering with the State University of New York, Stony Brook, from Sep. 2012 to Sep. 2013. She is currently a professor at Hunan University, Changsha, China. Her research interests include wireless network and mobile computing, network management and control, cloud computing and mobile cloud, and big data.

Xueping Ning received the MS degree in computer application from Hunan University, China, in 2016. Her research interests include compressive sensing and matrix completion.

Xin Wang (M'10) received the PhD degree in electrical and computer engineering from Columbia University, New York, NY. She is currently an associate professor in the Department of Electrical and Computer Engineering, State University of New York at Stony Brook, Stony Brook, NY. Before joining Stony Brook, she was a member of technical staff in the area of mobile and wireless networking with Bell Labs Research, Lucent Technologies, New Jersey, and an assistant professor in the Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY. Her research interests include algorithm and protocol design in wireless networks and communications, mobile and distributed computing, and as well as networked sensing and detection. She has served on the executive committees and technical committees of numerous conferences and funding review panels, and served as the associate editor of the *IEEE Transactions on Mobile Computing*. She achieved the NSF career award in 2005, and ONR challenge award in 2010. She is a member of the IEEE.

Dongliang Xie received the PhD degree from the Beijing Institute of Technology, China, in 2002. He is currently a visiting researcher in the Department of Electrical and Computer Engineering, State University of New York at Stony Brook. He is an associate professor in State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications (BUPT), China. His research interests focus on resource-constrained wireless communication and information-centric network, including architecture of ubiquitous and heterogeneous network, complex network analysis, and as well as content retrieval and service management.

Jiannong Cao (M'93-SM'05-F'14) received the PhD degree in computer science from Washington State University, Pullman, WA, USA, in 1990. He is currently a chair professor and head of the Department of Computing, Hong Kong Polytechnic University, Hung Hom, Hong Kong. His research interests include parallel and distributed computing, computer networks, mobile and pervasive computing, fault tolerance, and middleware. He has served as an associate editor and a member of the editorial boards of many international journals, including the *IEEE Transactions on Parallel and Distributed Systems*, the *IEEE Network,* Elsevier's *Pervasive and Mobile Computing Journal*, the *Springer Peer-to-Peer Networking and Applications*, and the Wiley's *Wireless Communications and Mobile Computing*. He has also served as a chair and member of organizing and technical committees of main international conferences, including PERCOM, INFOCOM, ICDCS, DSN, SRDS, ICNP, and RTSS. He is fellow of the IEEE.

Gaogang Xie received the BS degree in physics and the MS and PhD degrees in computer science from Hunan University, in 1996, 1999, and 2002, respectively. He is currently a professor and director of Network Technology Research Center with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include Internet architecture, packet processing and forwarding, and Internet measurement.

Jigang Wen received the PhD degree in computer application from Hunan University, China, in 2011. He worked as a research assistant in the Department of Computing, Hong Kong Polytechnic University, from 2008 to 2010. He is now a postdoctoral fellow in the Institute of Computing Technology, Chinese Academy of Sciences, China. His research interests include wireless network and mobile computing and high speed network measurement and management.