Int. J. Electron. Commun. (AEÜ) 68 (2014) 756–762

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

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

Baselining network-wide traffic by time-frequency constrained Stable Principal Component Pursuit☆

Kai Hu, Zhe Wang*, Baolin Yin

State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China

ARTICLE INFO

Article history:
Received 14 May 2013
Accepted 28 February 2014

Keywords:
Baseline of traffic matrix
Robust PCA
Time-frequency constraint
Numerical algorithm
Simulation

ABSTRACT

The Internet traffic analysis is important to network management, and extracting the baseline traffic patterns is especially helpful for some significant network applications. In this paper, we study on the baseline problem of the traffic matrix satisfying a refined traffic matrix decomposition model, since this model extends the assumption of the baseline traffic component to characterize its smoothness, and is more realistic than the existing traffic matrix models. We develop a novel baseline scheme, named Stable Principal Component Pursuit with Time-Frequency Constraints (SPCP-TFC), which extends the Stable Principal Component Pursuit (SPCP) by applying new time-frequency constraints. Then we design an efficient numerical algorithm for SPCP-TFC. At last, we evaluate this baseline scheme through simulations, and show it has superior performance than the existing baseline schemes RBL and PCA.

1. Introduction

The Internet traffic analysis is of critical importance to network operation and management. Usually, the total traffic is modeled by the superposition of diverse components corresponding to different user behaviors [1–4]. The baseline traffic represents the most prominent traffic patterns [3], which is quite helpful for many significant network applications such as capacity planning, load balancing, and anomaly detection. In the past, most studies devoted to estimating the trend of the single-link traffic [5], rather than extracting the common traffic pattern of the whole network, however, the later is more informative for the manager of a large-scale network. Recently, as the network-wide traffic measurement is becoming increasingly popular, efficiently baselining the network-wide traffic turns into a practical and most urgent problem.

The traffic matrix a kind of network-wide traffic data, and it represents the traffic exchanged between each Origin-Destination (OD) pair[1] of the network. Compared with other traffic data such as the link loads, it has a significant advantage [6]: the OD flows are invariant under the changes of topology and routing. Thus the traffic matrix shows the true intensity of the relationships between the OD pairs, and hence is quite helpful for archiving the optimization in capacity planning and traffic engineering, and detecting the network-wide anomalies more accurately. The traffic matrix is obtained either by indirect estimation or by direct measurement [7]. Until very recently, the estimation approach was still an active research topic. But some things have changed around 2010 [8], because the netflow-enabled routers [9] are increasingly deployed, a large percentage of today's networks are able to measure themselves. Hence in many cases, it is feasible to directly measure the traffic matrix now.

The baseline of a traffic matrix captures the common patterns among OD flows, and it should be stable against the disturbance of anomaly traffic. The Principal Component Analysis (PCA) was used for traffic matrix analysis first in [1], and showed the low-rank nature of the baseline (i.e. the deterministic) traffic component, but it performed poorly when the traffic matrix contains large anomalies [4,10]. Recently, the Robust Principal Component Analysis (RPCA) theory [11], which focus on recovering the low-rank matrix contaminated by the sparse matrix whose non-zeros entries may have large magnitudes, has attracted wide attentions. Candes et al. [11] presented the Principal Component Pursuit (PCP) method, and proofed that it could recover the low-rank matrix accurately under very board conditions. Interestingly, the empirical characteristics of the traffic matrix are close to the structural hypotheses of the RPCA theory, since the trends of different OD flows, which are highly correlated, fit for the low-rank hypothesis, and the network anomalies, which rarely appear in time, fit for

☆ This work is supported by the National Science Foundation of China (Nos. 61073013 and 90818024) and the Open Project of State Key Laboratory of Software Development Environment (Nos. SKLSDE-2012ZX-15 and SKLSDE-2013ZX-34).

* Corresponding author. Tel.: +86 10 8231 1642.
E-mail addresses: hukai@buaa.edu.cn (K. Hu), wangzhe@cse.buaa.edu.cn (Z. Wang), yin@nlsde.buaa.edu.cn (B. Yin).

[1] The traffic traversing each OD pair is named an OD flow.

http://dx.doi.org/10.1016/j.aeue.2014.02.016
1434-8411/

the sparse hypothesis. Inspired by this fact, Abdelke et al. [12] first adopted the PCP method for network traffic analysis, while their work mainly considered the anomaly detection problem. Later, Bandara and Jayasumana [3] proposed the Robust Base Line (RBL) scheme, which was also based on PCP and followed the exact "low-rank and sparsity" assumption, and they argued that RBL performs better than several existing traffic baseline schemes.

Even so, it still makes sense to work on this topic more intensively. In fact, the exact "low-rank and sparsity" traffic matrix model in [3,12] is quite simple, and not very realistic. On the one hand, considering the baseline time-series of each OD flow, as it represents the long-term and deterministic traffic trends such as the diurnal pattern, this time-series should be smooth enough. But this feature cannot be characterized by the low-rank assumption. On the other hand, the empirical OD flow traffic also contains the short-term fluctuations behavior with small magnitudes [13]. In this case, the traffic matrix does not exactly meet the "low-rank and sparsity" assumption, instead, it has a noise component. In [4], we modeled the noise traffic, but did not consider the smoothness of the baseline traffic. Consequently, it is necessary to build a more realistic traffic matrix model, and consider the related traffic baseline problem. In addition, the evaluations of traffic baseline schemes were not very sufficient in the previous studies. A key hurdle is that obtaining the ground-truth baseline of the real-world traffic matrix is impossible, as a result, one could neither measure the accuracy of a baseline scheme, nor compare different baseline schemes trustworthily. Hence the simulation approach which contains the ground-truth information is needed in the evaluation process.

In this paper, we study on the baseline problem under a more realistic traffic matrix model, and propose a novel baseline scheme to enforce the smoothness of the baseline traffic component. Our contributions are listed as follows.

- We present a refinement of the traffic matrix decomposition model in [4], which extends the descriptions of the baseline traffic to characterize its smoothness.
- We propose a novel traffic matrix baseline scheme named Stable Principal Component Pursuit with Time-Frequency Constraints (SPCP-TFC). As an extension of the Stable Principal Component Pursuit (SPCP) [14], SPCP-TFC takes new time-frequency constraints.
- We design the Accelerated Proximal Gradient (APG) algorithm for SPCP-TFC, which has a fast convergence rate.
- We evaluate our baseline scheme through simulations and show it has superior performance than RBL and PCA.

2. Methodology

2.1. A Refined Traffic Matrix Decomposition Model

Suppose $X \in \mathbb{R}^{T\times P}$ is a traffic matrix, and each column $X_j \in \mathbb{R}^T$ $(1 \le j \le P)$ is an OD flow in T time intervals. In [4], we proposed the simple Traffic Matrix Decomposition Model (TMDM), assuming X is the sum of a low-rank matrix, a sparse matrix, and a noise matrix. This model is equivalent to the data model of the generalized RPCA problem [14], and the low-rank deterministic traffic matrix corresponds to the baseline traffic.[2] However, TMDM did not consider the temporal characteristics of the baseline traffic. Since the baseline traffic time-series of each OD flow represents the long-term and steady user behaviors, it tends to display a smooth curve. A number of mathematical tools, such as the wavelets and the splines [15], can formulate smoothness. As the most salient baseline traffic patterns are slow oscillation behaviors, in this paper, we formulate the baseline traffic time-series as the sum of harmonics with low frequencies, and thus establish a Refined Traffic Matrix Decomposition Model (R-TMDM):

Definition 1 (R-TMDM) *The traffic matrix $X \in \mathbb{R}^{T\times P}$ is the superposition of the deterministic (baseline) traffic matrix A, the anomaly traffic matrix E, and the noise traffic matrix N. A is a low-rank matrix, and for each column time-series in A, the Fourier spectra whose frequencies exceed a critical value f_c are zeros; E is a sparse matrix with most entries being zeros, but the non-zeros entries may have large magnitudes; N is a random noise matrix, and each column time-series is a zero-mean stationary random process with a relatively small variance.*

As the OD flows in the backbone network are highly aggregated by superimposing independent traffic processes, it is appropriate to model the noise traffic by the Gaussian processes following the central limitation theory [6,16]. For simplicity, we assume each time-series N_j $(1 \le j \le P)$ is the white Gaussian noise with variance $\sigma_j^2 > 0$ in this study.[3]

2.2. Stable Principal Component Pursuit with Time-Frequency Constraints

Let $\|\cdot\|_*$, $\|\cdot\|_1$, and $\|\cdot\|_F$ denote the nuclear norm, the l_1 norm, and the Frobenius norm, respectively. The Stable Principal Component Pursuit (SPCP) method for the generalized RPCA problem solves this convex program [14]:

$$\begin{aligned} &\text{minimize}_{A,E,N} \|A\|_* + \lambda\|E\|_1 \\ &s.t. \quad A + E + N = X, \|N\|_F^2 \le \delta, \end{aligned} \tag{1}$$

where $\lambda > 0$ is a balance parameter, and $\delta > 0$ is a constraint parameter. The objective function of (1) combines the nuclear norm and the l_1 norm, which are the convex relaxations of the rank function and the l_0 norm, respectively, to enhance the low-rank structure of matrix A, as well as the sparsity of matrix E.

In this study, in order to extract the baseline traffic (i.e. matrix A) more accurately, we extend SPCP by preserving its objective function and redesigning the constraint functions. Firstly, considering the R-TMDM model, it is necessary to add a constraint for the baseline traffic matrix based on its frequency-domain assumption. Let $W = [W_0 \cdots W_{T-1}]_{T\times T}$ denote the discrete Fourier basis matrix of length T. For each $0 \le k \le T-1$, the Fourier basis W_k is defined as

$$W_k(t) = \frac{1}{\sqrt{T}} e^{-i(2\pi k/T)(t-1)}, \quad 1 \le t \le T, \tag{2}$$

with frequency $f_k = \min\{(k/T), (T-k/T)\}$. Suppose W_H is made up of the high-frequency bases W_k in W satisfying $f_k \ge f_c$, and thus $W_H(W_H)^\top$ is the projection operator to the high-frequency subspace. Hence we add the following constraint for the baseline traffic:

$$W_H(W_H)^\top A = 0^{T\times P}. \tag{3}$$

Secondly, unlike the Frobenius norm inequality in (1), we use a different constraint strategy for the noise traffic matrix N. For each column vector N_j $(1 \le j \le P)$, consider its periodogram function $\{I_{N_j}(k)\}_{k=0}^{T-1}$:

$$I_{N_j}(k) = |W_k^\top N_j|^2 = \left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} N_j(t) e^{-i(2\pi k/T)(t-1)} \right|^2, \quad 0 \le k \le T-1. \tag{4}$$

[2] In the following discussion, the words "deterministic traffic" and "baseline traffic" are used interchangeably.

[3] In future work, we plan to consider a more general model, i.e. each noise traffic time-series is a fractional Gaussian noise, whose temporal characteristics are more similar to the real-world backbone traffic.