Multi-Low-Rank Approximation For Traffic Matrices

Saurabh Verma, Arvind Narayanan, Zhi-Li Zhang
Department of Computer Science and Engineering, University of Minnesota, USA
Email: {verma,arvind,zhzhang}@cs.umn.edu

***Abstract*—With the Internet applications become more complex and diverse, simple network traffic matrix estimation or approximation methods such as gravity model are no longer adequate. In this paper, we advocate a novel approach of approximating traffic matrices with multiple low-rank matrices. We build the theory behind the MULTI-LOW-RANK approximation and discuss the conditions under which it is better than Low-Rank SVD in terms of both matrix approximation and preserving the local (and global) structure of the traffic matrices. Further, we develop an effective technique based on spectral clustering of column/row feature vectors for decomposing traffic matrices into multiple low rank matrices. We perform a series of experiments on traffic matrices extracted from a synthetic dataset and two real world datasets – one that represents nationwide cellular traffic and another taken from a tier-1 ISP. The results thus obtained show that: 1) MULTI-LOW-RANK approximation is superior for traffic classification; 2) it can be used to predict complete or missing entries of traffic matrices over time; 3) show it's robustness against noise; and, 4) demonstrate that it closely follows the optimal solution (i.e., low-rank SVD solution).**

***Index Terms*—Traffic matrix approximation, Low rank SVD, Traffic classification.**

I. INTRODUCTION

The wide proliferation of various kinds of sensors in the physical and/or cyber worlds has enabled us to collect a whole gamut of (spatial-temporal) data, e.g., voice calls between users at various locations in a cellular network, traffic between different points-of-presence (PoPs) in an ISP (Internet service provider) network, human mobility or commuting behaviors across different locations in a transport network. Traffic matrices such as origin-destination (OD) matrices are a natural way to represent many of these datasets arising in these application domains. Here, origins and destinations may refer to a person, a location or a physical object or an abstract entity. Each cell of an OD matrix quantifies the relation between a pair of origin and destination using certain metrics, e.g., traffic volume, activity counts, that is observed during a given time interval. We will be using the terms traffic matrices and OD matrices interchangeably. With abundance of such data, extracting meaningful patterns from OD matrices is an important data analysis task that has wide applications, from network traffic engineering to urban transportation management, smart city planning, social behavior analysis and cyber-physical world security.

Perhaps the most prominent area where OD matrices have been widely studied in the past is Internet traffic analysis, where the *gravity* model – originally proposed in traffic analysis in transportation networks [1] – and its extensions have been developed for PoP-level IP traffic matrix estimation [2], [3], [4]. These approaches essentially assume that a (PoP-level) OD traffic matrix can be approximated by a low rank matrix, and in the case of the standard gravity model, a $\mathrm{rank}-1$ matrix. Such an assumption may be justified, if traffic flows on different links of the network are roughly independent [5]. The goal is to characterize the *entire* OD matrix for the purpose of IP traffic matrix estimation. However, the emergence of large cloud-based application service providers such as Google, Facebook, Amazon, Netflix, coupled with the dominant role of content distribution networks (CDNs) in content delivery, has altered the Internet traffic dynamics. It is shown in [6] that the simple gravity model is no longer sufficient to capture and model the IP traffic matrices in a large ISP network.

In this paper, we start with a brief discussion on related work for traffic matrix approximation based on conventional methods such as PCA/SVD and NMF (non-negative matrix factorization). All these approaches implicitly assume that observed data points come from a *latent* linear subspace with "noises", and thus can be approximated using a low-rank matrix. However, as we have observed from many real-world network traffic matrices such as those from voice communications in cellular networks or data communications in large ISP backbone networks, this assumption is no longer valid. Instead, we postulate that the observed global traffic matrix in a network is likely an aggregate of many diverse traffic patterns, each of which reflects a distinct class of application/user communication structures or behaviors, and thus can be approximated by a low-rank matrix. In other words, the observed data points represent a mixture of several (latent) low-dimensional linear sub-manifolds. In such a setting, we argue that using the standard SVD/PCA to approximate the entire traffic matrix is not adequate; to preserve the local structures inherent in the data, it is best to approximate it using multiple low-rank matrices, each capturing one latent sub-manifold (see Section III).

Based on the above intuition, we develop a theory behind our proposed MULTI-LOW-RANK approximation method and discuss the conditions under which its better than a Low-Rank SVD in terms of both matrix approximation and preserving the local (and global) structure of the traffic matrices (see Section IV). In order to identify and extract sub-matrices corresponding to clusters of data points lying in various linear sub-manifolds, we develop an effective technique based on spectral clustering of row/column feature vectors. We first convert the original OD matrix to generate a new (feature) similarity graph using Gaussian kernel by treating each row (or column) as a feature vector. We then apply spectral clustering to project the feature space into lower dimensional manifolds for traffic matrix decomposition and MULTI-LOW-RANK matrix approximation (see Section VI). In Sections VII–XI, we perform a series of experiments on traffic matrices extracted from a synthetic dataset and two real world datasets – one that represents nationwide cellular traffic and another taken from a tier-1 ISP. The results thus obtained show that MULTI-LOW-RANK approximation is superior for traffic classification. It can be used to predict complete or missing entries of traffic matrices over time and also it is robust against noise and closely follows the optimal solution (via SVD).

II. BACKGROUND AND RELATED WORK

One of the earlier attempts in extracting meaningful patterns in traffic matrices is made by [7] where principal component analysis (PCA) was employed for approximating the low rank matrix by performing either eigenvalue decomposition of covariance or singular value decomposition (SVD) of the data matrix. Since then, many variants of the PCA-based approach (e.g., robust PCA for handling noise and outliers) and other related techniques such as latent semantic indexing (LSI) have been developed for anomaly detection and network tomography [8], [9], [10]. The basic premise of PCA (or SVD or LSI) is that the data lies in a *linear* subspace of a high dimensional space. This premise is invalidated when there are non-linear relations or patterns in the data, although such patterns may lie in multiple linear sub-manifolds of low-dimensional space and have low *intrinsic* dimensions. For such cases, PCA (or SVD) would fail to capture the local structure of the data and won't be able to approximate the low rank matrix accurately. Another matrix factorization approach is NMF [11] developed to address the interpretability issue associated with the low-rank matrix approximations. However, NMF also assumes that the entities lie in a lower linear subspace of the original high dimensional data space.

Besides analysis through matrix decomposition, much of the previous work in traffic matrix prediction focus on filling missing entries in a partially observed matrix under various assumptions such as sparsity or spatio-temporal constraints [12], [13], [14]. Most of these problems are solved using the techniques obtained from compressive sensing or matrix completion [15]. However, in our case we are dealing with partially observed matrices and are interested in predicting full matrix over time in future. This is possible, if the structure of traffic matrices remain intact over time in the sense that only few local structures of matrix changes at a particular time over different time domains. In [16], such a problem is addressed by first constructing partially observed matrix by identifying important OD flow links from past data and then converting the main problem into popular traffic matrix prediction with missing entries. This approach only takes global structure of the data into account and does not make an attempt to preserve local, but important, structures of the data.

III. MULTI-LOW-RANK INTRODUCTION AND MOTIVATION

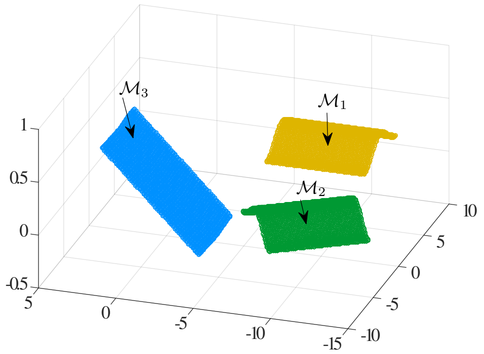


Fig. 1: Data consisting of three almost linear sub-manifolds $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$.

In this paper, we are interested in approximating a class of Internet traffic matrices composed of multiple linear sub-manifolds as shown in Figure 1. This special structure allows us to predict such traffic matrices over time domain based on the information available from present day traffic matrices. This is possible because, over the time domain, most local structures (or clusters) are preserved while few of them change in traffic matrices and can be accomplished by approximating matrices with multiple low-rank sub-matrices. Through series of experiments, we demonstrate and argue that it is more appropriate to account for these clusters while approximating the traffic matrices.

Consider a low rank $\widehat{\mathbf{A}} \in \mathbb{R}^{n\times m}$ matrix approximation of $\mathbf{A} \in \mathbb{R}^{n\times m}$ matrix. Then in general, $\widehat{\mathbf{A}}$ is obtained through minimizing following objective function, where $\|.\|_F$ is the Frobenius norm of matrix:

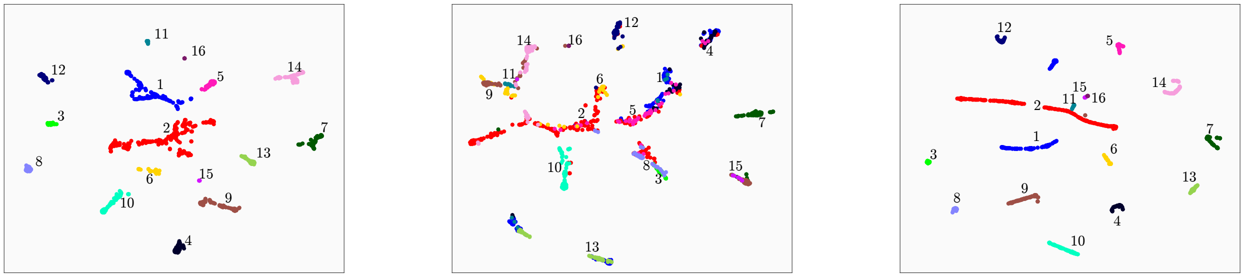
$$\begin{aligned} &\underset{\widehat{\mathbf{A}}}{\text{minimize}} && \|\mathbf{A} - \widehat{\mathbf{A}}\|_F^2 \\ &\text{subject to} && \operatorname{rank}(\widehat{\mathbf{A}}) \leq r, \end{aligned} \tag{1}$$

The optimal solution of the above function is obtained through Low-Rank SVD of $\mathbf{A}$ i.e., $\widehat{\mathbf{A}} = \widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T$, where $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{r\times r}$ contains top r singular values and $\widehat{\mathbf{U}} \in \mathbb{R}^{n\times r}, \widehat{\mathbf{V}}_r \in \mathbb{R}^{m\times r}$ are orthogonal matrices containing the corresponding r left and r right singular vectors, respectively.

SVD provides the best approximation of a matrix under rank constraint but does not guarantee preserving the important local structure (i.e., clusters) of the data. For example, consider a special case of block diagonal matrix $\mathbf{A}$ commonly seen as origin-destination (OD) matrices in representing Internet traffic data. Here, we are interested in a low rank matrix approximation while also being able to preserve the useful local structure for this block diagonal matrix.

$$\text{Let} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & 0 & 0 \\ 0 & \mathbf{A}_2 & 0 \\ 0 & 0 & \mathbf{A}_3 \end{bmatrix} = \mathbf{U\Sigma V}, \quad \text{where,}$$

$$\mathbf{U\Sigma V} = \begin{bmatrix} \mathbf{U}_1 & 0 & 0 \\ 0 & \mathbf{U}_2 & 0 \\ 0 & 0 & \mathbf{U}_3 \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_1 & 0 & 0 \\ 0 & \mathbf{\Sigma}_2 & 0 \\ 0 & 0 & \mathbf{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \mathbf{V}_1 & 0 & 0 \\ 0 & \mathbf{V}_2 & 0 \\ 0 & 0 & \mathbf{V}_3 \end{bmatrix}$$



(a) Structure of a Real Traffic Dataset. Depicts total 16 local structures/clusters.

(b) Structure of the data after Low-Rank SVD approximation with error rate $= 41.28\%$.

(c) Structure of the data after Multi-Low-Rank SVD approx. with error rate $= 46.66\%$.

Fig. 2: Visualization of the local and global structure of a Real Traffic Dataset in $\mathbb{R}^2$ using t-SNE. Comparing Figures 2b & 2c reveal that MULTI-LOW-RANK SVD approx. (with each sub-matrix $\text{rank} = 1$) preserves the structure of the data much better than Low-Rank SVD approx. (having single matrix $\text{rank} = 16$) by sacrificing a small amount of approximation error rate.

Let $\widehat{\mathbf{A}}$ be the Low-Rank SVD approximation of $\mathbf{A}$ with $\text{rank} = r$ and let $\sigma_r(\mathbf{A}_1) \geq \sigma_1(\mathbf{A}_2)$, where $\sigma_j(\mathbf{A})$ is the top j^{th} singular value of $\mathbf{A}$ matrix. Then, $\widehat{\mathbf{A}}$ can be obtained as:

$$\widehat{\mathbf{A}} = \begin{bmatrix} \widehat{\mathbf{U}}_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{\Sigma}}_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{V}}_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

As a result, no information about $\mathbf{A}_2$ and $\mathbf{A}_3$ clusters are extracted from the matrix. Even, if we *increase the rank* r, it is quite possible that a dominant cluster with higher singular values can mask the extraction of less dominant but important local clusters of the data. To solve this issue, we propose MULTI-LOW-RANK approach to preserve the useful cluster information in the matrix approximation. Specifically, we consider approximating matrix $\mathbf{A}$ via multiple low rank sub-matrices $\{\widehat{\mathbf{A}}_s\}_{s=1}^{C}$, where each $\widehat{\mathbf{A}}_s$ contains some local structure information and C is total number of sub-matrices.

To motivate further, consider a Real Traffic Dataset (see dataset description and definition of approximate error rate in Section VIII). We apply the state-of-art visualization technique, t-SNE [17] in conjunction with spectral clustering to reveal the local and global structure of the data in $\mathbb{R}^2$ plane as shown in Figure 2a. Note that, we found a total 16 local structures/clusters in the data. Then, we proceed to apply t-SNE on the approximation data obtained through Low-Rank SVD and our proposed MULTI-LOW-RANK SVD approximation method. For MULTI-LOW-RANK SVD approximation, we set $\text{rank} = 1$ for each sub-matrix/cluster. To make a fair comparison with Low-Rank SVD, we approximate the (single) data matrix with $\text{rank} = 16$. This results in obtaining Figures 2b & 2c which clearly show that Multi-Low-Rank SVD is able to preserve the overall structure of the data much better than Low-Rank SVD without much compromising the quality of approximation (only $\sim 5.38\%$ of accuracy is lost).

IV. MULTI-LOW-RANK APPROXIMATION THEORY

In this section, we provide some theoretical results to help justify the approach taken in this paper and the condition under which MULTI-LOW-RANK SVD is superior than Low-Rank SVD, particularly for approximation purposes only. First, we state the following proposition:

Proposition 1. *Let* $\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}$ *be a matrix and* $\widehat{\mathbf{S}} = \begin{bmatrix} \widehat{\mathbf{S}}_1 \\ \widehat{\mathbf{S}}_2 \end{bmatrix}$ *be the approximation of* $\mathbf{A}$ *under the rank constraint* ($\text{rank} \leq r$) *on sub-matrices* $\widehat{\mathbf{S}}_1, \widehat{\mathbf{S}}_2$. *Then, the optimal approximation is given by* $\widehat{\mathbf{S}} = \begin{bmatrix} \widehat{\mathbf{A}}_1 \\ \widehat{\mathbf{A}}_2 \end{bmatrix}$, *where* $\widehat{\mathbf{A}}_1, \widehat{\mathbf{A}}_2$ *are* $\text{rank} \leq r$ *matrices obtained via Low-Rank SVD.*

Proof. Formally, the problem can be stated as:

$$\begin{aligned} \underset{\widehat{\mathbf{S}}_1, \widehat{\mathbf{S}}_2}{\text{minimize}} \quad & ||\mathbf{A} - \widehat{\mathbf{S}}||_F^2 \\ \text{subject to} \quad & \text{rank}(\widehat{\mathbf{S}}_1) \leq r, \text{rank}(\widehat{\mathbf{S}}_2) \leq r \end{aligned} \tag{2}$$

The proof is based on the following simple observation:

$$||\mathbf{A} - \widehat{\mathbf{S}}||_F^2 = \underbrace{||\mathbf{A}_1 - \widehat{\mathbf{S}}_1||_F^2}_{f_1(\widehat{\mathbf{S}}_1)} + \underbrace{||\mathbf{A}_2 - \widehat{\mathbf{S}}_2||_F^2}_{f_2(\widehat{\mathbf{S}}_2)} \tag{3}$$

Since $f_1(\widehat{\mathbf{S}}_1)$, $f_2(\widehat{\mathbf{S}}_2)$ and corresponding constraint variables are independent, equation 2 can be written as:

$$\begin{aligned} & \min_{\widehat{\mathbf{S}}_1} ||\mathbf{A}_1 - \widehat{\mathbf{S}}_1||_F^2 \quad + \quad \min_{\widehat{\mathbf{S}}_2} ||\mathbf{A}_2 - \widehat{\mathbf{S}}_2||_F^2 \\ & \text{s.t.} \quad \text{rank}(\widehat{\mathbf{S}}_1) \leq r, \qquad \text{rank}(\widehat{\mathbf{S}}_2) \leq r \end{aligned} \tag{4}$$

Thus, the optimal solution is straight forwardly obtained via SVD of $\mathbf{A}_1, \mathbf{A}_2$ as shown from Eq. (1).

Remarks. Proposition 1 shows that under the rank constraint on sub-matrices of a matrix, Low-Rank SVD still provides the best approximation when performed on the corresponding sub-matrices. The above results is valid for only Frobenius norm case and not necessarily holds for spectral norm.

Next, we derive the condition under which approximating matrix with multiple low-rank sub-matrices is better than approximating with single/global rank for overall matrix.

Theorem 1. *Let* $\mathbf{A}_{n \times m}$ *be a matrix of the given form:*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix}, \widehat{\mathbf{S}} = \begin{bmatrix} \widehat{\mathbf{A}}_1 & \widehat{\mathbf{A}}_2 \\ \widehat{\mathbf{A}}_3 & \widehat{\mathbf{A}}_4 \end{bmatrix}$$

and $\mathbf{A}_s$ *are sub-matrices. Let* $\widehat{\mathbf{A}}$ *be the Low-Rank SVD of* $\mathbf{A}$ *with rank* $r \leq \min(n, m)$. *Similarly, let* $\mathbf{S}$ *contain sub-matrices* $\widehat{\mathbf{A}}_s$ *of rank* $r_s \leq \min(n_s, m_s)$.

Then $\|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2 \leq \|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2$ holds, if it satisfies the condition $\big(\sum\limits_{s=1}^{4}\sum\limits_{j=1}^{r_s}\sigma_{sj}^2 - \sum\limits_{j=1}^{r}\sigma_j^2\big) \geq 0$, where σ_j is the top j^{th} singular value of $\mathbf{A}$ and σ_{sj} is the top j^{th} singular value of $\mathbf{A}_s$ matrix.

Proof: Low-Rank SVD matrix approximation error for rank r is given by (here $q_s = \min(n_s, m_s)$).

$$\begin{aligned}\|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2 &= \sum_{j=r+1}^{q}\sigma_j^2 \\ \|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2 &= \|\mathbf{A}_1-\widehat{\mathbf{A}}_1\|_F^2+\cdots+\|\mathbf{A}_4-\widehat{\mathbf{A}}_4\|_F^2 \qquad (5)\\ &= \sum_{s=1}^{4}\sum_{j=r_s+1}^{q_s}\sigma_{sj}^2\end{aligned}$$

Further, we can rewrite $\mathbf{A}$ in terms of sub-matrices as:

$$\begin{aligned}\text{trace}(\mathbf{A}^T\mathbf{A}) &= \text{trace}(\mathbf{A}_1^T\mathbf{A}_1)+...+\text{trace}(\mathbf{A}_4^T\mathbf{A}_4)\\ \sum_{j=1}^{q}\sigma_j^2 &= \sum_{s=1}^{4}\sum_{j=1}^{q_s}\sigma_{sj}^2 \qquad (6)\end{aligned}$$

Here, we used the fact that eigenvalues of $\mathbf{A}^T\mathbf{A}$ are square of singular values of $\mathbf{A}$ and $\text{trace}(\mathbf{A}^T\mathbf{A})$ corresponds to the sum of eigenvalues of a matrix. Finally, using Eq. (5) & (6):

$$\|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2 - \|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2 = \sum_{s=1}^{4}\sum_{j=1}^{r_s}\sigma_{sj}^2 - \sum_{j=1}^{r}\sigma_j^2 \qquad (7)$$

Then it follows, $\|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2 \geq \|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2$, if we have $\Big(\sum\limits_{s=1}^{4}\sum\limits_{j=1}^{r_s}\sigma_{sj}^2 - \sum\limits_{j=1}^{r}\sigma_j^2\Big) \geq 0$.

Remarks. Above result hold for a more general case of block matrices and proof can be derived in a similar manner as shown above. The above theorem states that in order to approximate the low rank matrix better than the Low-Rank SVD matrix of rank r, the sum of square of singular values of Low-Rank sub-matrices must be greater than the sum of square of singular values of Low-Rank SVD matrix of rank r. The result is intuitive in the sense that we tend to seek the top dominant directions in the data measured by singular values and higher values of these (singular values) under the constraint of preserving local structure suggest better approximation of the data.

Corollary 1.1. *Let $\mathbf{A}_{n\times m}$ be a matrix of the given form and let $\widehat{\mathbf{S}}$ contains sub-matrices $\mathbf{A}_s$ of rank r.*

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_1\\ \vdots \\ \mathbf{A}_c\end{bmatrix}, \widehat{\mathbf{S}} = \begin{bmatrix}\widehat{\mathbf{A}}_1\\ \vdots \\ \widehat{\mathbf{A}}_c\end{bmatrix}$$

Then $\|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2 \leq \|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2$, if $\big(\sum\limits_{s=1}^{c}\sum\limits_{j=1}^{r}\sigma_{sj}^2 - \sum\limits_{j=1}^{r}\sigma_s^2\big) \geq 0$, where σ_j is the top j^{th} singular value of $\mathbf{A}$ and σ_{sj} is the top j^{th} singular value of $\mathbf{A}_s$ matrix.

Proof. Directly can derived from Theorem 1.

Remarks. Here, the sum of (square of) eigenvalues of each sub-matrices would contribute towards the quality of matrix approximation. Therefore, identifying dominant (local) principal components will result in reduced matrix approximation error along with capturing the local structure of the data.

Corollary 1.2. *Let $\mathbf{A}_{n\times n}$ be a block diagonal matrix of the given form and let $\widehat{\mathbf{S}}$ contain sub-matrices $\mathbf{A}_s$ of rank r.*

$$\mathbf{A} = \begin{bmatrix}\mathbf{A}_1 & \dots & 0\\ & \ddots & \\ 0 & \dots & \mathbf{A}_c\end{bmatrix}, \widehat{\mathbf{S}} = \begin{bmatrix}\widehat{\mathbf{A}}_1 & \dots & 0\\ & \ddots & \\ 0 & \dots & \widehat{\mathbf{A}}_c\end{bmatrix}$$

Then, $\|\mathbf{A}-\widehat{\mathbf{S}}\|_F^2 \leq \|\mathbf{A}-\widehat{\mathbf{A}}\|_F^2$, is always satisfied.

Proof. For block diagonal matrix, it is easy to see that σ_j are top r singular values of the following set $\{\sigma_{sj} : \sigma_{sj} \in \sigma(A_s), \forall s\}$, where $\sigma(\mathbf{A}_s)$ are singular values of $\mathbf{A}_s$, and from there, it directly follows $\big(\sum\limits_{s=1}^{c}\sum\limits_{j=1}^{r}\sigma_{sj}^2 - \sum\limits_{j=1}^{r}\sigma_j^2\big) \geq 0$.

Remarks. Thus, if the data has a block diagonal form structure, it is always better to approximate matrix through MULTI-LOW-RANK method which will guarantee to provide lower approximation error than Low-Rank SVD. Moreover, it also helps to preserve both local and global structure of the data.

A. Computational Complexity

Time complexity of exact SVD of a $\mathbf{A}_{n\times d}$ matrix is bounded by $\mathcal{O}(\min\{n^2d, nd^2\})$ [18]. As a result, in the case of MULTI-LOW-RANK approximation, the complexity will be equal to $\sum\limits_{s=1}^{c}\mathcal{O}(\min\{n_s^2d, n_sd^2\})$ where $\sum\limits_{s=1}^{c}n_s = n$ and c is number of clusters. If $n{>}d$, then the time complexity of SVD and MULTI-LOW-RANK approximation would be same i.e. $\mathcal{O}(nd^2)$. But for high dimensional data, where $n{<}d$, the complexity of MULTI-LOW-RANK is less than SVD since $\sum\limits_{s=1}^{c}\mathcal{O}(n_s^2d){<}\mathcal{O}(n^2d)$.

V. MULTI-LOW-RANK BASED PROBABILISTIC MATRIX FACTORIZATION

We extend our MULTI-LOW-RANK approach to include matrices with missing entries. Since, the objective function of SVD under a matrix with missing entries (where eigenvalue decomposition is not possible anymore to give optimal solution and therefore need to perform optimization through iterative methods) is prone to over-fitting, we a seek more robust method, specifically, probabilistic matrix factorization technique (PMF) [19]. Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be a matrix composed of multiple $\{\mathbf{A}_c \in \mathbb{R}^{n_c\times m}\}_{c=1}^{C}$ sub-matrices and each of them derived from $\{\mathcal{D}_c\}_{c=1}^{C}$ different distributions to represent C clusters in the data. Then, probabilistic matrix factorization can be generalized to multi-distributions as follows:

$$p(\mathbf{A}|\{\mathbf{U}_c, \mathbf{V}_c, \sigma_c\}_1^C) = \prod_{c=1}^{C}\prod_{i=1}^{n_c}\prod_{j=1}^{m}\Big[\mathcal{N}(\mathbf{A}_{c_{ij}}|\mathbf{U}_{c_i}^T\mathbf{V}_{c_j}, \sigma_c^2)\Big]^{I_{c_{ij}}} \qquad (8)$$

where, $\mathbf{U}_c \in \mathbb{R}^{r \times n_c}, \mathbf{V}_c \in \mathbb{R}^{r \times m}$ are latent matrices of rank r with $\mathbf{U}_{c_k}, \mathbf{V}_{c_k}$ as k^{th} column in the matrix respectively. Also, $\mathcal{N}(x|\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. I_{ij} is indicator matrix with entry 0 for missing entries in matrix $\mathbf{A}_c$ or else 1. Assuming zero-mean spherical Gaussian priors on $\mathbf{U}_c$ & $\mathbf{V}_c$ and then deriving the log of posterior distribution of Eq. (8) and applying MAP inference, results in minimization of following objective function:

$$E = \sum_{c=1}^{C} \sum_{i=1}^{n_c} \sum_{j=1}^{m} I_{c_{ij}} (\mathbf{A}_{c_{ij}} - \mathbf{U}_{c_i}^T \mathbf{V}_{c_j}) + \sum_{c=1}^{C} \lambda_{c_u} \sum_{i=1}^{n_c} \|\mathbf{U}_{c_i}\|_F^2 + \sum_{c=1}^{C} \lambda_{c_v} \sum_{j=1}^{m} \|\mathbf{V}_{c_j}\|_F^2 \tag{9}$$

where, $\lambda_{c_u} = \frac{\sigma_c^2}{\sigma_{c_u}^2}$, $\lambda_{c_v} = \frac{\sigma_c^2}{\sigma_{c_v}^2}$ and σ_{c_u} & σ_{c_v} are prior variances on $\mathbf{U}_c$ & $\mathbf{V}_c$ respectively. Directly optimizing Eq. (9) is NP-hard, since the distribution assignment for each entry of $\mathbf{A}$ is unknown. But, we can work around this problem by first performing clustering which provides distribution assignment for each entry. Secondly, assuming distributions are independent enable us to decouple Eq. (9) into $E = \sum_{c=1}^{C} E_c$, where each E_c corresponds to PMF of each sub-matrix and can be solved independently. Also note that Eq. (9) reduces exactly to MULTI-LOW-RANK SVD under the condition that all prior variances $\sigma_{c_u}, \sigma_{c_v} \to \infty$.

VI. DECOMPOSING TRAFFIC MATRICES INTO MULTI-LOW-RANK MATRICES

For decomposing heterogeneous traffic matrices into multiple sub-matrices with lower ranks, we can adopt two approaches: a) in many practical applications such as labeled traffic classification, we can group data points with same labels into forming sub-matrices; b) in absence of labeled datasets, we can adopt a clustering approach to decompose the traffic matrices. Specifically, we advocate to use spectral clustering due to its sound theoretical foundation and ability to handle high dimensional as well as non-linearity in the data. Other reasons of not directly applying standard clustering algorithm such as k-means is due its dependency on having clusters to form convex regions which may not be true for the data containing multiple linear sub-manifolds. In this section, we provide details of applying our version of spectral clustering on traffic matrices which requires constructing similarity matrix and graph Laplacian.

Constructing Similarity Matrix: Computing appropriate similarity matrix $\mathbf{W}$ of spectral clustering from the data matrix $\mathbf{X}$ is a crucial step and needs careful consideration. For high dimensional data, Gaussian (or heat) kernel $\mathbf{W}$ is a suitable choice, for which theoretical motivation can be found in [20].

$$\mathbf{W}_{ij} = e^{-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma_i^2}} \tag{10}$$

Gaussian kernel can be interpreted in many different ways, from kernel density estimation (KDE) to representing conditional probability $p_{j|i}$ of picking x_j as the neighbor of x_i data point. As density can be different in different regions, choosing σ appropriately in Gaussian kernel is important and can greatly affect the mapping of embedded data points in low-dimensional space. Instead of setting constant σ for all data points, we propose to compute σ_i at each data point x_i such that the entropy of distribution:

$$-\sum_j p_{j|i} \log p_{j|i} = \log k \tag{11}$$

is equal to $\log k$, where k is a user defined perplexity parameter and can be interpreted as a smooth measure of effective number of neighbors. For calculating σ_i we performed a binary search over its value – so that gives $\log k$ entropy for each data point. It turns out that similarity matrix is robust for different values of k and typical, the values lie in the range of $5 - 50$.

Constructing Graph Laplacian: Different versions of graph Laplacians exist in literature but we adopt a symmetric normalized graph Laplacian (as shown below) proposed by [21] as it is less susceptible to bad clustering when different clusters are connected with varied degree.

$$\mathbf{L} = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2} \tag{12}$$

where $\mathbf{D}$ is the diagonal degree matrix whose elements are sum of the rows of similarity matrix. From eigen decomposition of $\mathbf{L}$, d largest eigenvectors are stacked as columns in a $\mathbf{Y}$ matrix which is renormalized to yield a low-dimensional representation of data in $\mathbb{R}^d$ space. There are several ways to estimate the intrinsic dimension d of the data (e.g., kernel PCA) but graph Laplacian implicitly provides a way to estimate d through examining drop in eigenvalues of $\mathbf{L}$. A better approach of approximating intrinsic dimension can be found in [22]. For our datasets, spectral clustering approach was sufficient enough to yield faithful results. After obtaining a low dimensional data $\mathbf{Y}$, we apply traditional clustering algorithms to obtain clusters. In our paper, we have applied DBSCAN for clustering due to its robustness against outliers. Finally using cluster labels, we construct sub-matrices and approximate them with low rank via SVD. Algorithm 1 shows the complete MULTI-LOW-RANK approximation method.

Algorithm 1 MULTI-LOW-RANK Approximation Algorithm

```
Input: Traffic Matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ and $k$ perplexity;
Compute matrix $\mathbf{W}$:
    for each $1 \le i, j \le N$ do
        compute $\sigma_i$ for a given $k$ using Eq. (11);
        compute $\mathbf{W}_{ij}$ using Eq. (10);
Compute $\mathbf{D} = \sum_j \mathbf{W}_{ij}$ and $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$;
Compute $\{v_1, v_2, .., v_d\}$ as $d$ largest eigenvectors of $\mathbf{L}$ and
    stack them to form $\mathbf{Y} \in \mathbb{R}^{N \times d}$ matrix;
Normalize $\mathbf{Y}$ to have unit length rows;
Apply DBSCAN algorithm to cluster points in $\mathbf{Y}$ matrix
    and obtain clusters labels;
Output: Construct $\{\mathbf{X}_c\}_{c=1}^{C}$ sub-matrices using cluster
    labels and approximate them via PMF/SVD.
```

VII. Application of Multi-Low-Rank Approximation in Traffic Classification

To demonstrate the importance of preserving local structures while approximating the data, we perform traffic classification on the approximated data obtained via Multi-Low-Rank SVD and compare the performance with Low-Rank SVD. For this purpose, we employed the widely used Nsl-Kdd [23] as a benchmark dataset which contains 15 different types of traffic (labels). We constructed a partial data matrix $\mathbf{X} \in \mathbb{R}^{1200 \times 2058}$ from the subset of data due to memory constraints and perform one-hot encoding for each categorical feature. Next, we follow the standard procedure of performing 10-fold cross validation with LIBSVM [24] library to test the classification performance. Following methods were considered for comparison: **1**) Multi-Low-Rank SVD via traffic labels to approximate $\mathbf{X}$ with total 15 sub-matrices and setting $\mathrm{rank} = 1$ for each of them. **2**) Multi-Low-Rank SVD via spectral clustering to get labels and create sub-matrices. In this case, we obtain total 11 such sub-matrices and set $\mathrm{rank} = 1$ to approximate each of them. **3**) Low-Rank SVD method, and to make fair comparison, we set $\mathrm{rank} = 15$ for matrix approximation. **4**) Full-Rank (without any approximation of the data) method.

Method	**Full-Rank**	**Low-Rank SVD**	**Multi-Low-Rank SVD via Spectral.**	**Multi-Low-Rank via Label.**
Accuracy	95.28%	90.10%	94.60%	96.58%

TABLE I: Traffic classification accuracy on Nsl-Kdd Data.

Table I shows the traffic classification accuracy of each method. It is clear from the results that Multi-Low-Rank SVD approximation outperforms the Low-Rank SVD on traffic classification task. The results conform with the reasoning of preserving local and global structure of the data. Furthermore, Multi-Low-Rank SVD approximation via traffic labels surprisingly outperforms the Full-Rank method as well. It is due to the fact that all the local structures are well approximated by only few dominant features which boost the performance of SVM as compare to just using raw features.

VIII. Comparison and Evaluation of Multi-low Rank Approximation on real datasets

Our first real dataset ($\mathcal{RD}_1$) represents a nationwide cellular traffic extracted from a call detail record (CDR) dataset. This dataset consists millions of (voice and text) call records captured by over a 1000 base stations (or towers) spanning an entire African nation for over a month. Two features make this dataset a good candidate for our traffic matrix analysis: 1) we have information about both the origin and destination towers associated with every call thereby capturing the amount of traffic transmitted or received by towers (or even traffic volume between towers), and; 2) the dataset represents an entire nation which makes the traffic matrix rich with diverse patterns. For this particular dataset, we consider the traffic matrix to be in the form of an origin destination matrix, where origins and destinations correspond to the cellular towers. The value inside the matrix denotes the number of calls made from the corresponding origin (or row) to some destination (or column). As a result, we obtain a data matrix $\mathbf{A} \in \mathbb{R}^{1214 \times 1214}$.

In order to evaluate the Multi-Low-Rank approximation method, we use the following relative Frobenius norm difference metric for calculating matrix approximation error rate: $\|\mathbf{A} - \widehat{\mathbf{A}}\|_F^2 / (\|\mathbf{A}\|_F^2 + \|\widehat{\mathbf{A}}\|_F^2)$ in order to bounded its range in $[0, 1]$. We compare our approach with existing algorithms specifically – Bi-clustering, non-negative factorization (NMF) and Lo-Rank SVD. Bi-clustering (or Spectral Co-Clustering) allows to cluster both rows and columns simultaneously while NMF is popular for factorizing with non-negative matrix in order to have better interpretation.

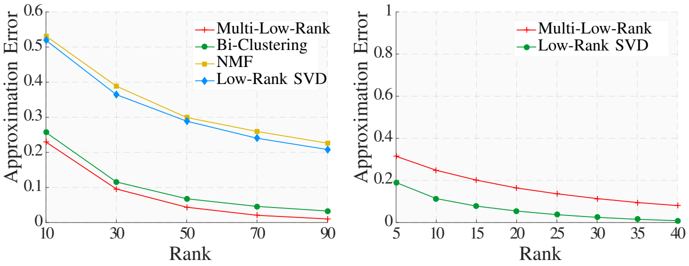
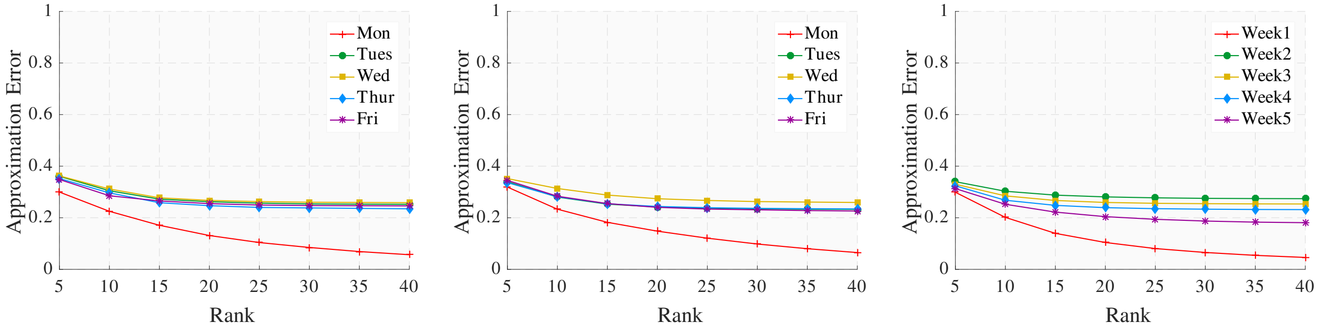


Fig. 3: (a) Comparing matrix approximation error rate. (b) Approximation error rate w.r.t. overall rank.

Figure 3a shows the matrix approximation error rate for different algorithms with respect to the rank. It is important to *distinguish the meaning* of "rank" in the context of different algorithms. For spectral and bi-clustering algorithms, rank corresponds to a sub-matrix representing a cluster. In our case, we found 16 clusters for this dataset. So here, matrix approximation error rate is the sum of approximation error of each sub-matrices. The rank for Low-Rank SVD and NMF (non-negative matrix factorization) represents the single (or global) rank of the full matrix. Figure 3b points out the fact that Low-Rank SVD or NMF requires higher rank to approximate traffic matrix to yield lower approximation error. In contrast, it would be better to approximate traffic matrix with multiple ranks corresponding to lower error rate that tend to preserve local as well global structure of the data. However, it is important to point out again that, we are no longer approximating matrix $\mathbf{A}$ via a single specified rank r matrix but rather approximating via 16 *small sub-matrices* of rank r (or as *rank* $16r$), so comparison may seem misleading. Therefore, we provide Figure 3b to be more thorough and discuss its implication later. But the point is that approximating a matrix via multiple sub-matrices can be made computationally cheaper and also be parallelize to provide a faster and better experience. Also, our proposed Multi-Low-Rank via spectral clustering approach outperforms bi-clustering algorithm in approximating the overall matrix.

We provide Figure 3b to understand how Multi-Low-Rank performs with respect to optimal solution and in order to make much more fair comparison with Low-Rank SVD. For this case, we evaluate the performance of Low-Rank SVD for



(a) Matrix prediction error rate over consecutive days in time period 10AM-1PM.

(b) Matrix prediction error rate over consecutive days in time period 8-9AM.

(c) Matrix prediction error over consecutive weeks for in time period 10AM-1PM.

Fig. 4: Variation of MULTI-LOW-RANK prediction error rate over different time domains on $\mathcal{RD}_1$ dataset.

which we set the (global) rank equal to the sum of rank of sub-matrices (sub-matrices are obtained from multi-low-rank method) and varies the individual (same) rank of sub-matrices. Figures 3b shows that approximating matrix with multiple ranks closely follows the optimal solution obtained via SVD with single rank (which is equal to the sum of multiple ranks) as we increase the rank of each sub-matrices and also tend to preserve the local structure information (due to how it is design) which is in contrast with Low-Rank SVD.

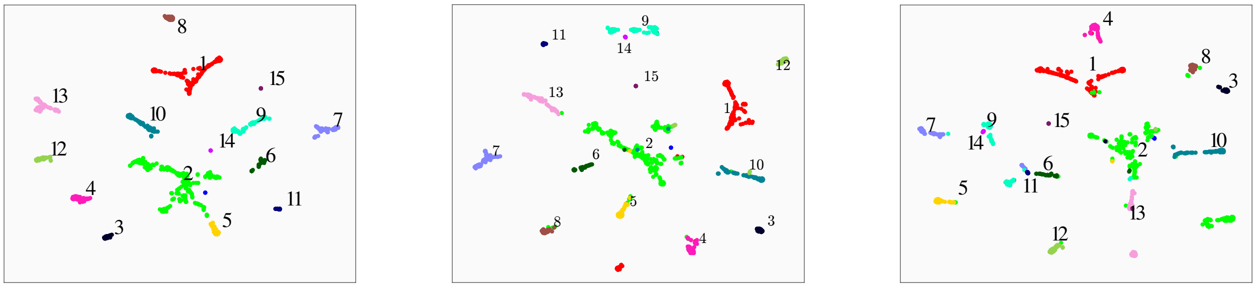
IX. MULTI-LOW RANK PREDICTION ON REAL DATASETS

In this section, we investigate the matrix prediction over different time domains by first evaluating the results over different days of a week. Here, we use the following matrix prediction error rate: $\|\mathbf{A}_i - \widehat{\mathbf{A}}_M\|_F^2 / (\|\mathbf{A}_i\|_F^2 + \|\widehat{\mathbf{A}}_M\|_F^2)$ where $\mathbf{A}_i$ is the i^{th} day matrix (e.g., Tuesday) and predicting this matrix using Monday $\mathbf{A}_M$ matrix with different ranks. Figure 4a shows that the prediction error over a week with respect to different ranks on $\mathcal{RD}_1$ dataset. We can observe that a rank in the range $10-15$ for each cluster is quite sufficient to approximate overall matrix for prediction purposes. Prediction error is also reasonable ranging around 0.25 which is close to the reference matrix. Similarly, this trend is consistent over different time periods as shown specifically for morning time in Figure 4b. Infact, the same trend can be observed around the four weeks of a month as shown in Figure 4c. The fifth week (or the first week of next month) is closest in approximation to the first week of a previous month suggesting the existence of some repetitive pattern happening each month.

We also compute the per cluster approximation error rate using the same metric (prediction error rate) as mentioned before on $\mathcal{RD}_1$ dataset. Figure 6a shows the error over different days. Few set of clusters such as $3, 5, 7$ exhibit low approximation error for all days suggesting that their behavior does not change over time domains and may be the core of traffic matrix. While some clusters show increment in approximation error suggesting that they are responsible for exhibiting certain pattern on that particular day. For instance, clusters $9, 10, 13$ on Wednesday have relatively higher approximation error compared to the other days. While on Friday most clusters have relatively low approximation error except 12 which suggests an existence of a unique pattern (may be because Friday is the last working day of a week). Similar trends are also observed over weeks as shown in Figure 6b.

To get a deeper understanding how the local structures of these MULTI-LOW-RANK matrices change over time domain, we adopt a visualization technique called t-stochastic neighbor embedding algorithm (t-SNE)[17], for projecting these data points into low-dimensional space (in $\mathbb{R}^2$ space). t-SNE projects the data in lower dimensional space by minimizing the KL-divergence between the distribution in higher dimensions and lower dimensions. Once we have data points in 2D space, we use cluster labels obtained from spectral clustering to label each points belonging to different clusters. Figure 5 shows the structure of clusters over different days for the same time period of the day. Comparing Figure 5a representing Monday with Figure 5b representing the next day, we can observe that most cluster structures remain intact (e.g., cluster labels $1, 3, 4, 6, ...$ etc.) over all the considered days. Only few clusters such as $9, 14$ seems to merge into a single cluster, thereby, dissolving a certain hidden pattern on Monday and giving rise to a new one on Tuesday. Similarly, cluster 12 seems to split into two parts where one part merges together with cluster 10 and other part remains separated. Based on Figures 5a & 5b , it is evident that the traffic matrix does not change much in the immediate future. Looking further into the time domain, specifically 2 days ahead (i.e. on Thursday), again most of cluster (or part of the cluster) structures remain intact and are stable. However, parts of some clusters break down and form new smaller clusters; while few other clusters seems to merge together to form a single new cluster. Nonetheless, the semantics behind the formation of these clusters is beyond the scope of this paper.

The second real world dataset ($\mathcal{RD}_2$) consists of (sampled) netflow records collected by a tier-1 ISP at various PoP locations in the US and Europe. For every netflow record, we have information such as the IP address, port, autonomous system number (ASN) for both the source and destination ends associated with the flow. To extract a traffic matrix from this dataset, we consider only web traffic (i.e. either source port or destination port is equal to 80) and construct an ASN to ASN matrix, where every cell in the matrix represents the number



(a) Structure of clusters on Monday.

(b) Structure of clusters on Tuesday.

(c) Structure of clusters on Thursday.

Fig. 5: Visualizing variation in local (and global) structure of traffic matrices over different time periods using t-SNE in conjunction with spectral clustering on $\mathcal{RD}_1$ dataset.

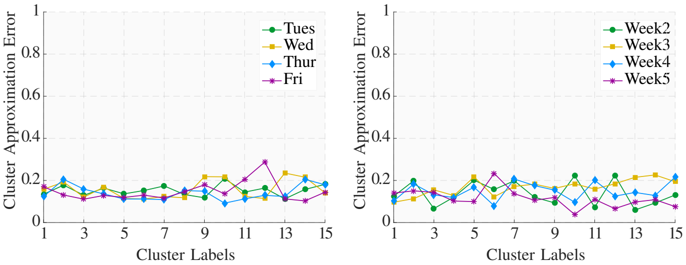


Fig. 6: (a) Relative cluster approx. error rate over days. (b) Relative cluster approx. error rate over weeks on $\mathcal{RD}_1$ dataset.

of netflow records. In other words, this matrix represents the traffic between different ASNs.

To show the efficacy of MULTI-LOW-RANK approximation on $\mathcal{RD}_2$ dataset, we predict the traffic matrices of subsequent hour. For instance, we use an ASN-to-ASN traffic matrix from 8PM to 9PM to predict that for 9PM to 10PM. Similarly, we use the ASN-to-ASN matrix capturing traffic between 9PM to 10PM to predict the same from 10PM to 11PM, and so on. Figure 7 shows the matrix approximation error for all the predictions are about $0.18-0.20$ which is quite acceptable. Such a task is especially useful to actively process streaming traffic data and predict the subsequent hour's requirement. Such a system could provide valuable insights to ISPs to dynamically provision resources according to the changing requirements as seen during different times of the day.

A. Multi-low Rank PMF Prediction on Real Datasets

In this subsection, we predict the missing entries of a matrix through MULTI-LOW RANK PMF discussed in Section V and compare the performance with Low-Rank PMF on $\mathcal{RD}_1$ and $\mathcal{RD}_2$ datasets. In our case, $\mathcal{RD}_1$ and $\mathcal{RD}_2$ contains the total of $1,473,796$ and $23,668,225$ records (or entries), respectively. For evaluation, we use 90% of records as the training data and set aside 10% of records for testing purpose. Following standard root mean square error (RMSE) is used to measure the performance on testing data: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i-\widehat{a}_i)^2} \in [0,\infty)$, where N is number of records, a_i & $\widehat{a}_i$ is actual and predicted value of the record. For fair comparison, we set number of latent features of Low-Rank PMF as the sum of latent features of each sub-matrix in MULTI-LOW-RANK PMF.

Dataset / Method	**LOW-RANK PMF**	**MULTI-LOW-RANK PMF**
$\mathcal{RD}_1$ **RMSE**	1.53	1.49
$\mathcal{RD}_2$ **RMSE**	2.34	2.11

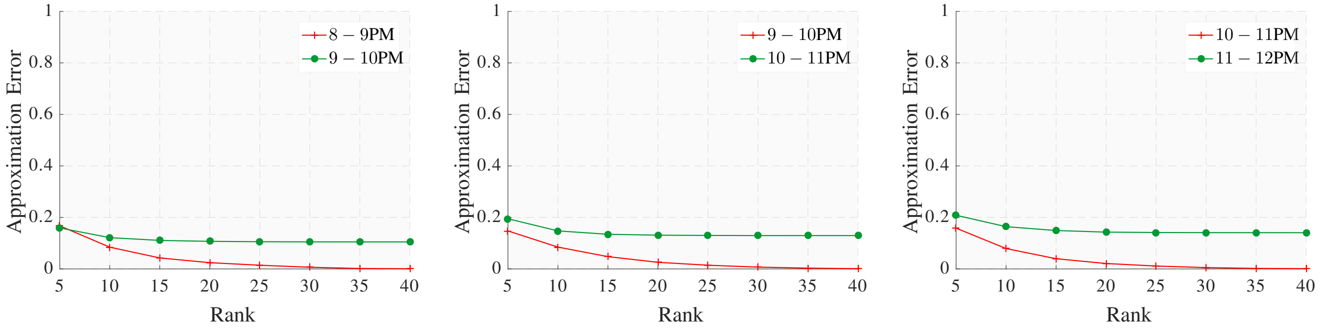
TABLE II: Prediction Error on $\mathcal{RD}_1$ & $\mathcal{RD}_2$ Datasets.

Table II shows that MULTI-LOW-RANK PMF is slightly better than predicting missing entries of a matrix than Low Rank PMF. This is due to ability of MULTI-LOW-RANK PMF to preserve local structures and also the ease in getting better local minima in optimization for smaller sub-matrices.

X. ROBUSTNESS OF MULTI-LOW-RANK APPROXIMATION

Finally, we evaluate the ability of MULTI-LOW-RANK to retrieve low rank matrices from the corrupted data with the help of proposed spectral clustering method. For this purpose, we generate a synthetic dataset as follows: Let $\mathbf{X}_i \in \mathbb{R}^{n_i \times d_i}$ be a low rank matrix of $\mathrm{rank} = r_i$ and corrupted by a Gaussian noise $\mathcal{N}(0,\sigma_i)$ with zero mean and σ_i variance, therefore, $\widetilde{\mathbf{X}_i} = \mathbf{X}_i + \mathcal{N}(0,\sigma_i)$. Let $\mathbf{D} \in \mathbb{R}^{n\times d}$ be a block diagonal matrix obtained as $D = \mathbf{diag}(\widetilde{\mathbf{X}}_1, ..., \widetilde{\mathbf{X}}_2)$, so that each $\widetilde{\mathbf{X}}_i$ lies in separate linear subspaces of high dimensional space. Finally, we obtain our noisy data matrix as: $\widetilde{\mathbf{D}} = \mathbf{P}_1\mathbf{D}\mathbf{P}_2 + \mathcal{N}(0,\sigma_D)$ where $\mathbf{P}_1$, $\mathbf{P}_2$ are random permutation matrices and σ_D is Gaussian noise variance added at the final stage of the data. We normalize each entry of $\mathbf{X}_i$ to $[0,1]$, so that the effects of noise variance is observable. Following parameters are set: $n = 800$, $d = 400$, $c = 8$, $\sigma_d = 0.5$, $r = 160$ (Low-Rank SVD) and $n_i = 100$, $d_i = 50$, $r_i = 20$, $\sigma_i = 0.5\ \forall i$.

Figure 8a shows the MULTI-LOW-RANK decomposition of the data where each cluster is obtained through proposed spectral clustering. This approach correctly retrieves all the low rank sub-matrices with appropriate $\mathrm{rank} = 20$ set in synthetic dataset. Figure 8b shows the effects of changing noise variance $\sigma = \sigma_i$ in MULTI-LOW-RANK matrix approximation. It reveals that MULTI-LOW-RANK matrix approximation error rate linearly decline with increasing noise variance upto certain level and then vary slowly with σ (here after $\sigma = 2$). Low-Rank SVD also follows the similar trend. But MULTI-LOW-RANK approximation is much better in resisting noise than Low-Rank SVD for any noise level.



(a) Prediction error for $9-10$PM hour matrix. (b) Prediction error for $10-11$PM matrix. (c) Prediction error for $11-12$PM matrix.

Fig. 7: Prediction error for next-hour matrix on $\mathcal{RD}_2$ dataset.

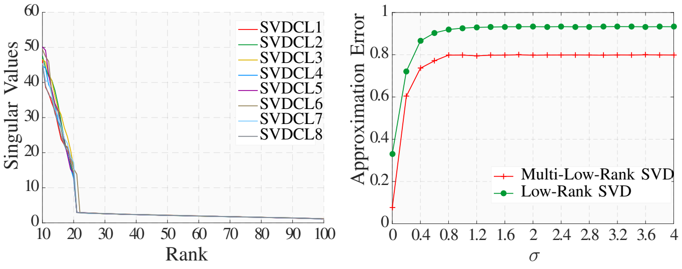


Fig. 8: (a) Multi-Low-Rank decomposition on synthetic dataset (b) Approximation Error Rate with respect to noise variance in synthetic dataset.

XI. Conclusion

In this paper, we have advanced a novel approach of approximating traffic matrices with multiple ranks as opposed to the popular single/global rank approximation approach. We established the theory behind the Multi-Low-Rank approximation and identified the conditions under which Multi-Low-Rank method is better than Low-Rank SVD in both matrix approximation and preserving the local (and global) structure of the traffic matrices. We developed an effective approach based on spectral clustering for Multi-Low-Rank matrix decomposition and approximation. Finally, with a series of experiments on synthetic and real datasets, we demonstrated that the Multi-Low-Rank approximation yields better results in traffic classification than Low-Rank SVD and can be used to predict the traffic matrices in the immediate future specially in case of streaming traffic data over different time domains and also show its robustness against noise.

Acknowledgment

This research was supported in part by DTRA grant HDTRA1-14-1-0040, DoD ARO MURI Award W911NF-12-1-0385, and NSF grants CNS-1411636.

References

[1] Y. Vardi, “Network tomography,” *Journal of the Amer. Stat. Assoc.*, 1996.

[2] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot., “Traffic matrix estimation: Existing techniques and new directions,” *ACM SIGCOMM*, 2002.

[3] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, “Fast accurate computation of large-scale ip traffic matrices from link loads.” *ACM SIGMETRICS*, 2003.

[4] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, M. C. A. Nucci, and C. Diot, “Traffic matrices: Balancing measurements, inference and modeling,” *ACM SIGMETRICS*, 2005.

[5] V. Erramilli, M. Crovella, and N. Taft., “An independent-connection model for traffic matrices.” *Internet Measurement Conference*, 2006.

[6] V. K. Adhikari, S. Jain, and Z.-L. Zhang, “From traffic matrix to routing matrix: Pop level traffic characteristics for a tier-1 ISP,” in *Proc. of ACM SIGMETRICS HotMetrics Workshop (in conjunction with ACM SIGMETRICS10 Conference)*, June 2010.

[7] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, and N. Taft, “Structural analysis of network traffic flows.” *ACM SIGMETRICS*, 2004.

[8] A. Lakhina, M. Crovella, and C. Diot, “Diagnosing network-wide traffic anomalies,” *ACM SIGCOMM*, 2004.

[9] Z. Wang, K. Hu, K. Xu, B. Yin, and X. Dong, “Structural analysis of network traffic matrix via relaxed principal component pursuit,” *Computer Networks*, 2012.

[10] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, “Network anomography,” *ACM SIGCOMM conference on Internet Measurement*, 2005.

[11] Y. Han and F. Moutarde, “Analysis of network-level traffic states using locality preservative non-negative matrix factorization,” in *Intelligent Transportation Systems*. IEEE, 2011.

[12] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu, “Spatio-temporal compressive sensing and internet traffic matrices,” in *ACM SIGCOMM Computer Communication Review*. ACM, 2009.

[13] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, “Traffic matrix estimation: Existing techniques and new directions,” *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 161–174, 2002.

[14] A. Soule, A. Lakhina, N. Taft, K. Papagiannaki, K. Salamatian, A. Nucci, M. Crovella, and C. Diot, “Traffic matrices: balancing measurements, inference and modeling,” in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1. ACM, 2005, pp. 362–373.

[15] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[16] Y. Song, M. Liu, S. Tang, and X. Mao, “Time series matrix factorization prediction of internet traffic matrices,” in *Local Computer Networks (LCN), 2012 IEEE 37th Conference on*. IEEE, 2012, pp. 284–287.

[17] L. van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research (JMLR)*, 2008.

[18] M. Holmes, A. Gray, and C. Isbell, “Fast svd for large-scale matrices,” in *Workshop on Efficient Machine Learning at NIPS*, 2007.

[19] R. Salakhutdinov and A. Mnih, “Probabilistic matrix factorization.” in *Nips*, vol. 1, no. 1, 2007, pp. 2–1.

[20] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” in *Neural Information Processing Systems (NIPS)*, 2002.

[21] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Neural Information Processing Systems (NIPS)*, 2002.

[22] L. Zelnik-Manor and P. Perona, “Self-tuning spectral clustering,” *Neural Information Processing Systems (NIPS)*, 2005.

[23] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in *Computational Intelligence for Security and Defense Applications, 2009*. IEEE, 2009.

[24] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.