Computer Communications 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

# Accurate traffic matrix completion based on multi-Gaussian models

Huibin Zhou, Dafang Zhang*, Kun Xie

*College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, PR China*

ARTICLE INFO

*Article history:*
Received 24 November 2015
Revised 23 October 2016
Accepted 24 November 2016
Available online xxx

*Keywords:*
Network measurement
Traffic matrix
Compressive sensing
Matrix completion
Multi-Gaussian models

ABSTRACT

Traffic matrix (TM) describes the volumes of traffic between a set of sources and destinations in a network. As an important parameter, TMs are used in a variety of network engineering tasks, such as traffic engineering, capacity planning and anomaly detection. However, it is a challenge to reliably measure TMs in practice. For example, due to flaws in the measurement systems and possible failure in data collection systems, missing values are unavoidable. It is important to recover the missing data from the partial direct measurements. Existing matrix completion methods do not fully consider network traffic behavior and traffic hidden characteristic. Their completion accuracy tends to be significantly worse when the data loss rate is high. In this paper, we perform a study on intrinsic characteristics of network traffic by analyzing real-world traffic trace data, which reveals that traffic has the features of temporal stability and spatial affinity. According to traffic spatial feature, we model TM as multi-Gaussian distributions, which describes the actual network traffic more accurately. Furthermore, we propose a novel matrix completion method based on multi-Gaussian models to estimate the missing traffic data. Finally, we utilize traffic temporal characteristic to further optimize traffic matrix completion for the missing data interpolation. Our proposed approach has been evaluated utilizing real-world traffic trace data. The extensive experiments demonstrate that our method achieves significantly better performance compared with the state-of-the-art interpolation methods.



## 1. Introduction

Traffic matrix (TM) is an overview of the whole network traffic, which denotes the volumes of traffic (in bytes, packets, or flows) between Origin and Destination (OD) pairs in the network [1]. It is a key input parameter of many network engineering tasks, such as traffic engineering [2], capacity planning and anomaly detection [3]. Unfortunately, it is usually difficult to obtain the complete TMs. Even good measurement systems can suffer from errors and missing data. Added to this, there are still serious impediments to large-scale TM data collection: data collection systems can fail, flow collectors often use an unreliable transport protocol, and legacy network components may not support flow collection or be resource-challenged [4,5]. Therefore, the network-wide TM measurements often have a significant number of missing values. How to cope with the missing data in TMs is still a main challenge. Since many network engineering tasks require the complete traffic information or they are highly sensitive to the missing values, it is important to accurately recover missing values from partial TM measurement data.

To recover the missing data, various studies have been developed, such as local interpolation method K-Nearest Neighbors (KNN) [6], Non-negative Matrix Factorization (NMF) [7,8] and Sparsity Regularized SVD (SRSVD) [4,5]. However, these methods do not scale well, and can not face a lot of missing data. Recently Compressive Sensing (CS) [9,10] theory provides a new paradigm for data acquisition, which utilizes the sparsity characteristic of data to infer the missing data. Many CS-based approaches have been proposed, including environment compressive sensing in sensor networks [11], spatio-temporal compressive sensing framework for traffic interpolation [4,5] and power laws-based compressive sensing reconstruction approach to network traffic [12]. Following CS, matrix completion (MC) [13,14] is a remarkable new field, which takes advantage of the low-rank structure of matrix to recover the missing entries. Many efficient algorithms for solving the low-rank matrix completion have been proposed,*e.g.*, the Singular Value Thresholding algorithm (SVT) [15], the Low-rank Matrix Fitting algorithm (LMaFit) [16] and the gradient descent algorithm on the Grassman manifold (OPTSPACE) [17]. However, these approaches can not been easily applied to estimate the missing traffic data in TM. Because network traffic behavior and traffic hidden characteristic are not being exploited, their completion accuracy tends to be significantly worse when the data loss rate is high.

* Corresponding author.
*E-mail addresses:* zhouhb317@hnu.edu.cn (H. Zhou), dfzhang@hnu.edu.cn (D. Zhang), xiekun@hnu.edu.cn (K. Xie).

http://dx.doi.org/10.1016/j.comcom.2016.11.011
0140-3664/

To design an accurate traffic data recovery algorithm, in this paper, we first analyze the data trace of real traffic, which reveals that there exist spatial-temporal characteristics in the data. Based on traffic spatial feature, TM is modeled as multi-Gaussian models. Inspired by Bayesian inference, we propose a novel matrix completion with multi-Gaussian models to estimate the missing data. Finally, taking advantage of traffic temporal feature, we further optimize traffic matrix completion for the missing data interpolation. Our main contributions are stated as below:

- We conduct a depth study on intrinsic characteristics of network traffic by analyzing the real-world traffic trace data collected from the Abilene network [18] and the GÉANT network [19]. The traffic data exhibits (i) temporal stability feature, *i.e.*, the elements of TM are usually similar at adjacent time slots, and (ii) spatial affinity feature, *i.e.*, some TM rows are close or similar to each other.
- Based on traffic spatial affinity feature, we partition the TM into many clusters by spectral clustering [20]. Thus, the subparts of TM with similar behavior are identified. We model each subpart as Gaussian distributions. Furthermore, we propose a novel matrix completion method based on multi-Gaussian models to estimate the missing traffic data.
- Taking advantage of traffic temporal stability feature, we employ the available neighbors of a missing data point and perform linear regression to interpolate the missing data, which further improves the data recovery performance.
- To evaluate the performance of the proposed approach, we have performed extensive experiments with real-world traffic trace data. The evaluations show that our algorithm can accurately recover the missing traffic data in TM with very low recovery error. Even when 98% of the data is missing, our approach can still reconstruct the TM with the error ratio less than 24%.

The rest of the paper is organized as follows. We introduce the related work and traffic matrix models in Section 2 and formulate the problem in Section 3. Experimental study on the real traffic data is presented in Section 4. We describe the proposed approach in detail in Section 5. Finally, we evaluate the performance of the proposed method through extensive experiments in Section 6, and conclude the work in Section 7.

## 2. Background

### 2.1. Related work

We review the related work on the missing data recovery. Interpolation is the mathematical term for filling in missing values. There are many studies devoted to interpolate the missing data. K-Nearest Neighbors (KNN) [6] is a classical and simple method, which uses the nearest neighbors of a missing data for interpolation. But KNN is not suitable for TM because the rows of TM are ordered arbitrarily, which results in weak correlation between neighbors. Non-negative Matrix Factorization (NMF) [7,8] is a group of algorithms where a matrix V is factorized into (usually) two matrices G and H, with the property that all three matrices have no negative elements. NMF is formulated as alternating nonnegative least squares problem for recovering missing entries in matrix. Sparsity Regularized SVD (SRSVD) [4,5] creates a SVD-like factorization of matrix, and applies regularization method to optimize the estimation of the missing values.

Compressive Sensing (CS) is a technique that can accurately recover a vector from a subset of samples given that the vector is sparse [9,10]. CS can be used to recover missing values with only a few sampled data. For instance, Kong et al. [11] developed the Environmental Space Time Improved Compressive Sensing (ESTI-CS) algorithm for environment data reconstruction in sensor networks. Zhang et al. [4,5] proposed the Sparsity Regularized Matrix Factorization (SRMF) algorithm, which utilizes low-rank structure of TM and traffic spatio-temporal properties to recover the missing traffic data. Nie et al. [12] proposed power laws and compressive sensing to reconstruct network traffic, in which traffic behavior (*i.e.* power-laws distribution) is considered.

Matrix Completion (MC) [13,14] is closely related to CS. It takes advantage of the low-rank structure of matrix to recover the missing entries. Recently, solving the low-rank matrix completion problem is concerned in mathematics. The Singular Value Thresholding algorithm (SVT) [15] is an iterative algorithm for solving the convex relaxation of the approximate matrix completion problem. The gradient descent algorithm on the Grassman manifold (OPTSPACE) [17] is based on singular value decomposition followed by local manifold optimization. The Low-rank Matrix Fitting (LMaFit) [16] is a low-rank factorization model and constructs a nonlinear successive over-relaxation algorithm. LMaFit can provide multi-fold accelerations over nuclear-norm minimization on a wide range of matrix completion or low-rank approximation problems.

Despite much recent progress in these interpolation methods, they utilize either the low-rank structure of matrix, or traffic characteristics, or in some combination. Only a few research investigate traffic models. Our extensive evaluations show that their recovery accuracy is still low when large amounts of data are lost.

To improve the traffic matrix completion accuracy, we fully exploit traffic spatio-temporal properties and traffic model for the missing data recovery. Instead of seeking the low-rank of TM, we model TM as multi-Gaussian models based on the spatial affinity feature, and develop a Bayesian approach to estimate the missing data. Meanwhile, taking advantage of traffic temporal stability feature, we optimize traffic matrix completion for the missing data interpolation.

### 2.2. Traffic matrix models

In order to better understand the network traffic, various traffic models have been developed. We examine some of these models for traffic matrix estimation to discuss how they relate to our approach.

Cao et al. [21] assumed that each OD pair follows a Gaussian distribution, which is utilized to inferring TM from link aggregated traffic data. They [21] also analyze the estimation problem using a moving window, which essentially captures the time-varying nature of the data. However, these are different from our approach. They [21] aim to inferring a TM from available link measurements. The consecutive windows further improve the estimates. In order to recover missing values in TM from partial flow-level measurements, we utilize the temporal feature to optimize traffic matrix completion for missing data recovery.

Vaton et al. [22] proposed to model each OD pair by a mixture of Gaussians, and developed a Bayesian approach to learn the prior distributions of the OD counts from the link counts only. The OD pair is distributed as a mixture of Gaussians: $p(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $\mathcal{N}(x|\mu_k, \Sigma_k)$ is the probability density function (PDF) of the Gaussian distribution with mean $\mu_k$ and standard deviation $\Sigma_k$. $w_k$ is the probability that $x$ stems from component $k$. The parameters $(w_k, \mu_k, \Sigma_k)$ are estimated by an EM algorithm [23]. For estimating the number $K$ of components in the mixture, Dempster et al. [23] used a BIC criterion. By contrast, we show here how spectral clustering method automatically divides the TM into $K$ clusters. Each subpart of the TM is modeled as Gaussian model, so that the TM follows multi-Gaussian distributions. The parameters $(\mu_k, \Sigma_k)$ are estimated by maximum likelihood estimate [24].

Zhao et al. [25] proposed a linear model to estimate the traffic matrix utilizing both flow-level measurements and SNMP data

Please cite this article as: H. Zhou et al., Accurate traffic matrix completion based on multi-Gaussian models, Computer Communications (2016), http://dx.doi.org/10.1016/j.comcom.2016.11.011