Spatio-Temporal Tensor Completion for Estimating Missing Internet Traffic Data

Huibin Zhou, Dafang Zhang, Kun Xie, Yuxiang Chen
College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, P. R. China
E-mail:{zhouhb317, dfzhang, xiekun, s1210w116}@hnu.edu.cn

***Abstract*—Network traffic data consists of Traffic Matrix (TM), which represents the volumes of traffic between Origin and Destination (OD) pairs in the network. It is a key input parameter of network engineering tasks. However, direct measurement of the OD pairs traffic is usually not feasible. Even good traffic measurement systems can suffer from errors, missing data. So obtaining the ODs traffic precisely is a challenge. Existing completion methods often perform poorly for network traffic estimation. Their recovery accuracy tends to be significantly worse when the data loss rate is high. Taking into account network traffic lower-dimensional latent structure and traffic hidden characteristic, a tensor (multi-way array) is introduced to model a time series of pure spatial traffic matrices in this paper. To recover the missing entries in tensors of traffic data, a novel spatio-temporal tensor completion method has been proposed. This approach not only takes advantage of tensor decomposition and its lower-dimensional representation, but also well takes into account traffic spatio-temporal properties. The extensive experiments with the real-world traffic trace data show that the proposed method can significantly reduce the missing traffic data recovery errors and achieve satisfactory completion accuracy comparing with the state-of-the-art completion methods.**

I. Introduction

The traffic data of network is essential to carry out better network management. These data consists of Traffic Matrix (TM), which represents the volumes of traffic between Origin and Destination (OD) pairs in the network [1]. As an overview of the whole network, it is a key input parameter of many networks engineering tasks, such as traffic engineering, capacity planning and anomaly detection [2]. Unfortunately the complete measurement of the OD traffic is usually difficult or even impossible (but expensive). Traffic data collection systems are affected by hardware and network transport protocol. Unreliable links and transport protocol (i.e., UDP) cause traffic data structural loss in the collection process [3]. How to cope with missing data that frequently arise in TMs is still a main challenge. Since many network engineering tasks are sensitive to missing values, it is important to accurately recover missing values from the partial direct measured OD pairs traffic data.

To infer the missing data, some research works have been developed. Non-negative Matrix Factorization (NMF) [4] use matrix decomposition technique to recover the missing entries in a matrix. Sparsity Regularized SVD (SRSVD) [5] utilizes matrix singular value decomposition to estimate the missing traffic data. Compressive Sensing (CS) [6] takes advantage of the sparsity of data to infer the missing values, such as spatio-temporal compressive sensing framework for traffic interpolation [5] and power laws and compressive sensing reconstruction approach to network traffic [7]. Following CS, matrix completion (MC) [8] exploits the low-rank structure of matrix to recover the missing entries, such as the Singular Value Thresholding algorithm (SVT) [9] and Low-rank Matrix Fitting algorithm (LMaFit) [10], etc.

Besides above research work, there are other studies which formulate data being processed as a form of tensor to estimate the missing values. A tensor is a multidimensional or N-way array, which preserve the multi-way nature of the data and extract the underlying factors in each dimension of tensor. J. Liu et al. [11] propose a high accuracy low rank tensor completion algorithm (HaLRTC) to estimate missing values in tensors of visual data. E. Acar et al. [12] develop an algorithm called CP-WOPT (CP Weighted OPTimization) to recover missing entries of a tensor. H. Tan et al. [13] propose a tensor decomposition based imputation method (TDI) to estimate the missing value in transportation traffic.

Despite much recent progress in these area, our extensive evaluation of the existing completion algorithms on real-world network traffic trace data shows that they do not perform well for the missing data estimation. Specifically, their recovery accuracy are still low when large amounts of data is missing. To overcome the shortcomings of these methods, we exploit traffic lower-dimensional latent space and traffic hidden characteristic to improve the quality of the missing data recovery.

In this paper, we model network traffic data as tensor pattern. Inspired by spatio-temporal compressive sensing [5], we propose a novel spatio-temporal tensor completion method to recover the missing entries in tensors of traffic data. We utilize spatio-temporal properties information to regularize the tensor decomposition procedure, resulting in a unified framework for traffic tensor completion. The main contributions of this paper include:

- A tensor is introduced to model a time series of pure spatial traffic matrices, which preserve the multi-way nature of the network traffic data and extract the latent structure of traffic via tensor factorization.
- By taking advantage of tensor decomposition, which project instances into a lower-dimensional latent space, and spatio-temporal information within-mode regularization, we propose a novel spatio-temporal tensor completion method to estimate the missing traffic data.
- Through extensive experiments with real-world traffic trace data, the evaluations show that our method can accurately recover the missing values with very low

978-1-4673-8590-9/15/$31.00 ©2015 IEEE

estimation error. Even when 95% of the data is missing, the proposed approach can still reconstruct the tensor with about 20% errors.

The remaining of this paper is structured as follows. Section II outlines the notation used in this paper. Section III introduces the related work. We formulate the problem in Section IV. Section V presents our proposed approach. Numerical results are given in Section VI. Conclusion and future work are discussed in Section VII.

II. NOTATIONS AND TENSOR BASICS

In this section, we partially adopt the notations of Kolda and Bader's review on tensor [14]. A tensor is a multidimensional array, whose essence is a mapping from a linear space to another. The order of a tensor is the number of dimensions, also known as ways or modes. Tensors of order $n \geq 3$ are denoted by Euler script letters ($\mathcal{X}$,$\mathcal{Y}$,$\mathcal{Z}$), matrices by capital letters (A, B, C), vectors by lowercase letters (a, b, c). An nth-order tensor is represented by $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_n}$ and its entries are denoted by $x_{i_1 i_2 \cdots i_n}$. The Frobenius norm of $\mathcal{X}$ is defined by $\|\mathcal{X}\| = \left(\sum_{i_1} \sum_{i_2} \ldots \sum_{i_n} x_{i_1 i_2 \ldots i_n}^2\right)^{\frac{1}{2}}$. The L1-norm of $\mathcal{X}$ is defined by $\|\mathcal{X}\|_1 = \sum_{i_1} \sum_{i_2} \ldots \sum_{i_n} |x_{i_1 i_2 \ldots i_n}|$.

The Hadamard product is the elementwise product of two vectors, matrices, or tensors of the same sizes. For instance, two tensors $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_n}$ and $\mathcal{Y} \in R^{I_1 \times I_2 \times \cdots \times I_n}$, their Hadamard product is denoted by $\mathcal{X} * \mathcal{Y}$ and defined as $(\mathcal{X} * \mathcal{Y})_{i_1 i_2 \cdots i_n} = x_{i_1 i_2 \cdots i_n} y_{i_1 i_2 \cdots i_n}$.

The Kronecker product of matrices $A \in R^{I \times J}$ and $B \in R^{K \times L}$, denoted by $A \otimes B$, is a $(IK) \times (JL)$ matrix defined by $A \otimes B = [a_{ij} B]_{IK \times KL}$.

The Khatri-Rao product of matrices $A \in R^{I \times J}$ and $B \in R^{K \times L}$, denoted by $A \odot B$, is a $(IJ) \times (K)$ matrix defined by $A \odot B = [a_1 \otimes b_1 \ a_2 \otimes b_2 \ \cdots a_K \otimes b_K]_{(IJ) \times K}$.

The mode-n unfolding, also known as matricization, of an nth-order tensor $\mathcal{X}$ is denoted by $X_{(n)}$ and arranges the mode-n one-dimensional fibers to be the columns of the resulting matrix (see Fig. 3).

An nth-order tensor $\mathcal{X} \in R^{I_1 \times I_2 \times \cdots \times I_n}$ is rank one if it consists of the outer product of N vectors, i.e., $\mathcal{X} = a^{(1)} \circ a^{(2)} \circ \cdots \circ a^{(N)}$. The symbol "$\circ$" represents the vector outer product.

The standard CANDECOMP/PARAFAC (CP) tensor decomposition factorizes a tensor into a sum of component rank-one tensors, which is expressed by

$$\mathcal{X} = \sum_{r=1}^{R} a_r^{(1)} \circ \cdots \circ a_r^{(N)} = [\![A^{(1)}, \ldots, A^{(N)}]\!] \qquad (1)$$

where $[\![\cdots]\!]$ is a shorthand notation of CP decomposition. $\left\{A^{(n)} | n = 1, \ldots, N\right\}$ are latent factor matrices and can be thought of as the principal components in each mode. The mode-n factor matrix can be denoted by $A^{(n)} = \left[a_1^{(n)}, \ldots, a_R^{(n)}\right] \in R^{I_n \times R}$. An illustration of CP decomposition for third-order tensor is given in Fig.1.

The rank of a tensor is the smallest R for which the above (1) holds, denoted by $R = rank(\mathcal{X})$.

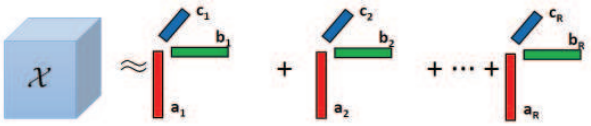

Fig. 1. CP decomposition of a third-order tensor.

III. RELATED WORK

We review the related work on the recovery of the missing data. Interpolation is the mathematical term for filling in missing values. There are many work devoted to interpolate missing data. From the view of data organizational structure, these interpolation methods can be divided into two categories, i.e., matrix-based interpolation methods and tensor-based interpolation methods.

A. Matrix-based Interpolation Methods

Matrix-based methods exploit two-dimensional global information to estimate the missing data. For instance, Non-negative Matrix Factorization (NMF) [4] in the presence of missing entries is formulated as alternating nonnegative least squares problem for recovering the missing values. Sparsity Regularized SVD (SRSVD) [5] creates a SVD-like factorization of matrix, and applies regularization method to optimize the estimation of the missing data.

Compressive Sensing (CS) is a technique that can accurately recover a vector from a subset of samples given that the vector is sparse [6]. CS can be used to recover the missing values with only a few sampled data. M. Roughan et al. [5] propose a novel spatio-temporal compressive sensing framework for TM interpolation, traffic prediction and anomaly detection, in which Sparsity Regularized Matrix Factorization (SRMF) is presented. SRMF leverages low-rank nature of traffic data and their spatio-temporal properties to estimate the missing traffic data. L. Nie et al. [7] propose a power laws-based and compressive sensing method to reconstruct end-to-end network traffic.

Matrix Completion (MC) [8] is closely related to CS. It takes advantage of the low-rank structure of matrix to recover the missing entries. The Singular Value Thresholding algorithm (SVT) [9] is an iterative algorithm for solving the convex relaxation of the approximate matrix completion problem. The Low-rank Matrix Fitting algorithm (LMaFit) [10] is a low-rank factorization model and constructs a nonlinear successive over-relaxation. LMaFit can provide multi-fold accelerations over nuclear-norm minimization on a wide range of matrix completion or low-rank approximation problems.

Matrix-based interpolation methods simply formulate the traffic data into two-dimensional matrix pattern by stacking the columns of TM. The muti-way nature of TM is naively discarded, which cause an unfaithful representation of structural properties for traffic data. Therefore, a matrix is still not enough to capture the comprehensive lower-dimensional latent space in the traffic data, and the data recovery accuracy based on matrix-based approaches is still low.

B. Tensor-based Interpolation Methods

Tensor-based interpolation methods can capture more global information than matrix-based methods due to the intrinsic multidimensional characteristics of tensor model. J. Liu et al. [11] first proposed a tensor completion method based on trace norm minimization and applied it on image completion, in which a high accuracy low rank tensor completion algorithm (HaLRTC) is used for estimating missing visual data. E. Acar et al. [12] develop a CP weighted optimization algorithm (CP-WOPT) that uses a first-order optimization approach to solve the weighted least squares problem and apply to estimate the missing network traffic data. H. Tan et al. [13] construct transportation traffic data as a tensor model and propose a Tucker decomposition based imputation algorithm (TDI) to impute the missing volumes.

A tensor is higher dimensional extension of matrix, which can preserve inherent structural properties in the data. Tensor-based methods exploit the multidimensional structure correlation properties of tensor to estimate the missing entries. The regular tenor completion methods minimize the trace norm of a tensor, i.e. the average of the trace norms of all matrices unfolded along each mode. However, their recovery performance degrades significantly in traffic tensor when the data missing ratio is high. On the other hand, there is no straightforward algorithm to determine the tensor rank of a specific given tensor. The problem is NP-hard [15]. A low-rank tensor decomposition based completion method can not achieve satisfactory prediction accuracy.

To improve the accuracy of the estimation results, we exploit the lower-dimensional latent structure in tensors of traffic data and traffic spatio-temporal properties for the missing data recovery. Based on low-rank tensor CP-decomposition and spatio-temporal information within-mode regularization, we propose to a novel spatio-temporal tensor completion method for recovering the missing traffic data.

IV. PROBLEM FORMULATION

A TM is a representation of the traffic volume flowing between a source i and a destination j. Considering a network with N nodes (computer, routers, etc.), the TM is an $N \times N$ matrix. Since TM evolves over time, a time series of pure spatial traffic matrices can be regarded as 3-dimensional array, i.e., $\mathcal{X} \in R^{N \times N \times T}$ (where there are T time intervals). For instance, the Abilene data [16] contains the traffic exchanged between 11 routers over 6 months collected using 10-minute intervals. This dataset forms a third-order tensor with source routers, destination routers and time modes (see Fig.2). Its each entry, $x_{ijt}(1 \leq i, j \leq N, 0 \leq t \leq T)$, denotes the amount of traffic sent from a source i to a destination j during a particular time interval t.

Definition 1. Binary Index Tensor $\mathcal{W} \in R^{N \times N \times T}$ is a 0-1 tensor, which indicates whether entries of $\mathcal{X}$ are missing,

$$w_{ijt} = \begin{cases} 0 & if \quad x_{ijt} \quad is \quad missing \\ 1 & otherwise \end{cases} \tag{2}$$

Our goal is to estimate the missing traffic data based on the partial direct measurements. To simplify the discussion,

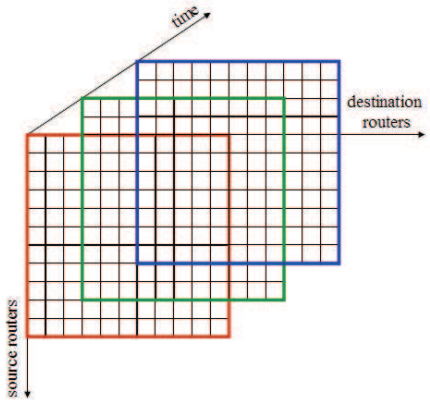

Fig. 2. A third-order tensor of traffic data.

we use the following system of linear equations to formulate the missing value estimation problem:

$$\mathcal{Y} = \mathcal{W} * \mathcal{X} \tag{3}$$

where $*$ denote Hadamard product, the tensor $\mathcal{Y}$ contains the measurements. Note that the presence of missing data is implicit in (3).

We seek an estimated tensor $\hat{\mathcal{X}}$ that satisfies the conditions imposed by the set of measurements. However, it is an underconstrained linear-inverse problem. To solve such problem, one possible approach is to introduce some constraints or prior knowledge about tensor $\mathcal{X}$, i.e., the tensor low-rank model and domain knowledge about relationship among data [17], i.e., spatio-temporal properties.

V. OUR SCHEME: SPATIO-TEMPORAL TENSOR COMPLETION

In this section, we propose a tensor completion method for estimating the missing traffic data. The proposed approach, namely spatio-temporal tensor completion (STTC), utilizes tensor CP-decomposition for completion. In addition, taking advantage of lower-dimensional representation in each mode of tensor, spatio-temporal within-mode regularization is used to improve the completion accuracy.

A. CP-decomposition for Completion

Given a third-order traffic tensor $\mathcal{X} \in R^{N \times N \times T}$, and its rank is R, the CP model can be express as $\mathcal{X} \approx [\![A, B, C]\!] = \sum_{r=1}^{R} a_r \circ b_r \circ c_r$, $A = [a_1, a_2, \cdots, a_R]$ and likewise for B and C, [see Fig.1]. We look for a factorization that satisfies the measurement (3). In the case of incomplete data, the interpolation model based on CP decomposition for missing traffic data can be formulated as,

$$\begin{aligned} minimize \quad f(A,B,C) = &\|\mathcal{W} * (\mathcal{X} - [\![A,B,C]\!])\|^2 \\ &+\lambda \left(\|A\|^2 + \|B\|^2 + \|C\|^2\right) \end{aligned} \tag{4}$$

where $*$ denote Hadamard product. This solution regularizes towards the tensor low-rank approximation but does not strictly enforce the measurement (3). The regularization parameter

λ allows a tunable tradeoff between fitting error and achieving tensor low-rank.

By seeking accuracy in the factors, i.e., A, B and C, it can be used to reconstruct the missing values. The objective of (4) is to find global low-rank structure in the tensor. In addition, we have a priori knowledge that the traffic data has intrinsical spatio-temporal structure.

B. Improvement with Spatio-Temporal

In real-world network, most of traffic data usually change slowly over time, that is, there is little mutation on measured value between adjacent time slots [18]–[21]. These traffic data exhibit temporal stability feature in time dimension. On the other hand, the OD pairs traffic is a combination of different classes of network traffic, which are not independent. At a single measurement interval, the traffic data are close or similar to each other, which exhibit spatial correlation feature [3], [19], [22].

We seek to exploit this insight to improve completion accuracy. Taking advantage of the CP factor matrices as the lower-dimensional representation in each mode of tensor and spatio-temporal within-mode regularization, we propose to combine with (4) and solve the following

$$\begin{aligned} minimize \quad f(A,B,C) &= \|\mathcal{W} * (\mathcal{X} - [\![A,B,C]\!])\|^2 \\ &+ \lambda \left(\|A\|^2 + \|B\|^2 + \|C\|^2\right) \\ &+ \alpha(\|[\![FA,B,C]\!]\|^2 \\ &+ \|[\![A,GB,C]\!]\|^2 \\ &+ \|[\![A,B,HC]\!]\|^2) \end{aligned} \tag{5}$$

where F and G are the spatial constraint matrices and H is the temporal constraint matrix, which expresses our knowledge about the traffic spatio-temporal properties.

To find F, G and H, we estimate an initial tensor $\tilde{\mathcal{X}}$ by interpolating three-dimensional means of known entries into the missing locations in the tensor $\mathcal{X}$. Unfolding third-order estimation tensor $\tilde{\mathcal{X}}$, the mode-1, mode-2 and mode-3 matricization are shown in Fig.3, denoted by $X_{(1)}$, $X_{(2)}$ and $X_{(3)}$ respectively.

The temporal constraint H captures temporal stability feature, i.e., the traffic data is similar at adjacent time slots in the tensor. Based on $X_{(3)}$, we set $H = Toeplitz(0, 1, -1)$ of the size $(T-1) \times T$, which denotes the Toeplitz matrix with central diagonal given by 1, the first upper diagonal given by -1, and the others given by 0, i.e.,

$$H = \begin{bmatrix} 1 & -1 & 0 & \cdots \\ 0 & 1 & -1 & \ddots \\ 0 & 0 & 1 & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

The factor matrix C is the principal components in time dimension of tensor, i.e., temporal lower-dimensional representation. By minimizing $\|[\![A,B,HC]\!]\|^2$, the temporal constraint H functions on the latent space of time dimension of traffic tensor, which approximate the property of having similar values at adjacent time slots.

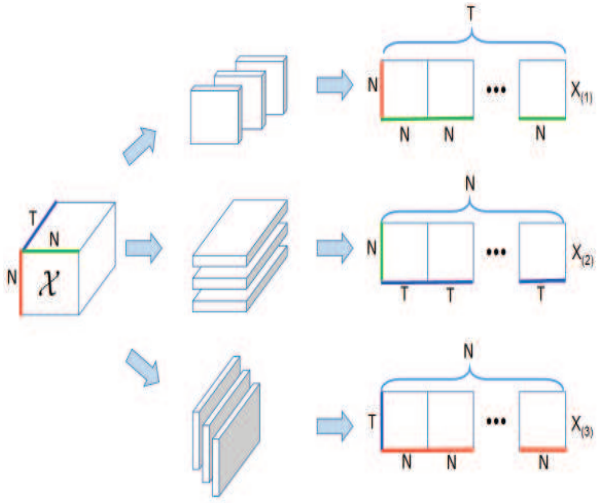

Fig. 3. Unfolding of the $(N \times N \times T)$ tensor to the $(N \times NT)$ matrix $X_{(1)}$, the $(N \times NT)$ matrix $X_{(2)}$ and the $(T \times NN)$ matrix $X_{(3)}$.

The spatial constraint F and G capture spatial correlation feature, i.e., the traffic data is similar in spatial dimension of tensor. Using $X_{(1)}$ and $X_{(2)}$, we chose F and G based on the similarity between rows of $X_{(1)}$ and $X_{(2)}$ respectively. For each row i of $X_{(1)}$, we perform linear regression to find a set of weights $w(n)(n = 1, 2, ..., N-1)$ such that the linear combination of rows j_n best approximates the row i, i.e., $X_{(1)}(i,*) \approx \sum w(n) X_{(1)}(j_n,*)$. Then we set $F(i,i) = 1$ and $F(i,j_n) = -w(n)$. For choosing G, the entire procedure is repeated using $X_{(2)}$.

The matrices F and G express spatial similar relationship of traffic. As the factor matrices A and B are lower-dimensional representation of the spatial dimension, by minimizing $\|[\![FA,B,C]\!]\|^2$ and $\|[\![A,GB,C]\!]\|^2$, the spatial constraint F and G function on the underlying latent structure of spatial dimension, which approximate spatial correlation feature.

C. STTC Algorithm

We propose an efficient STTC algorithm for estimating the missing entries in the traffic tensor. In order to reconstruct the missing data, we derive A, B and C from (5) using an alternating least squares (ALS) procedure. The detail pseudo code is described in Algorithm 1.

The ALS approach fixes B and C to solve for A, then fixes A and C to solve for B, then fixes A and B to solve for C, and continues to repeat the entire procedure until some convergence criterion is satisfied [14].

Having fixed all but one matrix, the problem reduces to a linear least-squares problem. For example, suppose that A and B are fixed. Then, from (5), we can rewrite the above minimization problem for C in matrix form as

$$\begin{bmatrix} W_{(3)} * \left((A \odot B) \times C'\right) \\ \sqrt{\alpha}\left[(F \times A) \odot B; A \odot (G \times B)\right] \times C' \\ \sqrt{\alpha}\,(A \odot B) \times C' \times H' \\ \sqrt{\lambda} C' \end{bmatrix} = \begin{bmatrix} W_{(3)} * X_{(3)} \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

(6) is inconsistent equation, which can be solved by numerical approaches such as the invMinL2 function in MATLAB. The subfunction myInverse in the pseudo code implement solving (6).

Algorithm 1 STTC Algorithm

Input:

$\mathcal{X} \in R^{N\times N\times T}$: input tensor
$\mathcal{W} \in R^{N\times N\times T}$: binary tensor
F, G: spatial constraint matrices
H: temporal constraint matrix
r: tensor rank
α: weight for constraints
λ: regularization parameter
$MaxIter$: max number of iterations

Initialize:

$A \in R^{N\times r}$, $B \in R^{N\times r}$, $C \in R^{T\times r}$;
fval0=f(A, B, C); // compute (5)

for 1 to $MaxIter$ **do**
  $A = myInverse(X_{(1)}, W_{(1)}, B \odot C, [(G \times B) \odot C; B \odot (H \times C)], F, A, \alpha, \lambda)$;
  $B = myInverse(X_{(2)}, W_{(2)}, C \odot A, [(H \times C) \odot A; C \odot (F \times A)], G, B, \alpha, \lambda)$;
  $C = myInverse(X_{(3)}, W_{(3)}, A \odot B, [(F \times A) \odot B; A \odot (G \times B)], H, C, \alpha, \lambda)$;
  $fval = f(A, B, C)$; // update compute (5)
  if $(fval \leq fval0)$ **then**
    $fval0 = fval$;
    $\hat{A} = A$; $\hat{B} = B$; $\hat{C} = C$;
  end if
end for
$\hat{\mathcal{X}} = [\![\hat{A}, \hat{B}, \hat{C}]\!]$;

Output: $\hat{\mathcal{X}}$

VI. EXPERIMENT RESULTS

We conduct extensive simulation experiments to evaluate the performance of the proposed STTC algorithm. We set up a series of missing scenarios, from low loss to high loss probability, and from random loss to highly structured loss patterns. Compared with the state-of-the-art interpolation methods including matrix-based and tensor-based methods, experiment results demonstrate the proposed method achieve significantly better performance.

A. Data Set

Our experiments are performed on two real-world traffic dataset. The first is the Abilene traffic data [16], which consists of 11 nodes in cities all over the United States. So there are total of $11 \times 11 = 121$ OD pairs flows. We use a complete one week traffic data collected using 10-minute intervals, i.e., $6 \times 24 \times 7 = 1008$ time intervals. This data set is used to build a tensor of size $11 \times 11 \times 1008$. The first mode stands for 11 source routers, the second mode for 11 destination routers, and the third mode for 1008 time intervals. The second is the G*É*ANT traffic dataset [23], which is the pan-European research network and composed of 23 routers. We also use a complete one week traffic data collected using 15-minute intervals. This data set forms a tensor of size $23 \times 23 \times 672$, corresponding to source mode, destination mode and time mode respectively.

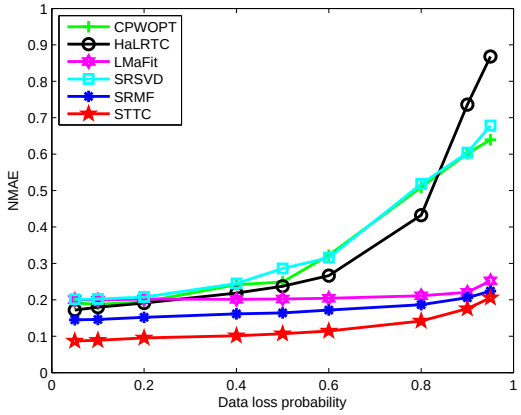

Fig. 4. Random Loss in Abilene data

B. Performance Metrics

The completion performance is evaluated by a popular metric Normalized Mean Absolute Error (NMAE), which measure errors only on the missing values. It is defined as follows,

$$NMAE = \frac{\|(1-\mathcal{W})*(\mathcal{X}-\hat{\mathcal{X}})\|_1}{\|(1-\mathcal{W})*\mathcal{X}\|_1} \tag{7}$$

where $\hat{\mathcal{X}}$ is the reconstructed tensor, $\|\cdot\|_1$ denote L1-norm. As Definition 1, $\mathcal{W}$ indicate the missing locations of entries. The NMAE is the relative error in the missing entries. The smaller the value, the better the performance.

C. Performance on Random Loss Patterns

To demonstrate the effectiveness of our proposed STTC, we compare the completion performance with two tensor-based methods (CP-WOPT and HaLRTC) and three matrix-based methods (SRSVD, LMaFit and SRMF). For the parameter settings of these method, best performance can be achieved according to the corresponding literature [5], [10]–[12]. We randomly drop the data independently with probability Pr ranging from 0.05 to 0.95 to evaluate the completion performance.

Fig.4 shows the comparison results in Abilene dataset. The X-axis presents the data loss probability, and the Y-axis presents the values of NMAE. There is an increase tendency to NMAE with the higher data loss probability.

Among the six interpolation methods, the proposed STTC achieves the best performance. Even when 95% data have been lost (that is, sampling rate is 5%), STTC still can reconstruct the missing data with about 20% errors. SRMF fall behind STTC a little, and perform well in a whole range. LMaFit achieves robust performance over the whole loss range. SRSVD, CP-WOPT, and HaLRTC have nearly the same performance under low loss probability, but the performance becomes worse for high loss probability. CP-WOPT, HaLRTC are not as good as we expected. The possible

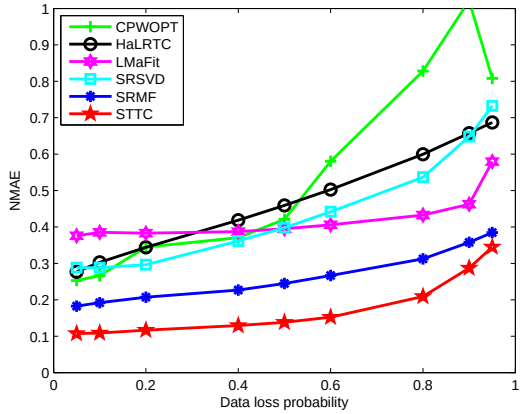

Fig. 5. Random Loss in GÉANT data

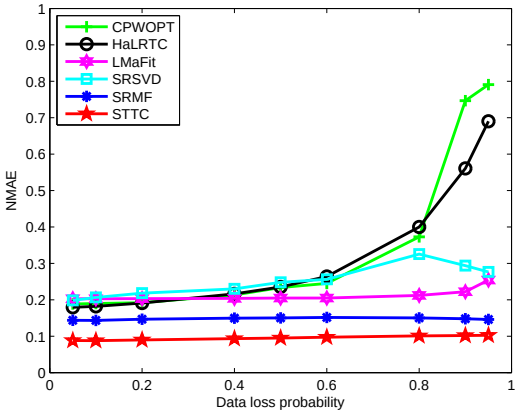

Fig. 6. Abilene data, TL, 50% time intervals chosen

reason is that traffic data tensor is approximation low rank and its rank varies in temporal or spatial domain while tensor-based methods are sensitive to the rank parameter. When the data loss rate is low, sufficient information is available. It is easy to reconstruct the missing data. With the increase of loss data, there is little available auxiliary information. Therefore, reconstructing the missing data become difficult. STTC and SRMF hold very good performance, which illustrate traffic spatio-temporal properties are effective in recovering the missing traffic data. More importantly, STTC outperforms SRMF, which demonstrate tensor can provide a faithful multidimensional structure correlation representation in traffic data and preserve spatio-temporal properties better than the stacked TM.

In Fig.5, we visit GÉANT dataset. It shows that the performance of STTC still outperforms all the other algorithms. Compared with Fig.4, the values of NMAE are larger as a whole. For low loss probability, CP-WOPT and HaLRTC exceed LMaFit and resemble SRSVD. Its performance degrades significantly when the missing ratio is higher than 50%. To a large extent, real-world traffic often exhibit multidimensional characteristics that violate the mathematical conditions under which existing tensor-based algorithms are designed to operate and are provably optimal. Since the scale of GÉANT network is larger, the intrinsic structure nature of the traffic data is weakened. More data are lost, more difficult it is to reconstruct. The best performance of STTC indicates the power of spatio-temporal properties and the multi-dimensional inherent correlation for the missing data recovery.

D. Performance on Structural Loss Patterns

We carry out simulation experiments on structural loss patterns. In practice, network traffic often shows highly structure loss due to software or hardware reasons. We simulate two typical data structure loss patterns.

- Time-mode Loss (TL): this pattern emulates random losses during certain times, which models that monitoring apparatus overloading cause data loss at time dimension of tensor. We randomly chose a certain proportion sampling time intervals, and drop data points with probability Pr from them.
- Spatial-mode Loss (SL): this pattern emulates certain nodes lose data, which models that unreliable transport protocol (i.e., UDP) cause data loss. In traffic tensor, a certain proportion OD pairs are randomly chosen, and data points are dropped with probability Pr from them.

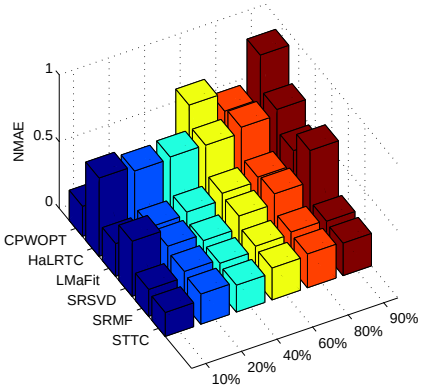

Fig. 7. GÉANT data, SL, Loss Pr=0.95

Fig.6 shows the TL pattern, where 50% sampling time intervals are randomly chosen. In each interval, loss probability Pr is from 0.05 to 0.95. In this case, STTC is very robust with superior performance to the other algorithms. When loss rate is less than 0.5, CP-WOPT and HaLRTC performs well. However, their NMAE increase rapidly for high loss. LMaFit show good performance. SRMF is close to STTC, which further prove that the importance of the spatio-temporal feature. While taking advantage of traffic intrinsic multidimensional structure and traffic characteristic, STTC achieve best performance.

Fig.7 plots the three dimensional histogram of six algorithms with SL pattern when Pr is 0.95. The X-axis presents the proportion of selected OD pairs, from 10% to 90%,

and the Z-axis presents the values of NMAE. Our STTC algorithm still outperforms CP-WOPT, HaLRTC, LMaFit, SRSVD and SRMF. In general, NMAE increases with the more chosen OD pairs. Specifically, when the loss rate is high, the performance degrades significantly except for SRMF and STTC. The reasons, like what has been analyzed previously, are that our proposed approach combine traffic data tensor model and traffic spatio-temporal feature for estimating the missing data.

In summary, STTC outperforms CP-WOPT, HaLRTC, L-MaFit, SRSVD and SRMF over a wide range loss rates and various loss patterns. These results clearly demonstrate that tensor can provide an efficient and faithful representation of structural properties for multidimensional traffic data and spatio-temporal is an important feature for the missing traffic data recovery.

VII. CONCLUSION

In this paper, we studied the inference of the missing network traffic data. To reduce data estimation error, we model network traffic data as tensor pattern. By taking advantage of tensor CP-decomposition and its factor matrices lower-dimensional representation combined with spatio-temporal within-mode regularization, we propose a spatio-temporal tensor completion method (i.e., STTC) to recover the missing traffic data. The extensive simulation experiments show that our proposed method can achieve promising completion accuracy over a wide range of loss scenarios and various loss probabilities.

As part of our future work, we plan to apply traffic spatio-temporal properties to investigate robust low-rank tensor recovery in the presence of missing traffic data, measurement errors, and anomalies.

ACKNOWLEDGMENT

This work is supported by The National Basic Research Program of China (973) under Grant No.2012CB315805 and the National Natural Science Foundation of China under Grant No.61173167, No.61472130 and No.61572184.

REFERENCES

[1] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 365–377, 1996.

[2] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, *Structural analysis of network traffic flows*, vol. 32. ACM, 2004.

[3] P. Tune and M. Roughan, "Internet traffic matrices: A primer," *Recent Advances in Networking*, vol. 1, 2013.

[4] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, pp. 556–562, 2001.

[5] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 3, pp. 662–676, 2012.

[6] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[7] L. Nie, D. Jiang, and L. Guo, "A power laws-based reconstruction approach to end-to-end network traffic," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 898–907, 2013.

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

[9] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[10] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," *Mathematical Programming Computation*, vol. 4, no. 4, pp. 333–361, 2012.

[11] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 208–220, 2013.

[12] E. Acar, D. M. Dunlavy, T. G. Kolda, and M. Mørup, "Scalable tensor factorizations for incomplete data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 41–56, 2011.

[13] H. Tan, G. Feng, J. Feng, W. Wang, Y.-J. Zhang, and F. Li, "A tensor-based method for missing traffic data completion," *Transportation Research Part C: Emerging Technologies*, vol. 28, pp. 15–27, 2013.

[14] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[15] J. Håstad, "Tensor rank is np-complete," *Journal of Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[16] "Abilene/internet2." http://www.internet2.edu/.

[17] A. Narita, K. Hayashi, R. Tomioka, and H. Kashima, "Tensor factorization using auxiliary information," *Data Mining and Knowledge Discovery*, vol. 25, no. 2, pp. 298–324, 2012.

[18] K. Xie, L. Wang, X. Wang, G. Xie, G. Zhang, D. Xie, and J. Wen, "Sequential and adaptive sampling for matrix completion in network monitoring systems," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*, pp. 2443–2451, IEEE, 2015.

[19] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *INFOCOM, 2013 Proceedings IEEE*, pp. 1654–1662, IEEE, 2013.

[20] Z. Huibin, Z. Dafang, X. Kun, and W. Xiaoyang, "Data reconstruction in internet traffic matrix," *Communications, China*, vol. 11, no. 7, pp. 1–12, 2014.

[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 *Distributed Computing Systems (ICDCS), 2014 IEEE 34th International Conference on*, pp. 176–185, IEEE, 2014.

[22] P. Tune and M. Roughan, "Spatiotemporal traffic matrix synthesis," in *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pp. 579–592, ACM, 2015.

[23] S. Uhlig, B. Quoitin, J. Lepropre, and S. Balon, "Providing public intradomain traffic matrices to the research community," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 83–86, 2006.