INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Estimation, Reconstruction and Analysis of Internet Traffic Data using Decomposition Techniques

by

Awnish Kumar

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Department of Computer Science and Engineering

Under the supervision of

Dr. Vijaya V. Saradhi

and

Dr. T. Venkatesh

2021-02-07

Declaration of Authorship

I, Awnish Kumar, hereby confirm that:

- The work contained in this thesis is original and has been done by myself under the general supervision of my supervisors.
- This work has not been submitted to any other Institute for any degree or diploma.
- Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to the authors/researchers by citing them in the text of the thesis and giving their details in the reference.
- Whenever I have quoted from the work of others, the source is always given.

Awnish Kumar
Research Scholar,
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
awnish@iitg.ac.in, k.awnish@gmail.com

Date: 2021-02-07
Place: IIT Guwahati

Certificate

This is to certify that the thesis entitled **"Estimation, Reconstruction and Analysis of Internet Traffic Data using Decomposition Techniques"** being submitted by **Mr. Awnish Kumar** to the department of *Computer science and Engineering, Indian Institute of Technology Guwahati*, is a record of bonafide research work under my supervision and is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute.

Dr. Vijaya V. Saradhi
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
saradhi@iitg.ac.in

Dr. T. Venkatesh
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
t.venkat@iitg.ac.in

Date: 2021-02-07
Place: IIT Guwahati

Dedicated to

Late Smt. Kusheshwari Devi & Shri Singheshwar Paswan (my grandparents)

and

my loving parents & all of my teachers.

Acknowledgements

It is a great privilege for me to thank all the people whose constant support and motivation has helped me for the successful completion of my research work. First and foremost, I wish to express my deepest gratitude to my supervisors, Dr. Vijaya V. Saradhi and Dr. T. Venkatesh for their valuable advice, guidance, and motivation. It is a privilege to have had the opportunity to work under their supervision. I have been blessed to have such supervisors who allowed me to explore on my own and provided the guidance at the same time to overcome any difficulty. Their constant support and motivation not only helped in my development as a research scientist but also developed my personality, ability, and nature. Besides my supervisors, I thank the rest of my thesis committee members: Prof. S. V. Rao, Prof. S. R. Mahadeva Prasanna, and Dr. S. Ranbir Singh, for their valuable comments and suggestions which helped me to widen my research work.

I sincerely express my heartful thanks to the director, the deans and other managements of IIT Guwahati whose unified efforts have made this institute a spot for world-class studies and education. I am grateful to all the faculty members and staff of the Department of Computer Science and Engineering for extending their technical and official support for the successful completion of my research work.

I am thankful to my friends Bala, Mritunjay, Dippojwal, Satish, Sibaji, Akash, Galib, Sukarn, Shubhrendu, and Rana for their motivation and support which helped me to overcome any problem either academic or non-academic. The numerous discussions and sharing of ideas have helped in improving our research work. I am also grateful to all my seniors especially Niladri, Shirshendu, Nilkanta, Mamata, Biswajit, Mritunjay, and Shounak for their unconditional help and support. You made my life at IIT Guwahati a smooth and memorable experience.

Abstract

With the ever-increasing demand for Internet services, the knowledge of Internet traffic data is very essential for addressing various network-wide applications within an Internet Protocol (IP) backbone network. However, the high-dimensional multivariate structure of Internet traffic data proves to be a major hurdle in addressing these applications. Decomposition techniques based on Principal Component Analysis (PCA), which provides a low-dimensional representation of Internet traffic data, have garnered widespread popularity among network researchers in the last few decades. Despite widespread applicability, PCA suffers from fundamental limitations stemmed in its assumption. This paves way for the research motivated towards exploring more apt decomposition techniques. In this thesis, we primarily focus on three applications, namely estimation, reconstruction, and analysis of Internet traffic data.

1. The first contribution of this thesis is to propose the use of matrix-CUR decomposition and a multi-view learning technique based on CCA for traffic matrix estimation to avoid the overheads associated with direct measurement of Internet traffic data.

2. The second contribution of the thesis is to propose the use of matrix-CUR decomposition for the reconstruction of missing values in the traffic matrix. For the reconstruction of missing values in traffic tensor, we introduce relative-error bound tensor-CUR (TCUR-REB) decomposition. Both these techniques are computationally inexpensive and TCUR-REB alleviates the limitation of having apriori knowledge of tensor rank.

3. The third contribution of the thesis is to propose the interpretable decomposition, namely correspondence analysis and matrix-CUR decomposition, for structural analysis and volume anomaly analysis of traffic matrices. Both these techniques alleviate the limitations of PCA associated with the lack of interpretability and assumption of continuous random variables.

We have performed extensive experimentation on real and synthetic Internet traffic data to demonstrate the efficacy of the proposed techniques. Results show that these proposed techniques outperform the state-of-the-art techniques in terms of their corresponding evaluation measures.

Contents

List of Figures

List of Tables

Abbreviations

ALS	Alternating Least Squares
AR	Augmented Reality
AS	Autonomous System
BPNN	Back-Propagation Neural Network
CA	Correspondence Analysis
CCA	Canonical Correlation Analysis
CDF	Cumulative Distribution Frequency
CHIN	Chicago
CNN	Convolutional Neural Network
Col-Rand	Column Random
CP	CANDECOMP/PARAFAC
CPWOPT	CP Weighted OPTimization
CS	Compressive Sensing
DBN	Deep Belief Network
DoS	Denial-of-Service
ELRA	ELement RAndom
HOSVD	Higher-order SVD
IE	Ingress-Egress
IP	Internet Protocol
KNN	k-Nearest Neighbour
LENS	Low rank, Error, Noise, and Sparse
L.H.S.	Left-hand Side
LMaFit	Low-rank Matrix Fitting
LOSA	Los Angeles
MGM	Modulated Gravity Model
NCIM	Non-stationary Conditionally Independent Model
NMAE	Normalized Mean Absolute Error
NMF	Non-negative Matrix Factorization
NMSE	Normalized Mean Squared Error

NRMSE	Normalized Root Mean Squared Error
OD	Origin-Destination
OSPF	Open Shortest Path First
PCA	Principal Component Analysis
PCAOM	PCA based Optimization Method
PCs	Principal Components
PoP	Point of Presence
R.H.S.	Right-Hand Side
RowRand	Row Random
RPCA	Robust Principal Component Analysis
RSE	Relative Spatial Error
RTE	Relative Temporal Error
RTME	Robust Traffic Matrix Estimation
RVs	Random Variables
SDN	Software-Defined Networking
SNMP	Simple Network Management Protocol
SRMF	Sparsity Regularized Matrix Factorization
SRSVD	Sparsity Regularized SVD
STC	Sequential Tensor Completion
STTC	Spatio-Temporal Tensor Completion
SVD	Singular Value Decomposition
SVT	Singular Value Thresholding
TCUR-AEB	Additive-Error Bound Tensor-CUR
TCUR-REB	Relative-error Bound Tensor-CUR
TIRA	TIme RAndom
TM	Traffic Matrix
TT	Traffic Tensor
UDP	User Datagram Protocol
UER	Unclassified Energy Rate
UEID	Unclassified Eigenflow ID
VANETs	Vehicular Ad-hoc NETworks
VoD	Video-on-Demand
VR	Virtual Reality

Chapter 1

Introduction

In the last few decades, there has been a rapid growth in the Internet Protocol (IP) traffic due to Internet video surveillance, video streaming, Virtual Reality (VR) and Augmented Reality (AR) traffic, mobile data traffic, Consumer Video-on-Demand (VoD) traffic, Internet gaming traffic, etc. According to Cisco's prediction [1], the global IP traffic will increase more than threefold between 2017 and 2022. Mobile data traffic will increase sevenfold between 2017 and 2022 globally. With the ever-increasing demand of the Internet, the knowledge of Internet traffic data has become very essential for handling various network-wide operations and management tasks within an IP backbone network.

An IP backbone network, or Autonomous System (AS), consists of a collection of IP routers and bidirectional layer-three links [2] as shown in Fig. 1.1. The arrangement of these routers and links define the topology of the underlying network. Each network is connected to a set of customers and a set of peer networks (other autonomous systems). The routers and links that are completely internal to the network are referred as backbone routers and backbone links (shown in green color); while others are referred as edge routers and edge links (shown in blue color). The edge routers through which traffic data enters the backbone network are referred to as ingress routers and the edge routers through which traffic data leaves the network are referred to as egress routers. Note that the same router can act as ingress as well as the egress router. For example, if traffic enters the network

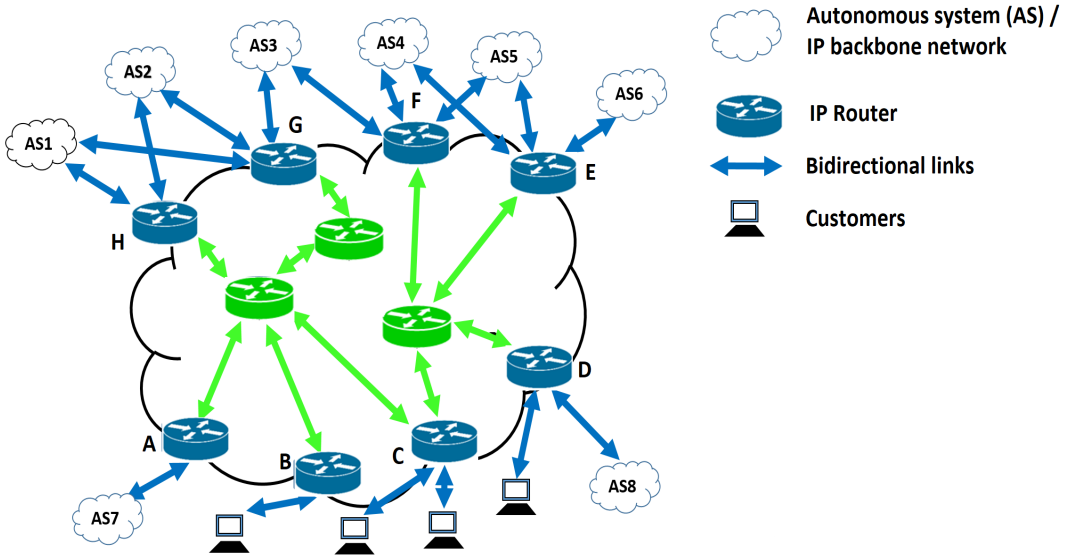


FIGURE 1.1: IP backbone network

through router A, A will be referred to as ingress router and if traffic leaves the network through router A, A will be referred to as the egress router. Traffic volume data (measured in the number of bytes or packets) enters the network through an ingress router, traverses multiple backbone links and leaves the network through an egress router. The traversal of the links is governed by the routing information. Network operators capture the volume of traffic flow across various Ingress-Egress (IE) router pairs during successive time-intervals in the form of IE Traffic Matrix (TM) or IE Traffic Tensor (TT).

An IE traffic matrix is a two-dimensional non-negative matrix with rows corresponding to time-intervals and columns denote various IE router pairs. The $(i,j)^{th}$ entry of TM denotes the volume of traffic flow during i^{th} time-interval across j^{th} IE router pair. The traffic matrix is also viewed as a set of Origin-Destination (OD) flows, where an OD flow is time-series of traffic volume flow across an IE node pair. On the other hand, an IE traffic tensor is a three-dimensional non-negative array where the first dimension, *i.e.*, the row, corresponds to ingress routers, the second dimension, *i.e.*, the column, denotes the egress routers and the third dimension, *i.e.*, the tube, corresponds to time-intervals. The $(i,j,k)^{th}$ entry of TT denotes the volume of traffic flow from i^{th} ingress router to j^{th} egress router during k^{th} time-interval. Direct measurement of Internet traffic volume data can be

performed using various monitoring technologies such as NetFlow [3], SFlow [4] and i-Perf [5]. However, direct measurements are associated with costs of communication, computational and storage overheads [6]. Obtaining an accurate traffic matrix is of value to network operators for addressing various network-wide applications such as *estimation, reconstruction, and analysis of Internet traffic data*.

Estimation problem refers to the estimation of traffic matrices using less expensive and easily available traffic information, to avoid the overhead costs associated with direct measurement.

Reconstruction problem refers to the reconstruction of missing values in the traffic matrix or traffic tensor which percolate due to multiple reasons including failure of monitoring equipment and use of unreliable transport protocol.

The *analysis* consists of two sub-problems.

- ***Structural analysis*** aims to identify the different structural patterns of the traffic matrix.
- ***Volume anomaly analysis*** aims to identify volume anomalies present in the traffic matrix, which are unexpected events.

1.1 Motivation

The major challenge in addressing the network-wide applications is that even a moderate-sized network consists of thousands of routers, resulting in a $1000 \times 1000 = 1{,}000{,}000$ IE router pairs. That is, the resulting traffic matrix will have 1 million columns which makes the analysis of the traffic matrix very cumbersome. So there is a need for a more compact representation of TMs. Thus, TMs are aggregated at the Point of Presence (PoP) level. A PoP is an artificial demarcation point representing a geographical location, which is a collection of physically co-located edge routers. The PoP level aggregation is readily accepted as it provides a simple visualization of the whole backbone network to its operators. For example, the

Abilene backbone network which is located in North America consists of 11 PoPs resulting in a total of 121 PoP pairs. That is, Abilene traffic matrices [7] consists of 121 columns with each corresponding to a PoP pair. Traffic matrices obtained from the Geant network [8] consists of 529 PoP pairs. Despite this aggregation, a moderate-sized network such as Abilene and Geant network consists of 121 and 529 PoP pairs respectively, which makes the resulting traffic matrix as a high dimensional multivariate structure. So, the central challenge in addressing various network-wide applications stems from the high dimensional multivariate structure of traffic matrices.

In practice, a common solution for dealing with high-dimensional multivariate structure is to find an alternate lower-dimensional approximation to the structure that preserves its important properties. It has often been observed that a high dimensional structure can be approximated using a smaller number of dimensions. Such an approach is termed as a dimensionality reduction technique. One can reconstruct the original high-dimensional structure from the low-dimensional representation with an associated reconstruction error. Decomposition techniques play a pivotal role in addressing the central challenge by decomposing the traffic matrix in terms of a few basis vectors. These basis vectors, in turn, provide a low-dimensional representation of traffic matrices. Thus, decomposition techniques are helpful in addressing various network-wide applications such as *estimation, reconstruction, and analysis of Internet traffic data.*

In the past two decades, Singular Value Decomposition (SVD), which forms the baseline of multiple decomposition techniques including Principal Component Analysis (PCA), has garnered popularity among network researchers for dealing with high-dimensional multi-variate structures. The basis vectors of PCA are termed as Principal Components (PCs) which are arranged in the decreasing order of importance (singular values). The projection of the traffic matrix onto the first few PCs provides a set of eigenflows which denote the low-rank approximation (or low-dimensional representation). Application of PCA and its variants have been

demonstrated extensively in literature for addressing various network-wide applications such as *traffic matrix estimation* [6], *compressive sensing* [9], *structural analysis* [10], and *volume anomaly analysis* [11].

Despite widespread usage, PCA suffers from a few fundamental limitations stemmed in its assumptions listed as follows:

1. *Assumption of continuous Random Variables (RVs)*: PCA is primarily employed for continuous random variables. In PCA, variables are assumed to be continuous random variables, while variables (PoP pairs) of TM are discrete random variables. Discreteness is of three types:

   (a) Binary data, which can take only two values such as 0 or 1.

   (b) Categorical data, which contain different categories such as days of a week.

   (c) Count data, which takes into account the number of occurrences such as the number of bytes flowing between a PoP pair during a given specified time-interval.

   PoP pairs belong to the category of count data type discrete RV. Although the variables in question are discrete RVs, PCA assumes them to be continuous RVs. Application of PCA on discrete RV by assuming them to be continuous RV results in *excessive skewness and kurtosis* [12]. In addition to this, the covariance matrix of assumed continuous RV doesn't match the true covariance matrix of the original discrete RV.

2. *Lack of interpretability*: The objective of PCA is to find directions along which maximum variance is achieved by projecting the data. The resulting directions, namely eigenflows, do not preserve meaning in terms of original dimensions (PoP pairs). It has been argued in the literature that recovering original input space from the projected space is an inherently challenging problem [13]. This is problematic when one is interested in obtaining insights from the output of matrix decomposition.

These limitations question the applicability of PCA on Internet traffic data and demands for decomposition techniques which can alleviate these limitations. Motivated by this, the present thesis proposes the use of decomposition techniques which are well suited for application on Internet traffic data and obtains improved performance for estimation, reconstruction, and analysis of Internet traffic data.

1.2 Objectives and Contributions of the Thesis

Objectives of the thesis are to perform the (i) *estimation* (ii) *reconstruction*, and (iii) *analysis* of Internet traffic data using decomposition techniques. For each of the objectives, we present the problem definition, limitations of existing techniques, and contributions of the thesis.

1. *Estimation of Internet traffic data*: Direct measurement of Internet traffic data is performed at each ingress node using NetFlow. However, using NetFlow, one obtains local TMs (that is the traffic which enters through the corresponding ingress node). These local TMs need to be shipped to a central location for appropriate processing to obtain a global TM (which contains traffic across *all* PoP pairs). The shipping cost of local TMs translates to communication overhead, while the processing cost at the central location translates to computational overhead. In addition to this, there is storage overhead at the central location as well. These overheads make the direct measurement an expensive affair. The availability of real traffic matrices becomes scarce and limited. This has motivated researchers to address the *estimation problem* which deals with the estimation of traffic matrices using easily available traffic information.

   For a network consisting of M links and P PoP pairs, the TM estimation problem [14] is posed as a system of linear equations which is as follows:

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t^{\mathsf{T}} \tag{1.1}$$

where $\mathbf{y}_t \in \mathbb{R}^{M\times 1}$ is a known vector containing volume of traffic flow across M links measured during a specific time interval t, obtained using Simple Network Management Protocol (SNMP) reports. $\mathbf{A} \in \mathbb{R}^{M\times P}$ is the routing matrix of binary values, defined such that $A(i,j) = 1$, if traffic flow across j^{th} PoP pair traverses i^{th} link, otherwise zero. $\mathbf{x}_t \in \mathbb{R}^{1\times P}$ is an unknown vector that represents the traffic volume across P PoP pairs during t^{th} time-interval. By solving Eq. (1.1), $\mathbf{x}_t$ is estimated from the known quantities $\mathbf{y}_t$ and $\mathbf{A}$. The complete TM is obtained by solving Eq. (1.1) for each time-interval. Since the number of unknowns (P) is more than the number of known parameters (M), this becomes an ill-posed inverse problem and multiple solutions exist for the unknown variables.

Traditional approaches for TM estimation assume an initial model for the TM and minimize the deviation of estimated TM from the modeled one subject to link and routing constraints given in Eq. (1.1). TM estimation techniques are classified into three generations based on the models used. The third-generation techniques, which also include the PCA-based method, have been shown to outperform the first and second-generation techniques in terms of achieving low estimation errors. The PCA-based method assumes a low-rank model for the traffic matrix. PCA [6] solves the TM estimation problem by estimating first few (*say* k) eigenflows instead of estimating P unknown variables, where k is the rank parameter. Since $k \ll P$, the problem becomes well-posed. Although PCA and other third-generation techniques have shown accuracy in estimation of traffic matrices as compared to the first two generations of techniques, they suffer from the following limitations:

(a) *Use of expensive multiple sources of information*: The reduction in error is achieved because of the use of multiple sources of information including *partial direct TM measurements* for TM estimation which is an expensive affair to obtain.

(b) *Rank parameter sensitivity of PCA-based method*: The rank parameter of the PCA-based method is a deciding factor for TM estimation and

stands for the number of eigenflows to be estimated instead of estimating traffic volume flow across all PoP pairs. These estimated eigenflows are then used to reconstruct the complete traffic matrix. However, the estimation accuracy is found to vary significantly with change in the rank parameter.

(c) *Naive use of multiple sources of information*: A recent approach [15] namely Robust Traffic Matrix Estimation (RTME) makes simultaneous use of multiple sources of information by vertically concatenating link count matrix and traffic matrix. However, it has been stated in [16] that this method of naive concatenation causes over-fitting and is not physically meaningful as the specific statistical property of each source of information is lost.

The first objective of the thesis is to alleviate these identified limitations and propose the use of two well-known decomposition techniques for traffic matrix estimation.

Contribution 1: Estimation of Internet Traffic Data

(a) Limitation (a) is addressed by the use of *multiple inexpensive estimated TMs* as input for TM estimation.

(b) Limitation (b) is addressed by employing matrix-CUR decomposition. In this decomposition, given a rank parameter k, the input TM is decomposed into three factor matrices. First factor matrix contains a small number of original columns of TM; third factor matrix contains a small number of original rows of the traffic matrix. The second matrix is computed in such a way that the product of three factor matrices is equal to the input TM. Experimentation using Abilene traffic matrices [7] suggests that the matrix-CUR decomposition is observed to exhibit less sensitivity to estimation accuracy and outperforms state-of-the-art techniques in terms of obtaining low estimation error.

(c) Limitation (c) is addressed by employing a multi-view subspace learning technique. This technique considers the *multiple estimated TMs* as

multiple *views* and simultaneously learn a common subspace shared by all the views. We make use of Canonical Correlation Analysis (CCA) for subspace learning of multiple views. Experimentation using Abilene traffic matrices suggests that CCA reduces the estimation error by more than 80% as compared to the single-view learning techniques.

2. *Reconstruction of Internet traffic data*: Internet traffic data, be it obtained through direct measurements or through solving the estimation problem, suffers from the missing value problem because of multiple reasons [17]. Missing values percolate in TM during direct measurement because flow-level measurements collected locally by network operators are incomplete and cannot form the global view of the TM. Direct measurement method is constrained by the use of legacy hardware components whose failure results in missing values in the local TMs. For complete list of causes for missing values in TM, please refer to [9]. TM estimation techniques, on the other hand, make use of known link counts obtained through Simple Network Management Protocol (SNMP) to estimate unknown TM. SNMP data itself is susceptible to error as it uses unreliable User Datagram Protocol (UDP) packets for transmission, thereby resulting in missing link measurements. Using SNMP data with missing values for estimation of TM will again result in missing values in TM. Since many network-wide applications are either intolerant or highly sensitive to missing data, addressing *the reconstruction problem* which deals with the reconstruction of missing values is of prime importance to network operators.

   Compressive sensing (CS) is a generic methodology for the reconstruction of these missing values by taking into account the presence of certain types of structures in the input data. For the reconstruction of missing values in traffic matrix or traffic tensor, CS techniques take advantage of their low-rank structure. The methodology is to obtain a low-rank approximation of the traffic matrix or traffic tensor with missing values followed by replacing the missing values with its corresponding entry from low-rank approximation.

Existing compressive sensing techniques employ (i) SVD-like factorization for traffic matrices and (ii) Higher-order SVD (HOSVD)-like factorization for traffic tensors to obtain a low-rank approximation. The output factor matrices are obtained using Alternating Least Squares (ALS) procedure which starts with random initialization of the factor matrices. ALS procedures are computationally expensive and are observed to be sensitive to the initialization procedures. That is, the ALS procedure exhibits large variation in errors in the reconstructed values for different initializations. Apart from this, tensor decomposition techniques require apriori knowledge of tensor rank. However, determining the tensor rank is an NP-hard problem [18], and there is no straightforward algorithm to determine the rank.

The second objective of the thesis is to alleviate these identified limitations and propose the use of decomposition techniques for compressive sensing of Internet traffic data which reconstructs missing values with least reconstruction error.

Contribution 2: Reconstruction of Internet Traffic Data

(a) We demonstrate the sensitivity (in terms of error in reconstruction and convergence time) of ALS-based procedures under varying initializations of factor matrices.

(b) To avoid the sensitivity of ALS-based procedures and the requirement to have apriori knowledge of tensor rank, we *introduce a novel tensor decomposition technique*, namely Relative-Error Bound Tensor-CUR (TCUR-REB) decomposition, for compressive sensing of traffic tensor. For loss rates ranging from 0% to 98%, TCUR-REB outperforms existing tensor decomposition techniques in terms of reconstruction of missing values in the Abilene [7] and Geant [8] traffic tensors.

(c) To avoid the computationally expensive ALS-based procedures for compressive sensing of traffic matrices, we employ matrix-CUR decomposition for the reconstruction of missing values in the traffic matrix. For loss rates ranging from 0% to 95%, matrix-CUR decomposition

outperforms existing matrix decomposition techniques in terms of reconstruction of missing values in the Abilene [7] and Geant [8] traffic matrices.

3. *Analysis of Internet traffic data*: Once the missing values are reconstructed, the traffic matrices are served as an input to other network-wide applications for the analysis of traffic matrices. In this thesis, we present two such analysis of traffic matrix, namely structural analysis and volume anomaly analysis. The structural analysis deals with the identification of different structural patterns present in the traffic matrix, which is important in the current scenario when there is a scarcity of publicly available datasets. To overcome this scarcity, the identified structural patterns are used in the generation of synthetic traffic matrices for carrying out various research-related activities. One such activity is volume anomaly analysis which makes use of synthetic traffic matrix generated using spikes structural patterns for ground truth information. Volume anomaly analysis deals with the identification of time-intervals and PoP pairs associated with the volume anomaly, where a volume anomaly refers to a large and sudden positive or negative change in the traffic volume flow. These changes are typically caused by unexpected events such as outages coming from network equipment failures and network attacks like Denial-of-Service (DoS) attacks. Volume anomaly analysis is crucial for network operators to correctly manage large and unexpected congestion problems at the backbone network caused by traffic volume anomalies.

   The objective of structural analysis is to decompose a traffic matrix as a sum of its constituents or structural patterns, which are later classified into different categories based on their behavior. This systematic decomposition sheds light on the intrinsic structure of the traffic matrix, and consequently on the behavior of the network as a whole. For structural analysis, PCA was exercised by Lakhina *et al.* [10]. They have considered the eigenflows as the structural patterns of the traffic matrix and expressed traffic flow across each PoP pair as a weighted sum of these structural patterns. These structural patterns are later classified into three categories namely, periodic, spikes and

noise. However, Wang *et al.* [19] demonstrated the inefficiency of the categorization in terms of completeness (each pattern has to be classified into at least one category) and orthogonality (each pattern must not be classified into more than one category at the same time). Orthogonality is determined by the number of indeterminate patterns[1] and completeness is determined by the number of non-determinate patterns[2]. They have proposed a robust variant of PCA, namely Robust Principal Component Analysis (RPCA), for structural analysis of the traffic matrix. The RPCA problem is to decompose a given matrix as a sum of a low-rank matrix (corresponding to the periodic pattern), a sparse matrix (corresponding to spikes pattern), and a random matrix (corresponding to noise pattern). However, this decomposition fails in case if the sparse matrix is also a low-rank matrix or if the low-rank matrix is also sparse. This results in confusion between the identification of periodic and spikes pattern. Although RPCA has resolved the issue concerning the completeness, however, the problem of orthogonality persists. Besides, PCA and its variants suffer from two fundamental limitations stemmed in its assumption as stated earlier.

For anomaly analysis, PCA was exercised by Lakhina *et al.* [11]. Once the principal components (PCs) are obtained using PCA, the principal component space is separated into normal and anomalous subspaces. The first k PCs comprise the normal subspace and remaining PCs comprise anomalous subspace. The PCA-subspace method identifies a time-interval as anomalous if the magnitude of projection onto the anomalous subspace exceeds an associated Q-statistic threshold. Various other authors [20, 21, 22] have used PCA-subspace method for anomaly analysis. Huang *et al.* [22] have proposed a distributed PCA-subspace method to handle the issues of scalability and communication overhead. Rubinstein *et al.* [20] analyzed poisoning and defense techniques for the PCA-subspace detector in backbone networks.

The following limitations are identified for the PCA-subspace method for volume anomaly analysis.

[1]Indeterminate patterns are those patterns which fall into more than one category.
[2]Non-determinate patterns are those patterns which do not fall into any of the categories.

(a) The obtained factors through PCA lack interpretability. As a consequence, identification of PoP pairs which triggered the volume anomaly becomes fundamentally difficult because of the inability of recovering original input space.

(b) Ringberg *et al.* [23] have demonstrated that the false positive rate is very sensitive to small differences in the number of principal components in the normal subspace.

(c) A large anomaly may inadvertently pollute the normal subspace as a result of which it goes undetected [23].

The third objective of the thesis is to alleviate the above identified limitations associated with the analysis of traffic matrix and propose the use of interpretable matrix decomposition techniques.

Contribution 3: Analysis of Internet Traffic Data

(a) *Structural analysis*: We employ Correspondence Analysis (CA) and matrix-CUR decomposition to perform structural analysis. Both of these techniques address the above-identified limitations. In addition, we show that matrix-CUR decomposition addresses the problems specific to the classification of structural patterns namely completeness and orthogonality.

(b) *Volume anomaly analysis*: We propose the use of CA and matrix-CUR decomposition for volume anomaly analysis of the traffic matrix. All the identified limitations in volume anomaly analysis are addressed by each of the employed techniques. Experimentation using Abilene [7] and synthetic [24] traffic matrices shows that matrix-CUR and CA not only identifies the time-interval associated with a volume anomaly but also the PoP pairs which triggered the volume anomaly.

1.3 Organisation of the Thesis

The rest of this thesis is organized as follows. Chapter 2 presents notations and background on matrix and tensor decomposition techniques. In Chapter 3, we present the first contribution of the thesis which proposes the use of matrix-CUR decomposition and multi-view subspace learning technique for *estimation of Internet traffic data*. In Chapter 4, we discuss the second contribution of the thesis, *i.e.*, to propose the use of matrix-CUR decomposition and relative error-bound tensor CUR decomposition for *reconstruction of Internet traffic data*. The third contribution of the thesis which proposes the interpretable decomposition technique, namely correspondence analysis and matrix-CUR decomposition, for *analysis of Internet traffic data* is presented in Chapter 5. Chapter 6 presents conclusions.

Chapter 2

Matrix and Tensor Decomposition Techniques

This chapter presents the notations and the decomposition techniques related to the thesis.

2.1 Notations

Scalars are denoted by non-bold small letters/capital letters with two indices/calligraphic letters with three indices such as x or $X(i,j)$ or $\mathcal{X}(i,j,k)$. Vectors are denoted by small bold letters, *say* $\mathbf{x}$. Matrices are denoted using bold capital letters, *say* $\mathbf{X}$. Higher-order tensors, *i.e.*, tensors of order three or higher, are denoted by bold calligraphic letters, *say* $\boldsymbol{\mathcal{X}}$. A TM with P PoP pairs and T time-intervals is represented as $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}\ \mathbf{x}_2^\mathsf{T}\ \cdots\ \mathbf{x}_T^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{T\times P}$, where $\mathbf{x}_i \in \mathbb{R}^{1\times P}$ is the traffic volume flow across P PoP pairs during i^{th} time-interval and $(.)^\mathsf{T}$ is the transpose of a matrix/vector. The corresponding traffic tensor is represented using bold calligraphic letter $\boldsymbol{\mathcal{X}}$. $\mathbf{Y} = [\mathbf{y}_1\ \mathbf{y}_2\ \cdots\ \mathbf{y}_T] \in \mathbb{R}^{M\times T}$ denote link count matrix of M links obtained using SNMP reports where $\mathbf{y}_i \in \mathbb{R}^{M\times 1}$ denote the volume of traffic flow across M links during i^{th} time-interval. $\mathbf{A} \in \mathbb{R}^{M\times P}$ is the routing matrix of binary values, defined such that $A(i,j) = 1$, if traffic flow across j^{th} PoP pair

traverses i^{th} link, otherwise zero. $\mathbf{1}$ denote an identity vector of size either $(T \times 1)$ or $(P \times 1)$ depending on the context. $\|.\|_F$ denote the Frobenius norm.

2.2 Principal Component Analysis

Principal Component Analysis (PCA) [25] is the most popular matrix decomposition technique which provides a low dimensional representation with the aim to retain most of the variation present in the data. PCA was originally introduced by Karl Pearson [26] in 1901 and was developed independently by Harold Hotelling [27] in the 1933. PCA has been used as a dimensionality reduction technique in various fields including image processing, genomic analysis, and information retrieval. In [28], authors have performed k-means clustering using PCA on DNA gene expression and Internet newsgroups. In the context of Internet traffic data, PCA has been employed for estimation [6], reconstruction [9], and analysis [10, 11] of Internet traffic data.

PCA can also be viewed as a co-ordinate transformation method that maps the high dimensional data onto a fewer set of axes, known as principal components. These principal components (PCs) are orthogonal, uncorrelated and points in the direction of maximum variation remaining in the data. Given a matrix $\mathbf{X} \in \mathbb{R}^{T \times P}$, the set of vectors $\{\mathbf{v}_i\}_{i=1}^{P}$, represent PCs where $\mathbf{v}_i$ denotes ith PC and are computed as follows:

$$\mathbf{v}_1 = \underset{|\mathbf{v}|=1}{\operatorname{argmax}} \|\mathbf{X}\mathbf{v}\|$$

$$\vdots$$

$$\mathbf{v}_P = \underset{\substack{|\mathbf{v}|=1 \\ \mathbf{v} \perp \mathbf{v}_1,\ldots,\mathbf{v}_{P-1}}}{\operatorname{argmax}} \|\mathbf{X}\mathbf{v}\|$$

These PCs are arranged in such a way that the first PC $\mathbf{v}_1$ captures most of the variation present in the data, the second PC $\mathbf{v}_2$ captures the second most

variation in the data and so on. It has been observed that computing the PCs is equivalent to solving the eigenvalue problem for the covariance matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. PCA is conducted by performing Singular Value Decomposition (SVD) on $\mathbf{X}$. Given a matrix $\mathbf{X} \in \mathbb{R}^{T \times P}$, SVD decomposes $\mathbf{X}$ into a product of three matrices as follows:

$$\mathbf{X} = \mathbf{E}\mathbf{S}\mathbf{V}^\mathsf{T} \tag{2.1}$$

where the columns of $\mathbf{E} \in \mathbb{R}^{T \times P}$, *i.e.*, $\{\mathbf{e}_i\}_{i=1}^{P}$, denote the set of left singular vectors (or the set of eigenflows) and are the eigenvectors of $\mathbf{X}\mathbf{X}^\mathsf{T}$. The columns of $\mathbf{V} \in \mathbb{R}^{P \times P}$, *i.e.*, $\{\mathbf{v}_i\}_{i=1}^{P}$, denote the set of right singular vectors (or the set of PCs) and are the eigenvectors of $\mathbf{X}^\mathsf{T}\mathbf{X}$. $\mathbf{S} \in \mathbb{R}^{P \times P}$ is a diagonal matrix containing singular values $\{\sigma_i\}_{i=1}^{P}$. The diagonal elements in this matrix are arranged in the descending order such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_P$.

$\mathbf{E}$, $\mathbf{S}$, and $\mathbf{V}$ are also termed as output factor matrices of SVD. Using these output factor matrices, one can obtain the following:

$$\hat{\mathbf{X}} = \sum_{i=1}^{P} (\sigma_i \mathbf{e}_i)\mathbf{v}_i^\mathsf{T} = \mathbf{X} \tag{2.2}$$

where $\hat{\mathbf{X}}$ is a reconstructed matrix obtained using the factors $\mathbf{E}$, $\mathbf{S}$, and $\mathbf{V}$ respectively. Note that the reconstructed matrix $\hat{\mathbf{X}}$ will be identical to the original matrix $\mathbf{X}$ as all the P PCs are used in obtaining $\hat{\mathbf{X}}$. As most of the variation present in the data is captured by the first few PCs (*say* k where $k \ll P$), using these k PCs alone in reconstructing $\hat{\mathbf{X}}$ will result in rank-k approximation of the original matrix $\mathbf{X}$ as given in Eq. (2.3).

$$\mathbf{X}_k = \sum_{i=1}^{k} \sigma_i \mathbf{e}_i \mathbf{v}_i^T \approx \mathbf{X}. \tag{2.3}$$

$\mathbf{X}_k$ is rank-k approximation of $\mathbf{X}$.

2.3 Correspondence Analysis

Objective of PCA is to find a low-dimensional subspace such that every principal direction captures maximum variance (energy) present in the data when the data is projected on to the corresponding principal direction. The objective of *maximizing variance* in the case of continuous random variables leads to obtaining eigenvectors and eigenvalues of *covariance matrix*. In PCA, the data associated with every variable is assumed to be continuous and randomly drawn from the multivariate normal distribution. Sample PCA too follows this assumption. When the nature of random variables are discrete, this assumption has a direct impact on the covariance matrix computation which is subject to matrix decomposition. Violation of this fundamental assumption results in excessive skewness and kurtosis of the obtained factor matrices [12]. Therefore, the use of traditional PCA results in obtaining PCs that deviate from the actual PCs in case of discrete random variable. Correspondence analysis (CA) is employed to achieve dimensionality reduction when the random variables are discrete.

Throughout its history, this method has been referred by multiple names including contingency table analysis [29], RQ-technique [30], reciprocal averaging [31], correspondence analysis [32], reciprocal ordering [33], dual scaling [34], and homogeneity analysis [35]. The theory of CA dates back to 1935, originated by Hirschfeld [36]. Fisher [29] provided an application of CA on a dataset involving two discrete random variables namely hair color (taking values fair, red, medium, dark, and black) and eye color (taking values blue, light, medium, and dark) of 5387 children. The objective is to obtain the correlation between these two random variables in low dimensional (two dimensional) space. Bansard *et al.* [37] applied CA for understanding recent trends in the research over the timeline among bioinformatics and medical informatics fields. They have used the papers published during the year 2000 to the year 2005. These papers were arranged in the form of (document $\times$ word) matrix where the entries denote the frequency of occurrence of a particular word in a particular document. Kurt *et al.* [38] applied CA to microarray data to study the relationship between genes and hybridizations.

The microarray data is represented as a table with rows corresponding to genes and columns corresponding to hybridizations. They have demonstrated the application of CA on well-known Saccharomyces cerevisiae cell-cycle synchronization data [39].

Let n be the sum of the elements of the matrix $\mathbf{X}$. Every element of the matrix $\mathbf{X}$ is divided with n to obtain correspondence matrix $\mathbf{P}$. An $(i,j)^{th}$ element of the correspondence matrix $\mathbf{P}$ denotes the joint probability of row event i and column event j occurring simultaneously. When two events e_1 and e_2 are *independent*, then the joint probability model is written as: $P(e_1 \cap e_2) = P(e_1) \times P(e_2)$. That is $P_{ij} = P_{i\bullet} \times P_{\bullet j}$; where P_{ij} denote the joint probability of occurrence of i^{th} row variable event and j^{th} column variable event, $P_{i\bullet}$ denote the probability of occurrence of i^{th} row event, and $P_{\bullet j}$ denote the probability of occurrence of j^{th} column event.

When the total independence gets deviated, the model is re-written as: $P_{ij} = \alpha_{ij} \times P_{i\bullet} \times P_{\bullet j}$, where α_{ij} denotes the amount of deviation. If $\alpha_{ij} = 1$, then row event i and column event j are independent. When row features have a certain relation with respect to column features, α_{ij} takes a value less than 1. For every row event and for every column event, $(i,j)^{th}$ entry of the matrix $\boldsymbol{\alpha}$ is given by:

$$\alpha_{ij} = \frac{P_{ij}}{P_{i\bullet} \times P_{\bullet j}} \tag{2.4}$$

Eq. (2.4) is well known as Pearson's ratio. Pearson's Chi-square statistic is then expressed as:

$$\mathcal{X}^2 = n \times \sum_i \sum_j P_{i\bullet} \times P_{\bullet j} (\alpha_{ij} - 1)^2 \tag{2.5}$$

Eq. (2.5) assumes smaller value when $\alpha_{ij} \to 1$; which indicates that row features are independent of column features. Higher the value of this quantity, the stronger is the relationship between row and column features.

Note however that both row features and column features are in high dimensional

space. CA [40], a multivariate statistical technique, is used to obtain a low dimensional subspace that contains the row features and column features. In order to obtain the low dimensional space, CA minimizes the sum of the squared χ^2-distance (a weighted Euclidean distance given as $\chi^2 = \sum \frac{(observed-expected)^2}{expected}$) from the subspace to each of the profile points. To obtain the solution to this minimization, SVD [41] is applied on the normalized and centered matrix $\mathbf{X}$ to obtain the principal components of row features ($\mathbf{F}$) and principal components of column features ($\mathbf{G}$). The complete CA algorithm is presented in Algorithm 1.

ALGORITHM 1: Correspondence Analysis

Require: A matrix $\mathbf{X}$ of size $T \times P$

1: Matrix sum: $n = \sum_{i=1}^{T} \sum_{j=1}^{P} X(i,j)$
2: Correspondence matrix: $\mathbf{P} = \frac{1}{n}\mathbf{N}$
3: Row masses: $\mathbf{r} = \mathbf{P1}$
4: Diagonal matrix of row masses: $\mathbf{D}_r = diag(\mathbf{r})$
5: Column masses: $\mathbf{c} = \mathbf{P}^\mathsf{T}\mathbf{1}$
6: Diagonal matrix of column masses: $\mathbf{D}_c = diag(\mathbf{c})$
7: Standardized residuals: $\mathbf{S}_r = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{rc}^\mathsf{T})\mathbf{D}_c^{-\frac{1}{2}}$
8: Singular value decomposition: $\mathbf{S}_r = \mathbf{J}\Sigma\mathbf{K}^\mathsf{T}$
9: Standard coordinates of rows: $\Phi = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{J}$
10: Standard coordinates of columns: $\Gamma = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{K}$
11: Principal coordinates of rows: $\mathbf{F} = \Phi\Sigma$
12: Principal coordinates of columns: $\mathbf{G} = \Gamma\Sigma$
13: **return** $\mathbf{F}$ and $\mathbf{G}$.

In order to visually interpret the relationship between row features and column features, the first two principal directions of $\mathbf{F}$ and $\mathbf{G}$ are obtained from algorithm 1 and plotted on a two-dimensional plot as vectors. This method is well known as biplot [42]. In the biplot, the row and column vectors having the highest inner product values are the closest vectors.

2.4 Matrix-CUR Decomposition

Although PCA decomposes an input matrix as a product of three orthogonal matrices, the decomposed matrices *lack interpretability* in terms of the original dimensions of the input matrix. This is problematic when one is interested in

obtaining insights from the output of matrix decomposition. For example, each eigenflow (left singular vector) of traffic matrix derived using PCA is a weighted linear combination of all OD flows and hence can't be interpreted in terms of original PoP pairs. When applications demand *interpretability*, the use of PCA is of limited scope. In this context, decomposition of the original matrix which retains interpretability is of prominence. Matrix-CUR decomposition is employed to achieve the interpretability using the output factor matrices.

Matrix-CUR decomposition [13] is a low-rank matrix decomposition technique that decomposes an input matrix in terms of a small number of *original* columns and *original* rows of the input matrix. Specifically, matrix-CUR decomposes input data matrix $\mathbf{X}$ as a product of three matrices: $\mathbf{C}$, $\mathbf{U}$, and $\mathbf{R}$, such that $\mathbf{C}$ comprises of a small number c of original columns of $\mathbf{X}$, $\mathbf{R}$ comprises of a small number r of original rows of $\mathbf{X}$, and $\mathbf{U}$ is a small precisely computed matrix such that the product of matrices $\mathbf{C}$, $\mathbf{U}$, and $\mathbf{R}$ provides an approximation of the input matrix. That is, $\mathbf{X} \approx \mathbf{C} \times \mathbf{U} \times \mathbf{R}$. Note that the matrices $\mathbf{C}$ and $\mathbf{R}$ can replace left and right singular vectors of SVD [13] retaining the *interpretability* in terms of the original columns and rows of the input matrix.

Matrix-CUR decomposition has been applied in diverse domains which includes Internet term-document data analysis [13], genetics [13] and Intelligent transportation system [43]. Mahoney *et al.* [13] employed matrix-CUR decomposition on a (document $\times$ term) matrix of size (139 $\times$ 15170) for the classification of documents based on two topics. They have also classified patients according to three cancer types using a (patient $\times$ genes) matrix of size (31 $\times$ 5520). In [43], matrix-CUR decomposition has been employed for compression and compressive sensing of transportation data matrix of size (8928 $\times$ 6024). The rows of the matrix denote time-intervals, the columns denote road segments and the entries denote the volume of traffic flow. The compression problem deals with the representation of the input matrix in the form of a few time-intervals and few road segments. The compressive sensing problem deals with the reconstruction of missing values in the input matrix using a limited number of time-intervals and road segments.

Decomposing the matrix $\mathbf{X}$ using SVD is known to take polynomial time $O(T^2P)$ or $O(P^2T)$. Whenever T and P are large, the decomposition will be prohibitive in practice. To overcome this issue, Frieze *et al.* [44] have proposed a method of obtaining a low rank approximation of $\mathbf{X}$ which is qualitatively faster. Frieze *et al.* [44] have sampled columns of input matrix $\mathbf{X}$ based on the Euclidean norms to form the matrix $\mathbf{C}$ such that $\mathbf{CC}^{+}\mathbf{X}$ is approximately equal to $\mathbf{X}$, where $(.)^{+}$ denote Moore-Penrose generalized inverse [45]. When the number of selected columns is polynomial in k and $1/\epsilon$ where k and ϵ are rank and error parameter respectively, then using Euclidean norm selection criteria one can obtain the following worst-case additive error guarantees with high probability.

$$\|\mathbf{X} - \mathbf{P}_c\mathbf{X}\|_F \leq \|\mathbf{X} - \mathbf{X}_k\|_F + \epsilon\|\mathbf{X}_k\|_F \tag{2.6}$$

where $\mathbf{P}_c\mathbf{X}$ denote the projection of $\mathbf{X}$ onto the subspace spanned by columns of $\mathbf{C}$ and $\mathbf{P}_c = \mathbf{CC}^{+}$. $\mathbf{X}_k$ is best rank-k approximation of $\mathbf{X}$ obtained using PCA as shown in (2.3).

In addition to selecting the original columns of the matrix $\mathbf{X}$, [46] have proposed to select rows simultaneously to construct two matrices $\mathbf{C}$ and $\mathbf{R}$ using 2 passes of (2.6). In the first pass, matrix $\mathbf{C}$ is obtained using $\mathbf{X}$. In the second pass, matrix $\mathbf{R}$ is obtained using $\mathbf{X}^T$. The matrix $\mathbf{U}$ is constructed such that the following additive-error bound holds with high probability.

$$\|\mathbf{X} - \mathbf{CUR}\|_F \leq \|\mathbf{X} - \mathbf{X}_k\|_F + \epsilon\|\mathbf{X}_k\|_F \tag{2.7}$$

Bounds of this form are generally weaker because the additive error term is usually large in practice. It is suitable either for large-sized data or for a small percentage of missing data or when limited resources are available [47]. However, if the input matrix is exactly low rank, the left-hand side will become zero and the first term on the right-hand side will also be zero and the additive-error term will remain non-zero. Thus, additive-error bound fails to reflect the case of perfect reconstruction in case the input matrix is exactly low-rank.

In order to address the above issue with the additive error bound which is due to the column selection strategy, Drineas *et al.* [48] proposed a new column selection strategy based on *normalized statistical leverage scores* (2.8) to obtain $\mathbf{C}$ given the input matrix $\mathbf{X}$.

$$\pi_i = \frac{1}{k}\sum_{j=1}^{k}(v_{ij})^2 \tag{2.8}$$

where π_i denote the normalized statistical leverage score of i^{th} column of $\mathbf{X}$. Algorithm 2 presented below constructs columns of $\mathbf{C}$ based on *normalized statistical leverage scores* given (i) the input matrix $\mathbf{X}$, (ii) rank parameter k, (iii) number of columns c' that one wish to select from $\mathbf{X}$, and (iv) an error parameter ϵ.

ALGORITHM 2: COLUMNSELECT($\mathbf{A}$,k,c',ϵ)

Input: A matrix $\mathbf{X} \in \mathbb{R}^{T\times P}$, rank parameter k, number of columns c' that one wish to select from $\mathbf{X}$, and error parameter ϵ.
Output: $\mathbf{C} \in \mathbb{R}^{T\times c}$ with c columns from $\mathbf{X}$ s.t. $\mathbb{E}[c] \leq c'$.

```
C = { } //empty set of columns
Compute top k right singular vectors {v_i}_{i=1}^k of X using (2.1).
for i ← 1 to P do
    compute normalized statistical leverage score using Eq. (2.8)
    probabilities p_i = min{1, c'π_i}, for c' = O(klogk/ε^2).
    generate a random number n between 0 and 1.
    if n <= p_i then
        C ← C ∪ X(:, i)
return the set of columns C as matrix
```

It was shown in [48] that selection of columns using statistical leverage scores satisfies

$$\|\mathbf{X} - \mathbf{P}_c\mathbf{X}\|_F \leq (1+\epsilon/2)\|\mathbf{X} - \mathbf{X}_k\|_F \tag{2.9}$$

with probability at least 0.99, where $\mathbf{P}_c = \mathbf{C}\mathbf{C}^+$ denotes a projection matrix onto the column space of $\mathbf{C}$. The above algorithm is limited to the selection of columns alone and obtains decomposition as $\mathbf{X} \approx \mathbf{P}_c\mathbf{X}$.

In addition to selecting the original columns of the matrix $\mathbf{X}$ based on *normalized statistical leverage scores*, Mahoney *et al.* [13] have proposed to select rows simultaneously to construct two matrices $\mathbf{C}$ and $\mathbf{R}$ using 2 passes of (2.9). In the

first pass, matrix $\mathbf{C}$ is obtained using $\mathbf{X}$. In the second pass, matrix $\mathbf{R}$ is obtained using $\mathbf{X}^T$.

Given $\mathbf{X}$ and a rank parameter k, construct 3 matrices: $\mathbf{C}$, $\mathbf{U}$, and $\mathbf{R}$, such that $\mathbf{C}$ comprises of a small number c of original columns of $\mathbf{X}$, $\mathbf{R}$ comprises of a small number r of original rows of $\mathbf{X}$, and $\mathbf{U} = \mathbf{C}^{+}\mathbf{X}\mathbf{R}^{+}$, where $^{+}$ denotes a Moore-Penrose generalized inverse [45] of the matrix. The product $\mathbf{CUR}$ guarantees the error bound of the following form

$$\|\mathbf{X} - \mathbf{CUR}\|_F^2 \leq (2+\epsilon)\|\mathbf{X} - \mathbf{X}_k\|_F^2. \tag{2.10}$$

Following are the advantages of matrix-CUR decomposition over PCA:

1. Matrix-CUR decomposition achieves an improved relative-error bound (as given in Eq. (2.10)), which states that the Frobenius norm of the error in reconstruction using matrix-CUR decomposition is always less than or equal to (2+ϵ) times the error incurred in reconstruction using PCA from its rank-k approximation.
2. The construction of $\mathbf{C}$ and $\mathbf{R}$ involves truncated SVD, *i.e.*, computation of top k right singular vectors, which has a time complexity of O(TPk), which is an improvement over PCA which involves full SVD with the complexity of O(min{$\mathrm{T}^2\mathrm{P}$, TP^2}).
3. The matrix $\mathbf{C}$ and $\mathbf{R}$ can replace left and right singular vectors, with an advantage of interpretability in terms of the original columns and rows of $\mathbf{X}$ because they are a subset of the original input matrix.
4. There is no distributional assumption involved in the selection of $\mathbf{C}$ and $\mathbf{R}$.

2.5 Canonical Correlation Analysis

Given *one set of data* consisting of P continuous random variables, PCA obtains P directions which have maximum variance among these random variables. In

Canonical Correlation Analysis (CCA), given *two sets of data* with each set having P continuous random variables, the objective is to obtain directions which have maximum *covariance* between these two sets of random variables. CCA [49], a well-known statistical method that captures the linear correlation between two sets of multivariate random variables, is one of the most representative techniques and has been widely used for many problems, such as classification, retrieval, regression, and clustering.

We explain the working of CCA using two sets $\mathbf{X}_1 \in \mathbb{R}^{T\times P}$ and $\mathbf{X}_2 \in \mathbb{R}^{T\times P}$ as input, with each set having P random variables. The objective of CCA is to maximize correlation between all the paired data points from the two sets so that they are unified in the correlated space. For this, we linearly project the data points onto an intermediate subspace that maximizes the correlation between them using CCA [49]. Mathematically, this is expressed as

$$\rho = \max_{\mathbf{w}_1,\mathbf{w}_2} \frac{\text{covariance}((\mathbf{w}_1)^\mathsf{T}\mathbf{X}_1, (\mathbf{w}_2)^\mathsf{T}\mathbf{X}_2)}{\sqrt{\text{variance}((\mathbf{w}_1)^\mathsf{T}\mathbf{X}_1)\text{variance}((\mathbf{w}_2)^\mathsf{T}\mathbf{X}_2)}} \tag{2.11}$$

where $\mathbf{w}_1$ and $\mathbf{w}_2$ are the linear projection directions for $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. As ρ is invariant to scaling of $\mathbf{w}_1$ and $\mathbf{w}_2$, we constrain the denominator in Eq. (2.11) to unity, to result in the following alternate optimization.

$$\begin{aligned} \rho = \max_{\mathbf{w}_1,\mathbf{w}_2} \ & \text{covariance}((\mathbf{w}_1)^\mathsf{T}\mathbf{X}_1, (\mathbf{w}_2)^\mathsf{T}\mathbf{X}_2) \\ \text{such that } \ & \text{variance}((\mathbf{w}_1)^\mathsf{T}\mathbf{X}_1) = 1, \ \ \text{variance}((\mathbf{w}_2)^\mathsf{T}\mathbf{X}_2) = 1. \end{aligned} \tag{2.12}$$

We can reformulate the problem as follows:

$$\begin{aligned} \rho = \max_{\mathbf{w}_1,\mathbf{w}_2} \ & (\mathbf{w}_1)^\mathsf{T}\Sigma_{12}\mathbf{w}_2 \\ \text{such that } \ & (\mathbf{w}_1)^\mathsf{T}\Sigma_{11}\mathbf{w}_1 = 1, \ \ (\mathbf{w}_2)^\mathsf{T}\Sigma_{22}\mathbf{w}_2 = 1 \end{aligned} \tag{2.13}$$

where Σ_{12} is the covariance between $\mathbf{X}_1$ and $\mathbf{X}_2$, Σ_{11} is the variance of $\mathbf{X}_1$, and Σ_{22} is the variance of $\mathbf{X}_2$.

Eq. (2.13) can be revised using the Lagrangian L as

$$L = (\mathbf{w}_1)^\mathsf{T}\Sigma_{12}\mathbf{w}_2 + \lambda_1((\mathbf{w}_1)^\mathsf{T}\Sigma_{11}\mathbf{w}_1 - 1) + \lambda_2((\mathbf{w}_2)^\mathsf{T}\Sigma_{22}\mathbf{w}_2 - 1)$$

where λ_1 and λ_2 are the Lagrangian multipliers. The optimization problem with respect to $\mathbf{w}_1$ and $\mathbf{w}_2$ results in the following generalized eigenvalue problem [50]

$$\begin{bmatrix} 0 & \Sigma_{12} \\ \Sigma_{21} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \lambda \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \tag{2.14}$$

where $\lambda = 2\lambda_1 = 2\lambda_2$ and $\Sigma_{21} = (\Sigma_{12})^\mathsf{T}$. The solution to Eq. (2.14), $\mathbf{w}_1$ are the eigenvectors of $(\Sigma_{11})^{-1/2}\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}(\Sigma_{11})^{-1/2}$ and $\mathbf{w}_2$ are the eigenvectors of $(\Sigma_{22})^{-1/2}\Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}(\Sigma_{22})^{-1/2}$ [50]. That is, Eq. (2.14) is repeatedly solved to obtain multiple directions for transforming the data. The number of such directions is bounded by the dimensions of $\mathbf{X}_1$ and $\mathbf{X}_2$. These directions are obtained by solving for $\mathbf{w}_1$'s and $\mathbf{w}_2$'s in every dimension. Let $\mathbf{W}_1 = [\mathbf{w}_{11}|\mathbf{w}_{12}|\cdots|\mathbf{w}_{1P}]$ be the projection matrix for $\mathbf{X}_1$ and $\mathbf{W}_2 = [\mathbf{w}_{21}|\mathbf{w}_{22}|\cdots|\mathbf{w}_{2P}]$ be the projection matrix for $\mathbf{X}_2$. Then $\hat{\mathbf{X}}_1 = \mathbf{X}_1 \times \mathbf{W}_1$ and $\hat{\mathbf{X}}_2 = \mathbf{X}_2 \times \mathbf{W}_2$ denote the projection of $\mathbf{X}_1$ and $\mathbf{X}_2$ onto the intermediate space respectively, such that the data points of $\mathbf{X}_1$ and $\mathbf{X}_2$ are maximally correlated in the intermediate (or common) subspace learnt using multiple views.

2.6 Tensor Decomposition Techniques

Tensor decomposition techniques have gained widespread attention over the last decade for various applications including recommender systems [51], image analysis [52], computer vision [53], and intelligent transportation systems [54, 55]. More recently, tensors of various dimensions have been utilized to represent Internet traffic data for anomaly detection [56], [57] and compressive sensing [58]. In this section, we present important definitions related to tensor decomposition techniques.

Definition 1. The **order** or **mode** of a tensor corresponds to the number of dimensions of the tensor.

Scalars are tensors of order zero. Vectors are tensors of order one. Matrices are tensors of order two. Higher-order tensors are tensors of order three or higher.

Definition 2. **Fibers** are one-dimensional sections of a tensor, determined by varying one index while fixing the others.

Columns and rows of matrices are considered as mode-1 and mode-2 fibers, respectively. Third-order tensors have columns, rows, and tubes, as mode-1, mode-2, and mode-3 fibers. Please refer to the tutorial material in [59] for further details.

Definition 3. **Slices** are two-dimensional sections of a tensor, determined by varying two indices while fixing the others.

A third-order tensor can be sliced in three distinct ways; horizontal, lateral, and frontal slices [59].

Definition 4. **Matricization** refers to the transformation of a tensor into a matrix.

The mode-n matricization of $\boldsymbol{\mathcal{X}}$, represented by $\mathbf{X}_{[n]}$, is defined as the arrangement of the mode-n fibers of a tensor as the columns of the $\mathbf{X}_{[n]}$. For details, please see [59].

Definition 5. The **n-mode product** (denoted by $\times_n$) of a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ with a matrix $\mathbf{U} \in \mathbb{R}^{J \times I_N}$ is defined as

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}} \times_n \mathbf{U} \Leftrightarrow \mathbf{Y}_{[n]} = \mathbf{U}\mathbf{X}_{[n]} \tag{2.15}$$

Definition 6. The **outer product** (denoted by $\circ$) of two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$\mathbf{a} \circ \mathbf{b} = \mathbf{a}\mathbf{b}^{\mathsf{T}} \tag{2.16}$$

Definition 7. **Rank-one tensors**: An N-dimensional tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is of rank one if it can be expressed as the outer product of N vectors, in the form of

$$\boldsymbol{\mathcal{X}} = \mathbf{a}_1 \circ \mathbf{a}_2 \circ ... \circ \mathbf{a}_N$$

Definition 8. **CANDECOMP/PARAFAC (CP) decomposition** decomposes a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ as a sum of k rank-one tensors as

$$\boldsymbol{\mathcal{X}} \approx \sum_{r=1}^{k} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = \tilde{\boldsymbol{\mathcal{X}}}$$

where k denotes the rank of a tensor and $\circ$ denotes the outer product. $\mathbf{A} = [\mathbf{a}_1$ $\mathbf{a}_2$ ... $\mathbf{a}_k]$, $\mathbf{B} = [\mathbf{b}_1$ $\mathbf{b}_2$ ... $\mathbf{b}_k]$, and $\mathbf{C} = [\mathbf{c}_1$ $\mathbf{c}_2$ ... $\mathbf{c}_k]$ are the factor matrices. $\tilde{\boldsymbol{\mathcal{X}}}$ denotes the reconstructed tensor using factor matrices.

Definition 9. **Tucker decomposition**, also termed N-mode Principal Component Analysis (PCA) or Higher-order Singular Value Decomposition (HOSVD), decomposes a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ as

$$\boldsymbol{\mathcal{X}} \approx \boldsymbol{\mathcal{G}} \times_1 \mathbf{A} \times_2 \mathbf{B} \times_3 \mathbf{C} = [[\boldsymbol{\mathcal{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]] = \tilde{\boldsymbol{\mathcal{X}}}$$

where $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{P \times Q \times R}$ is termed the core tensor, with $P \ll I_1$, $Q \ll I_2$, and $R \ll I_3$. $\mathbf{A} \in \mathbb{R}^{(I_1 \times P)}$, $\mathbf{B} \in \mathbb{R}^{(I_2 \times Q)}$, and $\mathbf{C} \in \mathbb{R}^{(I_3 \times R)}$ are the factor matrices, with P, Q, R being the number of components in the three factor matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, respectively. If $P = Q = R$, then Tucker decomposition is equivalent to CP decomposition.

Definition 10. **Tensor-CUR decomposition** decomposes an input tensor into a small number of original frontal slices and original tubes. For a given tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, tensor-CUR decomposition computes three carefully constructed sub-tensors: $\boldsymbol{\mathcal{C}}$, $\mathbf{R}$, and $\mathbf{U}$, where $\boldsymbol{\mathcal{C}}$ consists of a small number of frontal slices of $\boldsymbol{\mathcal{X}}$, $\mathbf{R}$ consists of a small number of tubes of $\boldsymbol{\mathcal{X}}$ as its columns, and $\mathbf{U}$ is a carefully constructed matrix such that 3-mode product of $\boldsymbol{\mathcal{C}}$, $\mathbf{U}$, and $\mathbf{R}$ gives an approximation to the input tensor.

2.7 Summary

This chapter briefly discussed four multivariate analysis techniques namely PCA, CA, CCA, and matrix-CUR. In addition, definitions related to tensor decomposition techniques are presented. PCA plays a vital role in the context of applications involving Internet traffic data namely *estimation, reconstruction, and analysis*. We identify limitations of applying PCA to each of these applications and show the utility of the CA, matrix-CUR, CCA and tensor decomposition techniques.

Estimation Estimated TMs using PCA are sensitive to the rank parameter k. We alleviate this limitation by employing matrix-CUR decomposition.

Estimation Uses multiple expensive sources of information for accurate estimates. We address this issue by utilizing inexpensive sources of information and employing CCA to estimate accurate TM.

Reconstruction Reconstructing missing values in the TM is achieved by employing matrix decomposition techniques which use computationally expensive alternate least squares (ALS) procedure. We employ matrix-CUR and tensor-CUR decomposition techniques which do not involve the use of ALS and obtain an accurate reconstruction of missing values

Analysis The eigenflows obtained by employing PCA do not have the interpretability. In applications like volume anomaly analysis and structural analysis, interpretability is required. We employ CA and matrix-CUR decomposition methods which have the interpretability property.

Chapter 3

Estimation of Internet Traffic Data

In this chapter, we present our first contribution which proposes the use of two well-known decomposition techniques to alleviate the identified limitations of existing estimation techniques (as presented in Chapter 1) and obtain *improved estimation of Internet traffic data* with least estimation error.

3.1 Introduction

Inferring accurate traffic matrix is of paramount importance in solving many problems that arise in network operations and management tasks, such as traffic engineering, network provisioning, and design [60]. Traffic matrices are obtained through direct measurement using various monitoring tools. One such example is NetFlow monitor [3] which performs the measurement at the ingress router. For each traffic flow, NetFlow maintains a record in router memory containing many fields including source and destination IP addresses, source and destination ASes, flow starting and finishing timestamps, the number of bytes and number of packets transmitted. The collection methodology using NetFlow for direct measurement involves the following steps:

1. The NetFlow leaving the egress links is examined to build a set of the destination AS numbers connected to each egress router.

2. For each destination AS number, all the egress routers corresponding to this AS is listed.

3. At each ingress router, and for each NetFlow record, the destination AS number is mapped to a set of egress routers using the list obtained in step 2. Open Shortest Path First (OSPF) is then simulated to choose the closest egress router.

These NetFlow measurements provide local TMs for each of the ingress routers. These measurements would be sufficient to obtain complete traffic matrix provided NetFlow data were collected for the entire network. However, the fact is that NetFlow is often only partially deployed because products from some vendors do not support NetFlow in a way consistent to our needs, *i.e.*, different from Cisco NetFlow specification, and some do not support it at all. Apart from this, the local TMs need to be shipped to a central location for appropriate processing to form a complete traffic matrix for the entire network. The shipping costs of local TMs translate to communications overhead and the processing cost at the central location translates to computational overhead. These overheads are coupled with an additional storage overhead at the central location. It has been widely established that these overheads are so high that the direct measurement of traffic matrices for the entire network is not feasible. This has been one of the main motivations behind recent research targeted toward estimation techniques that can estimate the traffic matrix from the readily available Simple Network Management Protocol (SNMP) link counts.

The TM estimation Problem: For a network consisting of M links and P PoP pairs, the TM estimation problem is posed as solving a system of linear equations [14]

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t^{\mathsf{T}}, \tag{3.1}$$

where $\mathbf{y}_t \in \mathbb{R}^{M\times 1}$ is a known vector of link counts obtained using SNMP reports. $\mathbf{x}_t \in \mathbb{R}^{1\times P}$ is an unknown vector that represents the traffic volume across P PoP pairs during t^{th} time-interval. By solving Eq. (3.1), $\mathbf{x}_t$ is estimated from the known quantities $\mathbf{y}_t$ and $\mathbf{A}$. Since the number of unknowns (P) is more than the number of known parameters (M), this is an ill-posed inverse problem and there exist multiple solutions for $\mathbf{x}_t$.

Existing solutions: The traditional approaches for TM estimation assume an initial model for the TM and minimize the deviation of estimated TM from the modeled one, subject to the constraints imposed by Eq. (3.1). TM estimation techniques in the literature are classified into three generations, based on the models used for the traffic flows. The first generation techniques [14, 61, 62] assume Poisson or Gaussian distribution for the TM and do not capture the temporal and spatial correlations in the TM. The second-generation techniques use better models, such as the gravity model [63], generalized gravity model [64], and the route change method [65] to capture the spatial and temporal correlation. These techniques show an improvement in minimizing the error of estimation, but the error is still large. Some techniques like the tomogravity, generalized tomogravity, and {1}-inverse technique [66], strive to minimize the error in estimation by capturing the spatio-temporal properties.

Third generation techniques [6, 67, 15] make use of partial direct TM measurements along with the SNMP data for the estimation. Soule *et al.* [6] have used direct measurements of the previous day along with $\mathbf{Y}$ and $\mathbf{A}$ for estimation of successive days TM using Principal Component Analysis (PCA) method. PCA method assumes a low-rank model for the initial traffic matrix. The key idea behind using PCA is to estimate k eigenflows instead of estimating traffic volume flow across P PoP pairs, where k is the rank parameter. Since, $k \ll P$, the number of unknown parameters becomes less than the number of known parameters, hence the TM estimation problem becomes well-posed. The third-generation techniques are shown to outperform second-generation techniques in terms of the estimation error.

Limitations: First and second-generation techniques are characterized by a high error in the estimated TM along with high bias in estimation error. Although the third-generation techniques have been tuned to reduce the estimation error, they incur high variance in the estimation error making the estimated TM utility limited. Besides, we identify the following limitations of third-generation techniques.

1. *Use of expensive multiple sources of information*: The reduction in error in the third generation techniques is achieved because of the use of multiple sources of information including partial direct TM measurements for TM estimation which is an expensive affair to obtain due to the additional cost of measurement, communication, and storage. A fundamental question arises that *are there inexpensive multiple sources of information worthy of incorporation in TM estimation models*?

2. *Rank parameter sensitivity of PCA-based method*: PCA is argued to be sensitive to various application tasks that are built over TM in the literature [23]. In this thesis, we identify another limitation of PCA in the context of TM estimation, *i.e.*, the estimation accuracy is sensitive to the input rank parameter of PCA. This limitation hasn't been explored yet in literature. We are the first one to identify it and demonstrate it experimentally using publicly available Abilene traffic matrices.

3. *Naive use of multiple sources of information*: In some earlier works [6, 15] of TM estimation, distinct multiple sources of information (direct measurements and link count matrix) were utilized to estimate the TM. However, they do not make simultaneous use of multiple sources of information. These methods fall in the realm of single-view learning because they learn a single subspace corresponding to a distinct source of data. A naive way to simultaneously use multiple sources of data is to concatenate the features extracted from multiple sources. The work in [15] concatenated the link count matrix and traffic matrix into a single view. However, as noted in [16], such a concatenation causes over-fitting and is not physically meaningful, because

the statistical properties of each view are lost. Another fundamental question arises that *are there techniques that construct multiple learners from the identified sources of data, instead of naively concatenating information from multiple sources*?

Multiple sources of data are natural in some application domains, such as image processing, computer vision, and information retrieval. For example, in computer vision, the pose of a person is considered to be a distinct view of data [68]. To collect distinct poses, one uses multiple sensors (cameras) at different places to collect information about the *same subject*. In the audio-visual speaker clustering [69], the data corresponding to the audio and the video features are considered to be two distinct sources of data (obtained with different sensors). However, for clustering Wikipedia documents, each webpage is considered as one source of data, while the hyperlink structure across the webpages is considered as another source of information. In this case, multiple *derivations* of the existing data are presented as views of data (same data source). To summarize, we note that multiple views of data are typically obtained from either different sources or constructed as derivations of the same underlying data. Multi-view learning techniques construct *multiple learners*, one from each source of data [70]. Application of such multi-view learning methods, that consider multiple views of the same underlying data, has seen success in diverse application domains [71, 72, 73, 74].

3.1.1 Main Contributions

The summary of first contribution of the thesis is as follows:

1. We address the first limitation concerning *the use of expensive multiple sources of information* by making use of traffic matrix estimated using inexpensive sources of information only such as SNMP link counts. In particular, we make use of estimated traffic matrix (obtained using second-generation techniques such as tomogravity and generalized tomogravity) as input to the proposed decomposition techniques for TM estimation.

2. To overcome the *rank parameter sensitivity of PCA-based method*, we propose the use of matrix-CUR decomposition for TM estimation. The key idea for TM estimation using matrix-CUR decomposition is to estimate (a few) c columns instead of estimating all P columns of the traffic matrix.

3. To avoid the overfitting because of the *naive use of multiple sources of information*, we propose the use of multi-view subspace learning technique which considers the traffic matrices estimated using less expensive and easily available information as inexpensive multiple sources (or *views*) of information and simultaneously learns a common subspace shared by all views. We make use of Canonical Correlation Analysis (CCA) for subspace learning of multiple views.

3.1.2 Summary of Experimental Results

We use real-world traffic matrices from the Abilene network [7] for experiments. The summary of the experimental result is as follows:

1. Experimentation on Abilene traffic matrices [7] suggests that CUR decomposition
   - Exhibits less sensitivity to the rank parameter as compared to PCA-based approaches.
   - Achieve low spatial and temporal error in the estimation of Abilene TMs.
   - Exhibits bias value close to zero and low value of variance.

2. Two distinct sets of multi-view data are constructed for TM estimation: $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$ and $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, where $\mathbf{X}_g, \mathbf{X}_{gg}, \mathbf{X}_t$, and $\mathbf{X}_{gt}$ are the traffic matrices estimated using gravity, generalized gravity, tomogravity, and generalized tomogravity methods, respectively. We compute the correlation among the two sets using CCA to find that the set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$ has a maximum correlation along 39 canonical directions. Hence, we use this set for multi-view learning.

We observe the following from experimental results that also demonstrate the advantages of the proposed multi-view learning technique.

- CCA has the ability to reconstruct one view of TM using the TM from another view and all the projection directions. For the set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, the reconstruction error of $\mathbf{X}_t$ using $\mathbf{X}_{gt}$ and the projection directions is as low as 2.98%.
- The proposed multi-view learning technique gives accurate estimates of TM to reduce the temporal and spatial error by more than 80% compared to all the other techniques, including the third-generation techniques.
- CCA outperforms single-view learning techniques in terms of bias and variance values. The obtained bias value is close to zero which indicates an accurate estimation of TM. There is also a low variance in the estimation error with the proposed technique.

3.2 Related Work

As mentioned earlier, any TM estimation technique starts by assuming an initial model for the TM and then tries to estimate another TM with the measured data, as closely as possible to the modeled TM. Earlier literature on TM estimation can broadly be categorized into three generations, based on the model they use for the initial TM.

First and Second Generation Techniques: First-generation techniques assume that the entries in the TMs follow Poisson [14, 61] or Gaussian [62] distributions. These models do not capture the spatial and temporal correlations in the TMs, thus resulting in a high error in the estimated TMs. To reduce the error, second-generation techniques made use of TM models that account for the spatiotemporal correlations in the data. The authors of [63] used a gravity model for the TM that assumes that the traffic flow across an IE pair is independent

of the traffic flow across the other pairs. In [75], probabilistic methods based on the gravity model were used that outperformed the gravity model for TM estimation. Gunnar *et al.* [76] critically evaluated these estimation techniques and demonstrated that the gravity model fails as an initial TM model because of the violation of underlying traffic flow patterns.

To overcome the limitations of the gravity model, the generalized gravity model was proposed in [64] which imposes a conditional independence criterion between various IE node pairs and also takes into account the asymmetric nature of the traffic matrix. The generalized gravity model has been shown to attain a low estimation error of traffic matrices as compared to the gravity model. However, the reduction in error is not significant and needs to be further lowered. Subsequently, the authors of [77] have proposed the tomogravity method, that combines network tomography and gravity model to select a TM closely following the gravity model, while satisfying the link constraints in Eq. (3.1). Different from these approaches, the work in [66] formulated the TM estimation problem as an optimization problem and solved it by computing {1}-inverse of the routing matrix. For any given matrix *say* $\mathbf{B}$, there exists a matrix $\mathbf{B}^-$ which is not necessarily unique such that the following holds: $\mathbf{B}\mathbf{B}^-\mathbf{B} = \mathbf{B}$. $\mathbf{B}^-$ is called a {1}-inverse of $\mathbf{B}$. The general form of {1}-inverse of $\mathbf{B}$ is given as follows:

$$\mathbf{B}^- = \mathbf{B}^+ + \mathbf{K} - \mathbf{B}^+\mathbf{B}\mathbf{K}\mathbf{B}\mathbf{B}^+ \tag{3.2}$$

where $^+$ denotes the Moore-penrose generalized inverse [45] and $\mathbf{K}$ denotes an arbitrary matrix.

Third Generation Techniques: Traditional estimation techniques use a single view of network-wide traffic volume flow data to estimate the TMs. Third-generation techniques [6, 15, 78] make use of two views of traffic data; partial direct TM measurements and SNMP link data. Fanout method in [78] does not require routing matrix and does not need to solve Eq. (3.1). A node fanout is represented as a vector that denotes the fraction of total incoming traffic volume that it forwards to other egress PoPs during a given time interval. The key idea

behind the fanout method is that these node fanouts show strong periodicity and can be used to estimate traffic volume flow during successive weeks at the same time interval. Node fanouts are computed using direct measurements from the previous week to estimate the TM of next week.

Soule et. al. [6] used a subspace learning technique, namely PCA for TM estimation. They have assumed that the matrices $\mathbf{V}$ and $\mathbf{S}$, obtained after employing PCA on the previous one week TM measurement data, are stable and can be used to estimate traffic matrices for successive weeks. PCA-based approach relies on the fact that top few PCs (first few columns of $\mathbf{V}$) captures most of the energy of TM. They have relaxed the Eq. (3.1) to the following equation

$$\mathbf{y}_t = \mathbf{A}\mathbf{V}_k\mathbf{S}_k\mathbf{e}_t^{\mathsf{T}}, \tag{3.3}$$

where $\mathbf{V}_k$ is $(P \times k)$ matrix with only the top k principal components, $\mathbf{S}_k$ is the corresponding diagonal matrix of size $(k \times k)$ and vector $\mathbf{e}_t^{\mathsf{T}}$ has the values of the k most significant eigenflows during t^{th} time-interval. The matrices $\mathbf{V}_k$ and $\mathbf{S}_k$ can be obtained from SVD of partial direct TM measurement of the previous week. So, the problem defined in Eq. (3.3) is of estimating unknown top k significant eigenflows instead of estimating P OD flows. As $k \ll P$, the number of unknown parameters becomes less than the number of known parameters, the TM estimation problem becomes well-posed. Eq. (3.3) can be solved by taking pseudo-inverse of $\mathbf{A}\mathbf{V}_k\mathbf{S}_k$ to obtain an estimate of top significant eigenflows $\mathbf{e}_t^{\mathsf{T}}$, which can be used to estimate the traffic matrix measurement at time-interval t of successive week using following equation

$$\hat{\mathbf{x}}_t^{\mathsf{T}} = \mathbf{V}_k\mathbf{S}_k\mathbf{e}_t^{\mathsf{T}}, \tag{3.4}$$

for t = 1,..., T. $\hat{\mathbf{x}}_t^{\mathsf{T}} \in \mathbb{R}^{P\times 1}$ denote the estimated traffic volume flow across P PoP pairs during t^{th} time-interval. In [6], the authors used partial direct TM measurements of the previous day along with link data, thereby their technique outperforms the other estimation techniques that use only link data.

After Soule's study [6], various authors have used PCA-based approaches for traffic matrix estimation. Zhang *et al.* [9] have solved the TM estimation problem using Sparsity Regularized Matrix Factorization (SRMF). Zhao *et al.* [79] have proposed a PCA based Optimization Method (PCAOM) for TM estimation where the objective is to minimize the Mahalanobis distance of estimated TM from the modeled one, subject to the constraints imposed by Eq. (3.1). The robust traffic matrix estimation (RTME) technique in [15], use a vertical concatenation of partial direct TM measurements and SNMP link data to solve the TM estimation problem using the least-squares estimator. It uses Singular Value Decomposition (SVD) to obtain a weighted least square estimate of the TM. SVD forms the basis for various subspace learning techniques including PCA. Results show that one can obtain a much more accurate estimate of the TMs using the two sources of information rather than a single source.

Apart from subspace learning techniques, deep learning techniques have also been employed for TM estimation. Jiang *et al.* [80] estimated the TM based on Back-Propagation Neural Network (BPNN) which uses SNMP link data as an input and TM as the output for training BPNN. The trained BPNN is then used to predict future TMs. Zhou *et al.* [81] have shown improvement over [80] by using the routing matrix along with SNMP link data as an input to the neural network. Authors in [82] have argued that with an increase in the size of the network, traffic flows are complex to characterize and the ill-posed nature of the estimation problem gets worse. Motivated by this, they proposed deep learning architecture based on the Deep Belief Network (DBN) for TM estimation. DBN is able to outperform subspace learning techniques, namely PCA, in terms of estimation accuracy. The authors of [83] have proposed to use the Convolutional Neural Network (CNN) to extract the spatiotemporal nature of traffic for traffic matrix estimation in the context of Vehicular Ad-hoc NETworks (VANETs).

Recent works in [67, 84] studied TM estimation for Software-Defined Networking (SDN)-based IP networks using SNMP link volume measurements, destination-based flow measurements, and origin-destination based flow measurements. An SDN router enables the measurement of different types of data, but the cost and

inaccuracy associated with gathering the data still remain. Our work considers the TM estimation only in the context of a conventional IP backbone network without the requirement of additional sources of data.

Motivation: First and second-generation techniques are characterized by a high error in the estimated TM along with high bias in estimation error. Third generation techniques make use of multiple sources (views) of data, *i.e.*, in addition to link counts and routing matrix, they used partial direct measurements to improve the estimation of TM. Subspace learning approaches based on PCA are argued to be sensitive to various application tasks that are built over TM in the literature [23]. In this thesis, we identify one limitation of PCA in the context of TM estimation, *i.e.*, the estimation accuracy is sensitive to the input rank parameter of PCA. The rank parameter is a deciding factor and stands for the number of eigenflows to be estimated instead of estimating all OD flows. Third generation techniques also lead to high variance in estimation error and the use of partial direct measurement is an expensive affair. Also, the vertical concatenation of multiple sources of data, as performed in one of the techniques [15] causes over-fitting. In this thesis, we (i) make use of inexpensive multiple sources of information for TM estimation, (ii) address the rank parameter sensitivity using matrix-CUR decomposition, and (ii) employ a multi-view subspace learning technique to avoid the issue of over-fitting.

3.3 Proposed Decomposition Techniques

In this section, we present the use of matrix-CUR decomposition to alleviate rank parameter sensitivity in estimating TM and multi-view subspace learning technique based on CCA for accurate estimation of TM. The proposed techniques do not make use of expensive sources of information for TM estimation.

3.3.1 Matrix-CUR Decomposition

The motivation for employing matrix-CUR decomposition is that the matrices **C** and **R** retain a subset of columns and rows of the original input TM. In the case of PCA, the columns of matrix **E** are the weighted linear combination of all the columns of the original input TM. By retaining the original rows and columns in the factored matrices lead to insensitivity in the estimated TM when the rank parameter is varied.

For the input TM, we make use of estimated TMs obtained using second-generation methods. Let $\mathbf{X}_{w-1}$ and $\mathbf{X}_w$ denote the previous week estimated TM and current week estimated TM obtained using the tomogravity method. PCA uses previous week's $(w-1)$ TM obtained using direct measurements to estimate the current week (w) TM. In order to eliminate the use of the direct measurement, we propose the use of $\mathbf{X}_{w-1}$ instead. Three steps are involved in estimating the current week (w) TM given previous week $(w-1)$ TM.

1. Obtain the column indices along with the factor matrices using $\mathbf{X}_{w-1}$ by employing matrix-CUR decomposition as given below

$$\mathbf{X}_{w-1} = \mathbf{C}_{w-1}\mathbf{U}_{w-1}\mathbf{R}_{w-1} \tag{3.5}$$

2. The columns in the $\mathbf{X}_w$ are selected according to the column indices obtained in step 1 to form $\mathbf{C}_w$.

3. Refine the estimate $\mathbf{X}_w$ by using $\mathbf{C}_w$ obtained in step 2 and $\mathbf{U}_{w-1}$ and $\mathbf{R}_{w-1}$ obtained in step 1 as given below

$$\hat{\mathbf{X}}_w = \mathbf{C}_w\mathbf{U}_{w-1}\mathbf{R}_{w-1} \tag{3.6}$$

where $\hat{\mathbf{X}}_w$ denote the TM of week w estimated using matrix-CUR decomposition.

We demonstrate that $\hat{\mathbf{X}}_w$ is insensitive to the rank parameter unlike the counterpart based on PCA. The proposed method eliminates the use of TM obtained using direct measurement method. Even though matrix-CUR decomposition provides a rank insensitive estimate, it uses a single source of information namely $\mathbf{X}_{w-1}$. In section 3.2, we presented the motivation for employing *multiple sources* of information for *better* TM estimation. In the following section, we identify the *multiple sources of information* and their use in TM estimation through multi-view subspace learning.

3.3.2 Multi-view Subspace Learning using Canonical Correlation Analysis

The class of learning techniques that take into account multiple feature sets to learn a hypothesis are termed multi-view learning techniques. For such techniques, the same data viewed from distinct perspectives can yield multiple independent feature sets. The multi-view learning techniques are shown to have better performance than those using a single view [70]. As discussed earlier, there are two ways to get multiple views of data; using distinct sensors or using different views of the same underlying data. For example, the text and image features extracted from a webpage are *distinct views* of the same underlying webpage, but with data from different sources [71]. On the other hand, two distinct views of the webpage can be constructed by considering the content as one view, and a graph with nodes representing the webpages and edges representing the hyperlinks as another view. However, the number of webpages and the number of nodes in the graph must be the same, because for every data point in one view (node in the graph) there must exist a corresponding data point in another view (the content of the webpage corresponding to the node), termed *pairing*.

Definition of a View:

In other applications of multi-view learning to information retrieval [71], computer vision [72], computational biology [73] and natural language processing [74], defining a view as well as the *pairing* of data points between multiple views occur

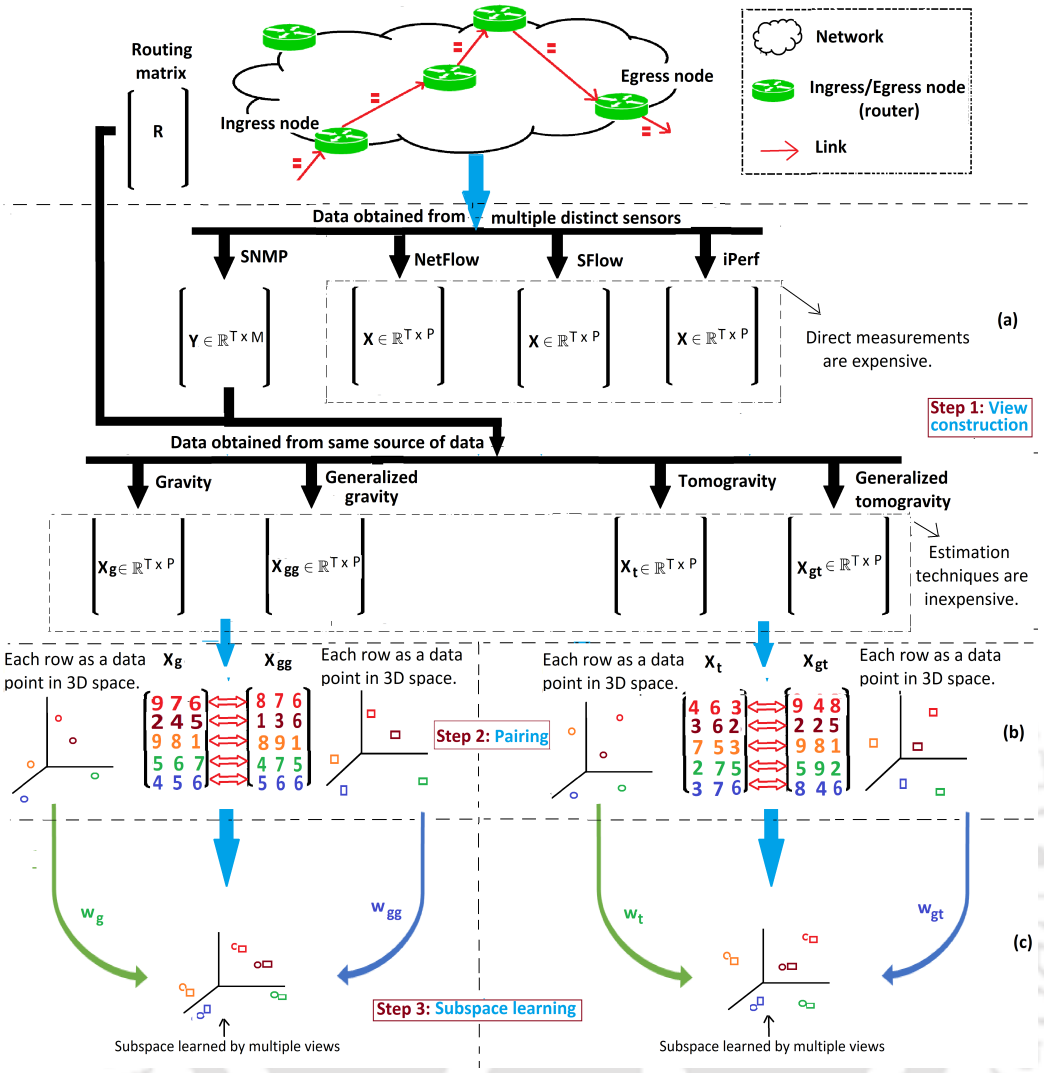


FIGURE 3.1: Schematic diagram of traffic matrix (TM) estimation using a multi-view subspace learning technique. For illustration purpose, sample TM matrices $\mathbf{X}_g, \mathbf{X}_{gg}, \mathbf{X}_t$, and $\mathbf{X}_{gt}$ are denoted with three IE node pairs and five successive time-intervals.

naturally. In the context of the network-wide traffic flows, defining a view as well as obtaining *paired* data points are challenging.

As shown at the top of Fig. 3.1, traffic enters the network through an ingress node, traverses multiple links and then leaves through an egress node. Each *view* of this data, captured as a matrix, contains the network-wide traffic volume, *i.e.*, the volume of traffic across different IE pairs or links, measured during a particular time interval. To obtain multiple views of the data, one could use tools like NetFlow, SFlow or iPerf to collect the traffic volume data across IE pairs, whereas SNMP can be used to obtain individual link count data. The upper part of Fig. 3.1(a)

illustrates how four distinct views of data can be obtained for learning. However, it is an expensive affair to collect data from direct measurements in large scale networks.

To avoid the use of expensive direct measurement, some TM estimation techniques use inexpensive link-level data along with the routing matrix to estimate the TM. Since they use the same source of information for TM estimation, we could consider the estimated traffic matrices as distinct *views* of traffic flow data. In the lower part of Fig. 3.1(a), we illustrate how to obtain four distinct views of TMs using different estimation techniques; gravity, generalized gravity, tomogravity and generalized tomogravity, labelled as $\mathbf{X}_g, \mathbf{X}_{gg}, \mathbf{X}_t$, and $\mathbf{X}_{gt}$, respectively. We utilize sets of estimated TMs $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$ and $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, as multiple views of data. As these views are TMs estimated using second-generation techniques, using them as an input for TM estimation gives us an additional advantage over techniques that use partial direct measurement, like PCA and RTME. Note that obtaining partial direct measurements is associated with the cost of measurement, communication, and processing.

Defining Pairing Between Views:

In multi-view learning, there has to be a correspondence between data points present in one view with the other view which is termed *data pairing*. The next step in our work is to define the *pairing* between multiple views of the data. In this work, we consider two views at a time as an input to the multi-view learning technique. Two views of the data, consisting of TMs estimated with two different techniques, estimate the traffic across the same IE node pairs for the same time intervals. Therefore, we propose to pair every row of the first view, obtained using one estimation method with the corresponding row from the second view, obtained using another estimation method.

In Fig. 3.1(b), we illustrate the pairing of data points (using red-colored bidirectional arrows) for the constructed multi-view dataset using $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$ and $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, where the view is for three IE pairs and five time-intervals. Each row of a view corresponds to a data point in three-dimensional space as shown in

Fig. 3.1(b). Consider the two views $\mathbf{X}_g$ and $\mathbf{X}_{gg}$ in the left part of Fig. 3.1(b). For set $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$, we pair row 1 of $\mathbf{X}_g$ with row 1 of $\mathbf{X}_{gg}$, row 2 of $\mathbf{X}_g$ with row 2 of $\mathbf{X}_{gg}$ and so on till row 5 of $\mathbf{X}_g$ with row 5 of $\mathbf{X}_{gg}$. This pairing is essential because the objective of multi-view subspace learning is to maximize the correlation between these paired data points corresponding to the same time-interval in the subspace. This pairing is natural because both the estimated TMs are for the same network topology, the traffic volume estimates are for the same IE pairs across the same time intervals, and every row in the estimated TM has a corresponding row in another TM, which essentially means that no data points are missing.

Intuition Behind Data and Learning Models:

We present the intuition behind defining the views of traffic matrix data and using a multi-view learning technique for TM estimation.

Answering "what constitute multiple views of traffic matrix data?" is a non-trivial task. In multi-view learning, information collected with regards to the same underlying object using different approaches can be considered as multiple views of the observed object. The multiple derivations of data in various application domains use different feature extraction methods. For example, in image processing, extracting edge features of image or texture features, obtaining line drawing of the image, extracting SIFT features, etc., are considered as multiple views. However, in the context of traffic matrix estimation, no such feature extraction methods are available. In this context, different estimation techniques give us different estimations of the TM. Since each technique uses the same underlying network topology, these estimates lead to different *derivations* of the traffic in the same network. Therefore, the proposed notion that each estimated TM serves as a view is consistent with the general definition of a *view* in multi-view learning. To verify the consistency across different views, we also show the correlation between them (see the result in Fig. 3.5), which also asserts the similarity.

In all the traffic matrix estimation methods that use subspace learning, single source of data is used and one set of projection directions is learned to be associated with a single source. Obtaining an accurate estimate of the TM from a set of

inexpensive and inaccurate estimated TMs is the central idea of this work. This is done by learning a common subspace associated with different views of data. Multiple projection directions are learned to be associated with multiple sources of data to derive a common subspace, that maximizes the correlation across different estimates.

The Problem:

Given two views constructed from different estimation techniques, the problem is to maximize the correlation between the paired data points in the subspace learnt using multiple views. Projection of the input views onto fewer dimensions of this subspace gives TMs estimated using multi-view learning, which also corresponds to the problem of joint dimensionality reduction. We define the joint dimensionality reduction problem as, given two views $\mathbf{X}_t$ and $\mathbf{X}_{gt}$, find low-dimensional mappings $\hat{\mathbf{X}}_t = \mathbf{X}_t\mathbf{W}_t$ and $\hat{\mathbf{X}}_{gt} = \mathbf{X}_{gt}\mathbf{W}_{gt}$ that *maximize the correlation* between $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{X}}_{gt}$, defined by

$$\rho(\hat{\mathbf{X}}_t, \hat{\mathbf{X}}_{gt}) = \frac{\mathbb{E}[(\hat{\mathbf{X}}_t)^\mathsf{T}\hat{\mathbf{X}}_{gt}]}{\sqrt{\mathbb{E}[(\hat{\mathbf{X}}_t)^\mathsf{T}\hat{\mathbf{X}}_t]\mathbb{E}[(\hat{\mathbf{X}}_{gt})^\mathsf{T}\hat{\mathbf{X}}_{gt}]}} = \frac{\mathbb{E}[(\mathbf{W}_t)^\mathsf{T}(\mathbf{X}_t)^\mathsf{T}\mathbf{X}_{gt}\mathbf{W}_{gt}]}{\sqrt{\mathbb{E}[(\mathbf{W}_t)^\mathsf{T}(\mathbf{X}_t)^\mathsf{T}\mathbf{X}_t\mathbf{W}_t]\mathbb{E}[(\mathbf{W}_{gt})^\mathsf{T}(\mathbf{X}_{gt})^\mathsf{T}\mathbf{X}_{gt}\mathbf{W}_{gt}]}} \quad (3.7)$$

where $\mathbf{W}_t$ and $\mathbf{W}_{gt}$ are the projection matrices for $\mathbf{X}_t$ and $\mathbf{X}_{gt}$, respectively. In the numerator $\mathbb{E}[(\mathbf{W}_t)^\mathsf{T}(\mathbf{X}_t)^\mathsf{T}\mathbf{X}_{gt}\mathbf{W}_{gt}]$ denotes the covariance between the two views and the denominator is the product of view-specific covariances. The low-dimensional mappings $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{X}}_{gt}$ correspond to the TMs estimated using multi-view subspace learning.

Canonical Correlation Analysis:

Once the views are constructed and pairing is done, the final step in multi-view learning is subspace learning, which seeks a unified space shared by all views. We make use of Canonical Correlation Analysis (CCA) for multi-view subspace learning. We provide two views at a time as an input to CCA, which in turn

maximizes the correlation between the input views and provides two sets of projection directions. As illustrated in the right part of Fig. 3.1(c), CCA takes $\mathbf{X}_t$ and $\mathbf{X}_{gt}$ as input and gives the set of projection directions $\mathbf{W}_t$ and $\mathbf{W}_{gt}$ as the output. Similarly, the left part of Fig. 3.1(c) illustrates working of CCA for $\mathbf{X}_g$ and $\mathbf{X}_{gg}$, which gives the set of projection directions $\mathbf{W}_g$ and $\mathbf{W}_{gg}$ as the output. The complete derivation of CCA using $\mathbf{X}_1$ (or $\mathbf{X}_t$) and $\mathbf{X}_2$ (or $\mathbf{X}_{gt}$) as input is presented in section 2.5.

The projection of $\mathbf{X}_t$ and $\mathbf{X}_{gt}$ onto the intermediate space, *i.e.*, $\hat{\mathbf{X}}_t = \mathbf{X}_t \times \mathbf{W}_t$ and $\hat{\mathbf{X}}_{gt} = \mathbf{X}_{gt} \times \mathbf{W}_{gt}$, denote the traffic matrices estimated using multi-view learning. The effectiveness of estimated TM using multi-view learning is measured by comparing it with the original TM. For this, the original TM is projected into intermediate space as

$$\breve{\mathbf{X}} = \mathbf{X}\mathbf{W}_t \tag{3.8}$$

or

$$\breve{\mathbf{X}} = \mathbf{X}\mathbf{W}_{gt} \tag{3.9}$$

The accuracy of the estimated matrix is studied by comparing it with the original TM using corresponding evaluation measures, the results for which are presented in the next section.

The intermediate subspace learnt using multiple views has the ability to transform one view to the other view. That is, given $\mathbf{X}_t$ and the mappings $\mathbf{W}_t$ and $\mathbf{W}_{gt}$, one can reconstruct $\mathbf{X}_{gt}$ by

$$\tilde{\mathbf{X}}_{gt} = \mathbf{X}_t\mathbf{W}_t\mathbf{W}_{gt}^{-1} \tag{3.10}$$

where $\tilde{\mathbf{X}}_{gt}$ is the reconstructed matrix $\mathbf{X}_{gt}$. Similarly, given $\mathbf{X}_{gt}$, $\mathbf{W}_t$, and $\mathbf{W}_{gt}$, one can reconstruct $\mathbf{X}_t$ by

$$\tilde{\mathbf{X}}_t = \mathbf{X}_{gt}\mathbf{W}_{gt}\mathbf{W}_t^{-1} \tag{3.11}$$

where $\tilde{\mathbf{X}}_t$ is the reconstructed matrix $\mathbf{X}_t$.

3.4 Experimental Results

In this section, we report the results of experiments conducted to measure the estimation error for the TMs using the proposed techniques, namely matrix-CUR decomposition and multi-view subspace learning technique using CCA. The data used for experimentation is collected from the Abilene backbone network [7] that consists of 11 PoPs across North America. The dataset has 24 TMs reflecting traffic volume data from March 1^{st} to September 10^{th}, 2004. The TMs are numbered **X01**, **X02**, ..., **X24**. Each TM is of size (2016×121) containing data across 121 PoP pairs during a week period (2016 intervals, each of 5 minutes). However, we use only 19 TMs corresponding to data from May 1^{st} to September 10^{th}, 2004 (**X06** - **X24**) as there are no missing periods. For third-generation techniques, we use direct TM measurement **X06** as input and estimate successive weeks TMs **X07**, ..., **X24**. For matrix-CUR decomposition, we use tomogravity estimates of 6^{th} week TM **X06** as input and estimate successive week TMs. For multi-view learning, we use tomogravity and generalized tomogravity estimates of current week TM as input views and estimate current week TM by learning a common subspace shared by input views. Note that the proposed decomposition techniques do not make use of direct measurement which is an expensive affair to obtain.

3.4.1 Evaluation Measures

We consider the four evaluation measures of a TM estimation technique which are widely reported in the majority of the literature [6, 15]. The four considered evaluation measures are discussed as follows:

1. Relative Temporal Error (RTE): Temporal error is computed for each time-interval such that it adds the estimation error during a given time interval across all PoP pairs. The relative temporal error of t^{th} time-interval is given by

$$RTE(t) = \frac{\sqrt{\sum_{j=1}^{P}(X(t,j) - \hat{X}(t,j))^2}}{\sqrt{\sum_{j=1}^{P}(X(t,j))^2}} \tag{3.12}$$

where P is the total number of PoP pairs. $\mathbf{X}$ and $\hat{\mathbf{X}}$ denote the original and estimated TM respectively. The relative temporal error (RTE) computation for multi-view learning is different, because the RTE is computed in the intermediate subspace. We use a modified definition for the RTE of multi-view learning as given below:

$$RTE(t) = \frac{\sqrt{\sum_{j=1}^{N}(\breve{X}(t,j) - \hat{X}_{gt}(t,j))^2}}{\sqrt{\sum_{j=1}^{N}(\breve{X}(t,j))^2}} \tag{3.13}$$

where N is the number of maximally correlated directions. $\breve{X}(t,j)$ is the $(t,j)^{th}$ element of $\breve{\mathbf{X}} = \mathbf{X}\mathbf{W}_{gt}$ which is the projection of original matrix onto intermediate subspace using $\mathbf{W}_{gt}$. $\hat{X}_{gt}(t,j)$ is the $(t,j)^{th}$ element of $\mathbf{X}_{gt}\mathbf{W}_{gt}$, which is the projection of TM estimated with generalized tomogravity technique in the intermediate subspace using $\mathbf{W}_{gt}$.

2. Relative Spatial Error (RSE): Spatial error is computed for each PoP pair such that it adds the estimation error across a given PoP pair during all time-intervals. Relative spatial error of jth OD flow is given by

$$RSE(j) = \frac{\sqrt{\sum_{i=1}^{T}(X(i,j) - \hat{X}(i,j))^2}}{\sqrt{\sum_{i=1}^{T}(X(i,j))^2}} \tag{3.14}$$

where T denote the total number of time-intervals. For multi-view learning,

the relative spatial error (RSE) for j^{th} OD flow is also computed in the intermediate subspace using the following equation on the same set of variables used in Eq. 3.13.

$$RSE(j) = \frac{\sqrt{\sum_{t=1}^{T}(\breve{X}(t,j) - \hat{X}_{gt}(t,j))^2}}{\sqrt{\sum_{t=1}^{T}(\breve{X}(t,j))^2}} \tag{3.15}$$

3. Bias of estimation error: A consistent difference between estimated and true values is called estimation bias. An ideal estimation technique should have an estimation bias close to zero. A positive value of bias corresponds to the overestimation of traffic volume flow and a negative value of bias implies an underestimation of traffic volume flow. Estimation techniques with bias value close to zero tend to characterize the traffic volume correctly and help in network capacity planning [6]. The bias of j^{th} OD flow denote the mean of total estimation error across j^{th} PoP pair and is given as

$$bias(j) = \frac{1}{T}\sum_{i=1}^{T}(\hat{X}(i,j) - X(i,j)) \tag{3.16}$$

For multi-view learning technique, bias is defined using the following equation on the same set of variables used in Eq. (3.13)

$$bias(j) = \frac{1}{T}\sum_{i=1}^{T}(\hat{X}_{gt}(i,j) - \breve{X}(i,j) \tag{3.17}$$

where T denotes the total number of time-intervals.

4. Variance of estimation error: Estimation techniques with low variance tend to track changes between individual time points very well and are useful for applications such as anomaly detection. The variance of estimation error for

j^{th} PoP pair is given by

$$Var(j) = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(\hat{X}(i,j) - X(i,j) - bias(j))^2} \tag{3.18}$$

For multi-view learning, the variance of estimation error of j^{th} OD flow is defined using the following equation on the same set of variables used in Eq. (3.13)

$$Var(j) = \sqrt{\frac{1}{T-1}\sum_{i=1}^{T}(\hat{X}_{gt}(i,j) - \breve{X}(i,j) - bias(j))^2} \tag{3.19}$$

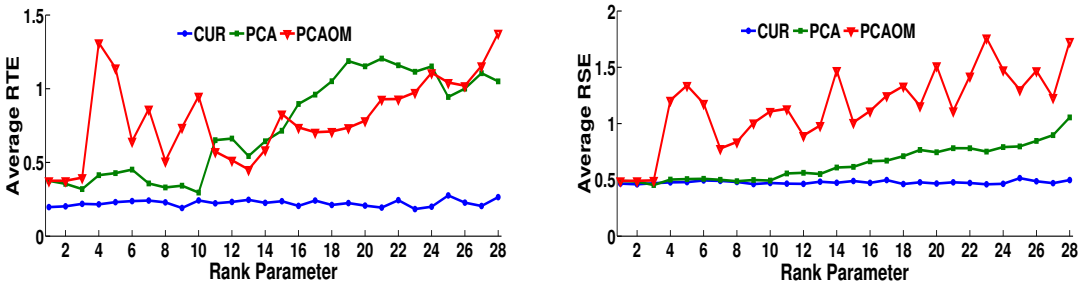
3.4.2 Matrix-CUR Decomposition

Here we present the experimental results of traffic matrix estimation using matrix-CUR decomposition. First, we demonstrate the rank parameter sensitivity of PCA-based approaches in terms of estimation error. Next, we present the efficacy of matrix-CUR decomposition through a comparison with the state-of-the-art techniques in terms of four evaluation measures discussed in the section 3.4.1.

Rank Sensitivity Analysis:

First, we demonstrate that the TMs estimated using PCA-based approaches, namely PCA [6] and PCAOM [79], are sensitive to the input rank parameter. Abilene network consists of 28 links and 121 PoP pairs. Since the number of links (known parameters) in the Abilene network is 28, the number of unknown parameters to be estimated should be less than or equal to 28 to make the problem well-posed. Authors of [6] have fixed the rank parameter of PCA to be 10. That is, they have proposed to estimate top 10 eigenflows, *i.e.*, first 10 columns of **E**, instead of estimating 121 columns of **X** using PCA. Authors of [79] have considered the rank parameter as the number of PCs which captures 90% of the variance in the input data.

Figure 3.2 shows rank parameter sensitivity of PCA [6], PCAOM [79], and matrix-CUR decomposition for estimation of **X07** in terms of average RTE and average



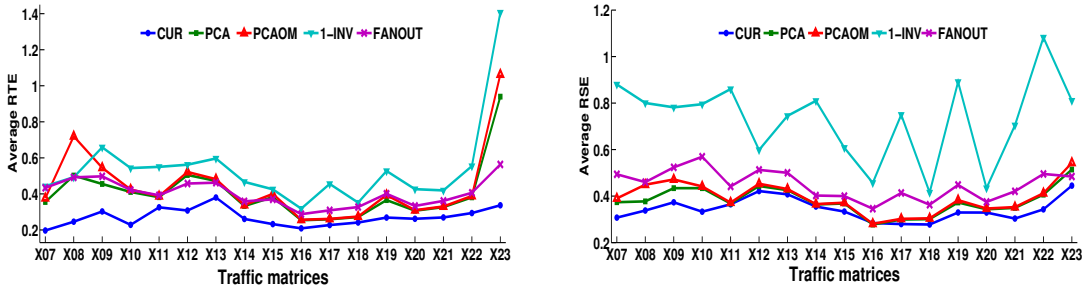
(a) X07 (b) X07

FIGURE 3.2: Rank parameter sensitivity for traffic matrix estimation.

RSE. In each of these plots, the x-axis denotes the rank parameter and the y-axis denotes average RTE or average RSE. The rank parameter is varied from 1 to 28. From these plots, we can observe that for every rank parameter matrix-CUR decomposition outperforms PCA and PCAOM in terms of average RTE and average RSE. We also observe that with the change in the rank parameter, the change in estimation error is very less in the case of matrix-CUR decomposition. On the other hand, PCA and PCAOM exhibit high variation in the estimation error with the change in the rank parameter. For PCA and PCAOM, the variation factor (maximum divided by minimum value) for RTE is 4.07 and 3.66 respectively. The average RSE value also exhibits a high variation factor of 2.32 and 3.58 for PCA and PCAOM respectively. For matrix-CUR decomposition, the variation factor of RTE and RSE is 1.50 and 1.11 respectively. Experimentation is performed for estimation of remaining TMs (**X08** - **X24**) as well and we observe high variation factor in estimation error for PCA and PCAOM and less variation factor for matrix-CUR decomposition. The high variation in estimation error with small changes in the rank parameter demonstrates the sensitivity of PCA-based approaches for TM estimation very well. On the contrary, CUR decomposition exhibits a small change in the value of temporal and spatial error with changes in the value of the rank parameter. This suggests matrix-CUR decomposition to be an apt technique for TM estimation as compared to PCA-based approaches.

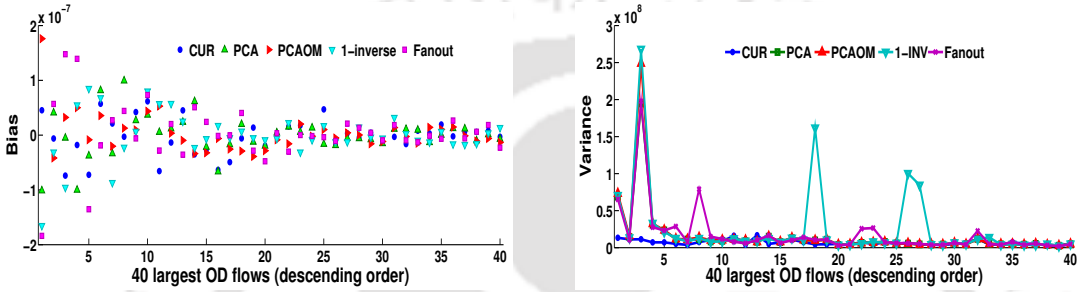
Measuring the Accuracy of Estimated TMs:

Here, we demonstrate the efficacy of matrix-CUR decomposition for TM estimation compared to state-of-the-art techniques namely PCA, PCAOM, Fanout, and



(a) Average RTE (b) Average RSE

FIGURE 3.3: Average RTE and RSE of Abilene Traffic Matrices.



(a) Bias (b) Variance

FIGURE 3.4: Bias and Variance of top 40 OD flows of **X07**.

{1}-inverse method. Figure 3.3 shows the average relative temporal and spatial error incurred in estimation of 17 TMs (**X07** - **X23**). The x-axis denotes the estimated TMs and the y-axis denotes the average RTE or average RSE value. We fix the rank parameter for estimating TM using PCA, PCAOM and matrix-CUR decomposition as 2 in this experiment. We observe that the CUR decomposition provides a more accurate estimate of traffic matrices as compared to the four techniques in terms of exhibiting the least average RTE and RSE value. Consider the estimation of **X07**. We observe that CUR decomposition attains **i)** an average RTE of 0.19; while for PCA, PCAOM, {1}-inverse, and Fanout method, it is 0.35, 0.38, 0.44, and 0.43 respectively. **ii)** an average RSE of 0.30; while for PCA, PCAOM, {1}-inverse, and Fanout method, it is 0.37, 0.39, 0.88, and 0.49 respectively.

We also demonstrate the efficacy of matrix-CUR decomposition in terms of the bias and variance of estimation error. Figure 3.4 shows the bias values of 40 largest OD flows of the estimated TM **X07**. From the plot of bias value, we observe that

the bias value for estimation of 1^{st} largest OD flow is closer to zero as compared to other techniques. In particular, PCA, {1}-inverse, and fanout method incurs a large negative bias value and PCAOM incurs a large positive bias value. This implies that PCA, {1}-inverse, and fanout method exhibits a large underestimation of 1^{st} largest OD flow; PCAOM exhibits a large overestimation of 1^{st} largest OD flow. On the other hand, CUR decomposition also exhibits overestimation but the difference is quite low and hence provides a more accurate estimate of 1^{st} largest OD flow. This observation holds for the majority of the large OD flows. From the variance plot, we observe that CUR decomposition exhibits a low value of variance for the majority of the large OD flows; while PCA, PCAOM, {1}-inverse, and Fanout method exhibit large variance for top largest OD flows. The obtained bias and variance values of estimation error suggest CUR decomposition as a better TM estimation technique as they yield estimates closer to the true value.

Through the experimental results, we learned that matrix-CUR decomposition provides a more accurate estimate of traffic matrices as compared to the recent second and third-generation techniques. An added advantage is that matrix-CUR decomposition does not make use of partial direct TM measurements which is an expensive affair to obtain. Rather it makes use of TM estimated using less expensive and easily available information, *i.e.*, link counts and routing matrix. Hence, TM estimation using matrix-CUR decomposition corresponds to second-generation techniques as it does not involve any use of partial direct TM measurements.

3.4.3 Multi-view Subspace Learning Technique

Here we present the experimental results of traffic matrix estimation using a multi-view subspace learning technique namely CCA. First, we present the correlation between different input views. Next, we present the error in the reconstruction of one input view TM using other input view TM to demonstrate the utility of common subspace learned using multiple input views. Finally, we present the

accuracy of the traffic matrix estimated using a multi-view learning technique in terms of the four evaluation measures.

Correlation between Different Views of TM:

First, we present the correlation between the different views of TM **X19**. These views, $\{view1, view2\}$ correspond to the TMs constructed with two second generation techniques discussed in Section 3.3.2. For quick reference, in the view set, $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$, $\mathbf{X}_g$ denotes the TM estimated using gravity model [63] while $\mathbf{X}_{gg}$ denotes the TM estimated using generalized gravity model [64]. Similarly, in the view set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, $\mathbf{X}_t$ is the TM estimated using tomogravity model [77] and $\mathbf{X}_{gt}$ is the one estimated using generalized tomogravity model technique [77].

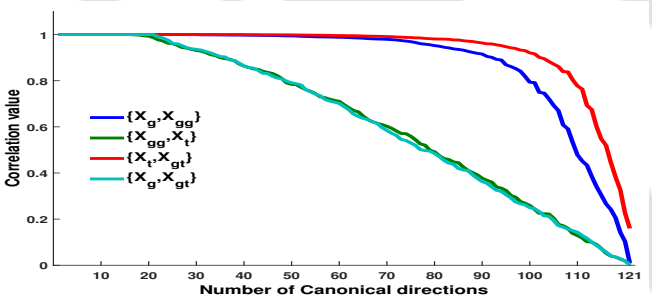


FIGURE 3.5: Correlation between sets of estimated TMs for **X19**.

Fig. 3.5 shows the correlation value, between the two TM estimates corresponding to different view sets, computed along the canonical directions. We observe that CCA employed on the set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, gives a larger number of maximally correlated canonical directions. In fact, for the TMs estimated with tomogravity and generalized tomogravity, 39 directions out of 121 directions are found to give a perfect correlation value of 1.0. Hence, we chose the TMs corresponding to view set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$ in most of the experiments.

The larger correlation values for the set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$ signifies that tomogravity and generalized tomogravity methods are very similar in nature and the variation of estimated traffic volumes between these two methods along the principal directions is low. Thus, when the estimated TMs are projected along these directions, it will reduce the variation in the estimated TMs. However, for the sake of completeness,

we also show the estimation error employing the other view combinations as well in Section 3.4.3.

Henceforth, we use the TMs estimated using the second-generation techniques as input for TM estimation. This avoids the inherent cost involved in techniques, like PCA and RTME that use partial direct measurement.

Reconstruction Error:

In this experiment, we reconstruct a TM estimated with generalized tomogravity technique using the TM estimated with tomogravity technique and projection matrices as specified by Eq. (3.10). The reconstruction error is computed by

$$\textbf{Reconstruction error (in \%)} = \frac{\|\mathbf{X}_{gt} - \tilde{\mathbf{X}}_{gt}\|_F}{\|\mathbf{X}_{gt}\|_F} * 100 \tag{3.20}$$

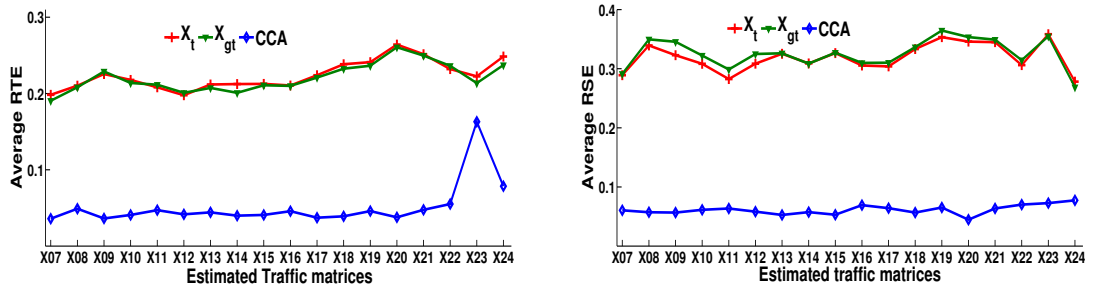
Similarly, the TM with tomogravity technique is reconstruted using the generalized tomogravity TM using Eq. (3.11). The reconstruction error for this is

$$\textbf{Reconstruction error (in \%)} = \frac{\|\mathbf{X}_{t} - \tilde{\mathbf{X}}_{t}\|_F}{\|\mathbf{X}_{t}\|_F} * 100 \tag{3.21}$$

For the view set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$ of **X19** TM, the reconstruction error of $\mathbf{X}_t$ using $\mathbf{X}_{gt}$ along the projection directions is 2.98%, whereas the reconstruction error of $\mathbf{X}_{gt}$ using $\mathbf{X}_t$ is 5.59%. Note that, despite the reconstruction being performed using different views of TM, there is a low error in reconstruction. For the view set $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$, the reconstruction error of $\mathbf{X}_g$ using $\mathbf{X}_{gg}$ is 11.32%, while it is 8.14% for $\mathbf{X}_{gg}$ using $\mathbf{X}_g$. The low error in reconstruction of one TM using the other is possible due to the common intermediate subspace constructed with CCA.

Measuring the Accuracy of Estimated TMs:

The CCA technique takes two distinct views, *say* $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$ as input, and gives two projection matrices ($\mathbf{W}_t$ and $\mathbf{W}_{gt}$) and two estimated TMs as output, *say* $\hat{\mathbf{X}}_t$ and $\hat{\mathbf{X}}_{gt}$. In this section, we characterize the accuracy of the estimated TMs using the four considered evaluation measures. We also compare with the accuracy obtained for the other techniques like PCA, Fanout, {1}-inverse, and RTME. It



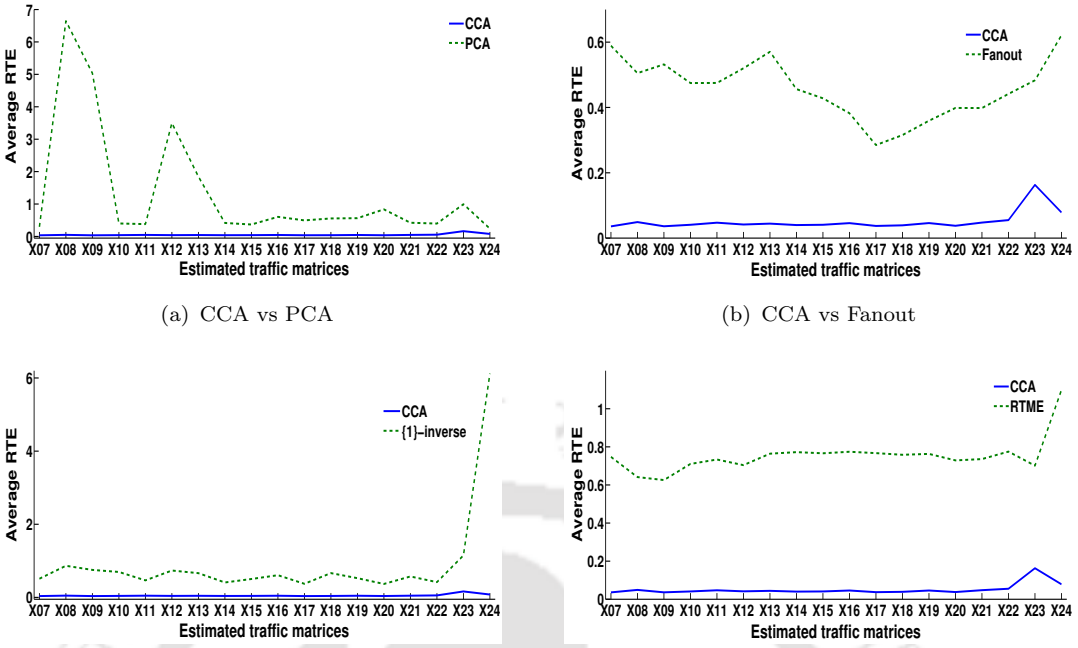
(a) Average RTE (b) Average RSE

FIGURE 3.6: Average RTE and RSE of Abilene Traffic Matrices.

was shown in the literature [15, 10] that few largest OD flows capture most of the energy present in the TM and they are of main interest to network operators, rather than the smaller flows. Hence, we present the temporal and spatial error for these largest OD flows only. For CCA, we sort the OD flows from the input views of TMs in the descending order of their average volume. With this, one can identify canonical directions along which these largest OD flows are maximally correlated. The relative temporal and spatial error for CCA is computed in the intermediate space, while for the other techniques it is computed in the original space.

Fig. 3.6 shows a comparison of the average RTE and average RSE for estimated TMs with tomogravity ($\mathbf{X}_t$), generalized tomogravity ($\mathbf{X}_{gt}$) and for $\mathbf{X}_{gt}$ projected onto the intermediate subspace using CCA. The x-axis indicates the estimated TMs **X07** through **X24** and the y-axis denotes the average RTE and average RSE value. It can be seen from the figure that the TM estimated using CCA exhibits much lower temporal error and spatial error. In particular, TMs estimated using CCA have an average RSE value of less than 0.08, while it is more than 0.26 with tomogravity and generalized tomogravity. Though not presented here, even for the view set $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$, TMs estimated using CCA exhibit least temporal and spatial error. The RTE value is less than 0.20, while the RSE value is less than 0.13.

Fig. 3.7 shows the RTE obtained with CCA compared to other single-view learning techniques; PCA [6], Fanout [78], {1}-inverse [66] and RTME [15] techniques for Abilene TMs (**X07** - **X24**). For PCA and Fanout, we utilize previous week TM to



(a) CCA vs PCA

(b) CCA vs Fanout

(c) CCA vs {1}-inverse

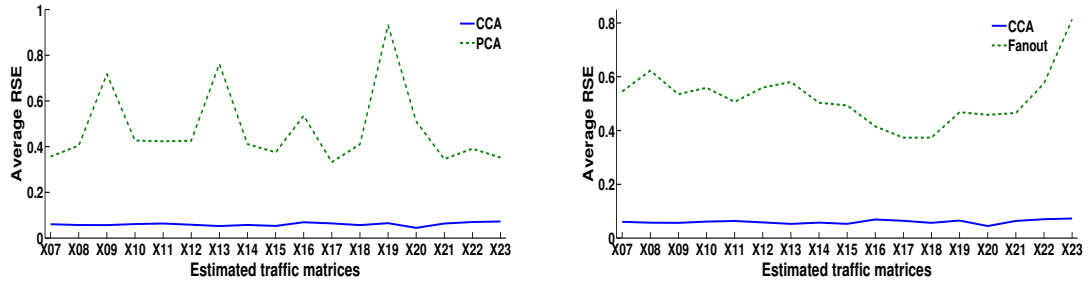
(d) CCA vs RTME

FIGURE 3.7: Average relative temporal error of Abilene TMs.

estimate the next week TM. We did not consider **X06** (May 1^{st} to May 7^{th}, 2004) due to the unavailability of previous week TM, without which PCA and Fanout do not work. CCA exhibits the least average RTE for all TMs, as high as 0.16, while the minimum value with the other techniques is 0.27 (PCA and Fanout), 0.37 ({1}-inverse) and 0.62 (RTME). For most of the cases, CCA gives more than 80% reduction in temporal error, by projecting generalized tomogravity estimates along maximally correlated canonical directions.

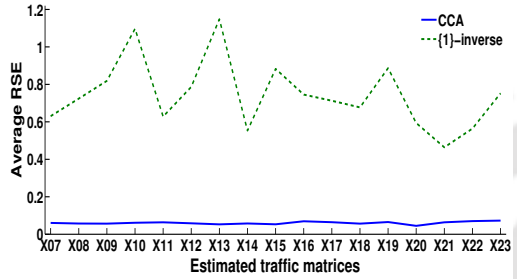
Fig. 3.8 shows the plot of the average RSE. We present the average RSE for 40 largest OD flows of each TM. It can be seen that CCA exhibits the least average RSE compared to the other single-view learning techniques. With CCA, the spatial error is not more than 0.07, while the minimum value of the average RSE for other techniques is: 0.30 (PCA), 0.34 (Fanout), 0.46 ({1}-inverse) and 0.70 (RTME). Experiments on the view set $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$ also show that the TMs estimated using CCA exhibit the least RSE and RTE compared to all the other techniques as well.

To show that there is no bias in the results due to the choice of techniques used for building the views, we conducted experiments with other view combinations as

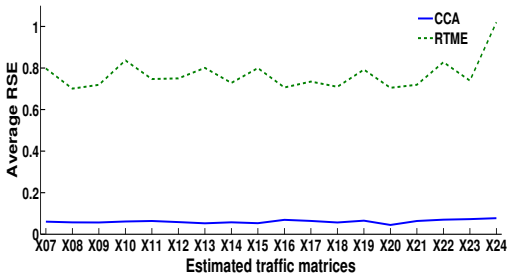


(a) CCA vs PCA

(b) CCA vs Fanout

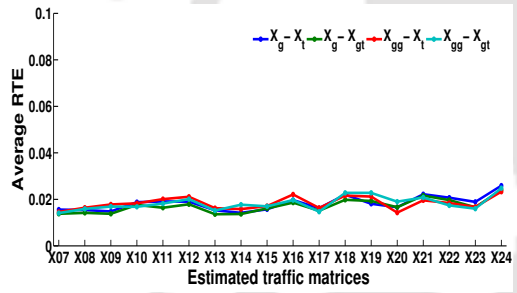


(c) CCA vs {1}-inverse

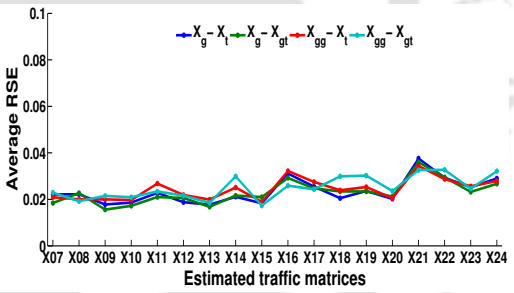


(d) CCA vs RTME

FIGURE 3.8: Average relative spatial error of Abilene TMs.



(a) Average RTE



(b) Average RSE

FIGURE 3.9: Average RTE and RSE of Abilene TMs using CCA on different view combination.

well. Fig. 3.9 shows the average RTE and average RSE value obtained with CCA based on other sets of views. From the results, we noticed that the estimation error with CCA did not largely vary due to the view combination.

For the sake of completeness, we also report the average RSE computed across all the 121 IE flows (not just the strongly correlated ones) for all the 17 weeks of data in Fig. 3.10. We do not report the error for 24^{th} week (TM **X24**), because of a very large volume anomaly (of the order 10^{13} bytes). When we consider all the IE flows, CCA has lower error than {1}-inverse and RTME technique, but slightly

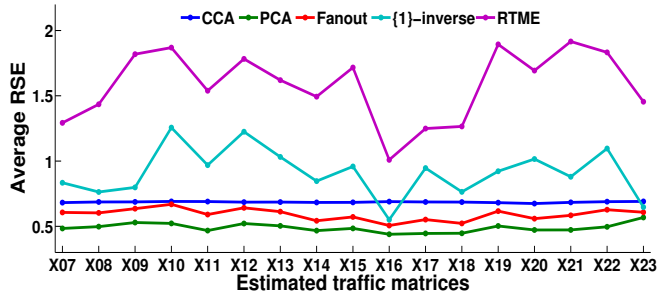
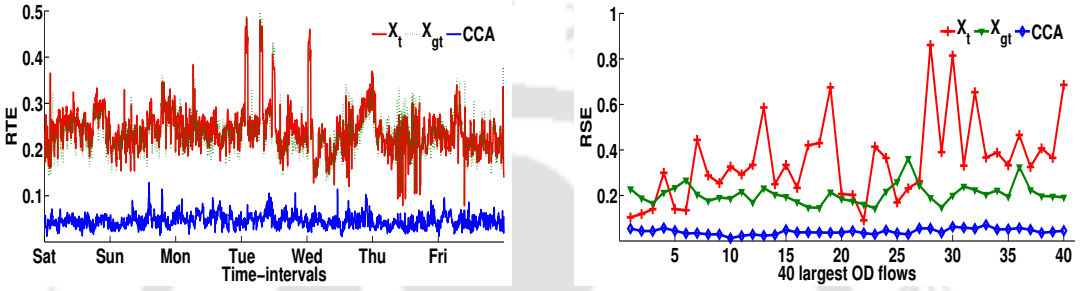


FIGURE 3.10: Average relative spatial error of all flows of Abilene TMs.



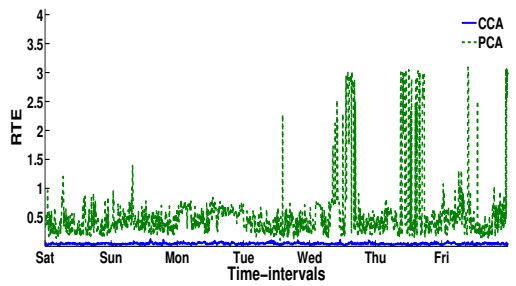
(a) RTE (b) RSE

FIGURE 3.11: RTE and RSE of **X19**.

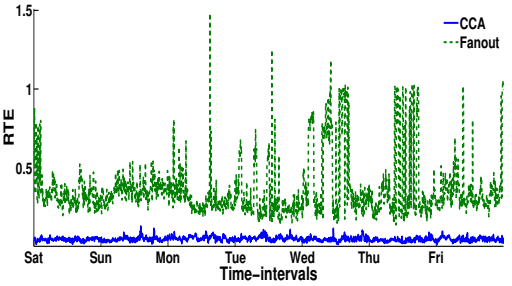
lags behind PCA and Fanout methods. However, it must be noted that most of the techniques employed for TM estimation only consider principal directions.

Detailed analysis of TM X19: For further analysis of the estimation error, we report the detailed results for one TM, **X19** and the view set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$. Fig. 3.11 shows the RTE and RSE for TM **X19** estimated with the techniques of tomogravity ($\mathbf{X}_t$), generalized tomogravity ($\mathbf{X}_{gt}$), and CCA ($\mathbf{X}_{gt}\mathbf{W}_{gt}$). The number of maximally correlated canonical directions (N) in this case is 39. The figure clearly shows that the TM estimated using CCA exhibits a low RTE across all the time-intervals. The average RTE using CCA is 0.05, while it is 0.24 for the others.

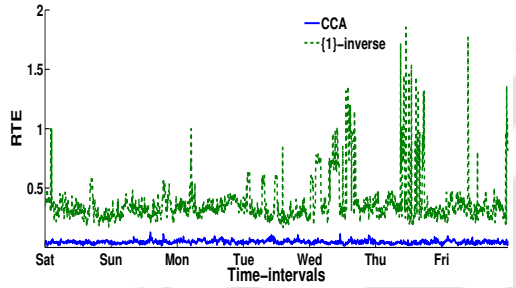
The x-axis of the RSE plot indicates the top 40 largest OD flows arranged in order of decreasing mean value and the y-axis indicates the RSE. We see that CCA has an average RSE of 0.06 for the top 40, while it is 0.35 for tomogravity and 0.36 for generalized tomogravity. It is similar for other TMs estimated with CCA. Experiments on the view set $\{\mathbf{X}_g, \mathbf{X}_{gg}\}$ also show that CCA gives much better estimates of the TMs.



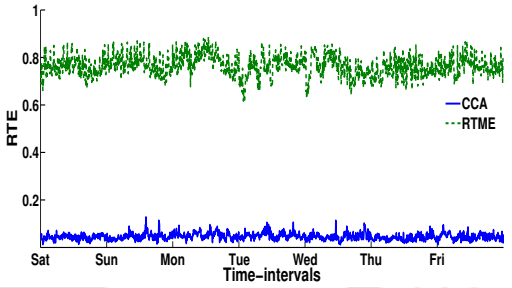
(a) CCA vs PCA



(b) CCA vs Fanout

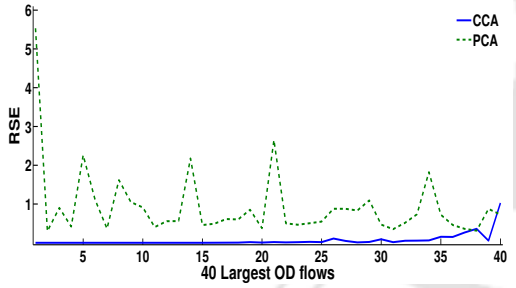


(c) CCA vs {1}-inverse

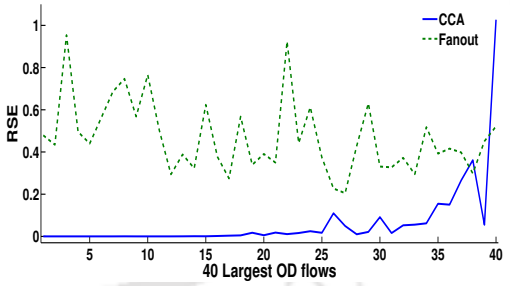


(d) CCA vs RTME

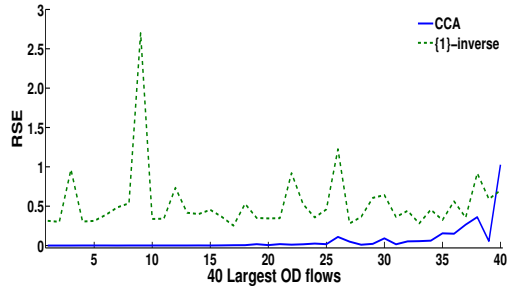
FIGURE 3.12: Relative temporal error of **X19**.



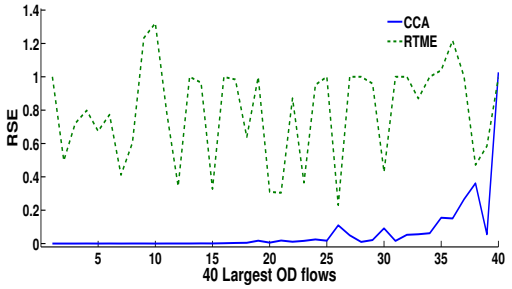
(a) CCA vs PCA



(b) CCA vs Fanout



(c) CCA vs {1}-inverse



(d) CCA vs RTME

FIGURE 3.13: Relative spatial error of **X19**.

Fig. 3.12 compares the RTE for CCA, PCA, Fanout, {1}-inverse, and RTME techniques for the TM **X19**. From the plot, we see that CCA outperforms all the other state-of-the-art single-view learning techniques across all time-intervals. CCA gives the least average RTE of 0.05, while PCA, Fanout and {1}-inverse techniques show high RTE on Wednesday, Thursday and Friday. TMs across all the time intervals estimated with RTME have high RTE. The average temporal error with PCA, Fanout, {1}-inverse and RTME techniques is found to be 0.49, 0.42, 0.38 and 0.76, respectively. Overall, CCA gives more than 80% reduction in temporal error compared to the other single-view learning techniques.

Fig. 3.13 compares the RSE for all the techniques for TM **X19**, which also shows that the CCA gives accurate estimates of the principal OD flows. The average error with CCA is only 0.06, while it is 0.45, 0.44, 0.54, and 0.79 for PCA, Fanout, {1}-inverse, and RTME techniques, respectively. Even in terms of spatial error, there is a reduction of 80% with CCA. We also observe that across all the time intervals, the RTE with CCA is lower than 0.13 and more than 80% of the top 40 OD flows estimated with CCA have RSE lower than 0.06. As discussed earlier, the significant improvement in the accuracy of estimation requires no additional sources of data. We use the TMs estimated with conventional techniques like tomogravity as input for multi-view learning.

Estimation Bias and Variance:

We also study the estimation bias and variance, computed for each OD flow estimated with the proposed technique. We present the results for the view set $\{\mathbf{X}_t, \mathbf{X}_{gt}\}$, but the results for the other view set are also similar. Fig. 3.14 shows the estimation bias in intermediate subspace for 40 largest OD flows of the TM **X19**. The bias values for CCA and {1}-inverse are shown on the left-side y-axis, while for the PCA, Fanout and RTME techniques, they are shown on the right-side y-axis. The x-axis indicates OD flows arranged in order of decreasing mean value. The bias values for CCA and {1}-inverse are found to be very low. For example, for 1^{st} largest OD flow the bias values for CCA, {1}-inverse, PCA, Fanout and RTME are 0, 0, -0.24, -0.20 and -0.11, respectively. This implies that CCA and {1}-inverse give an accurate estimation of the largest OD flow and hence are better

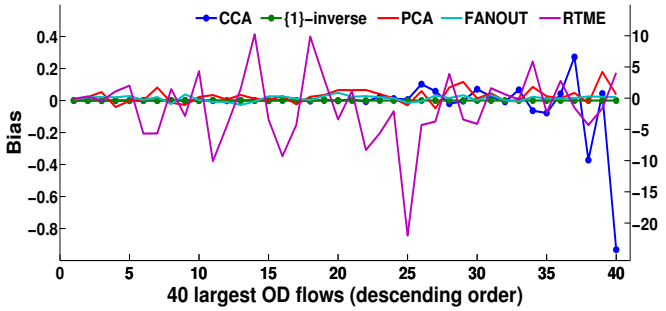


FIGURE 3.14: Plot of estimation bias for CCA and {1}-inverse using left y-axis and for PCA, Fanout and RTME using right y-axis.

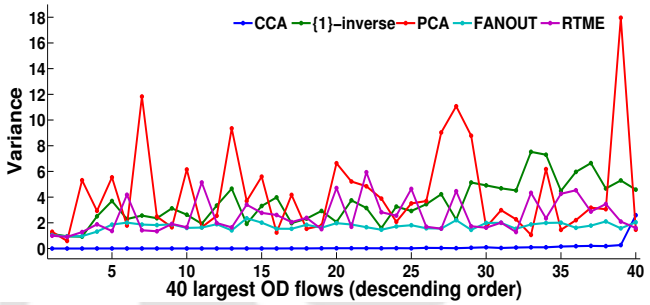


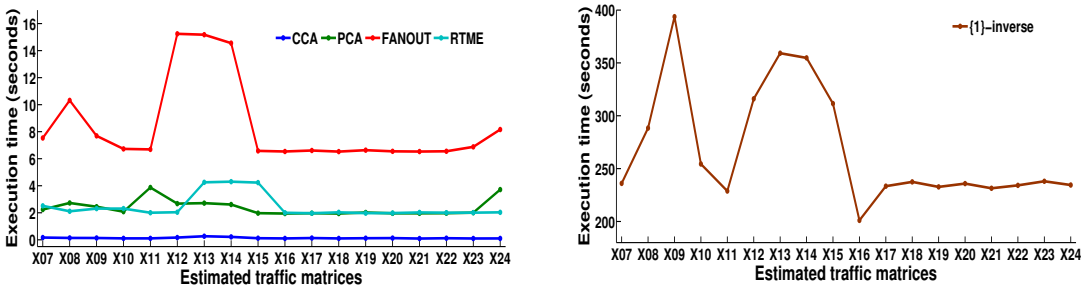
FIGURE 3.15: Variance of estimation error.

for capacity planning, whereas PCA, Fanout, and RTME underestimate the flow volumes.

Fig. 3.15 shows the variance of estimation error in intermediate space for 40 largest OD flows of TM **X19**. We observe that CCA exhibits the least variance in estimation error compared to PCA, Fanout, {1}-inverse, and RTME techniques. CCA shows almost zero variance in estimation error for the top 39 largest OD flows. We conclude that the proposed multi-view learning technique helps to estimate the TMs without bias and variance, a feature that can be very useful in applications like capacity planning and anomaly detection.

3.4.4 Execution Time

Finally, we compared the execution time for different techniques, run on a 1.10 GHz Windows machine with 2GB RAM using MATLAB software. Fig. 3.16 shows the execution time required to estimate the TMs using different techniques, based on Abilene dataset. We plotted the execution time for {1}-inverse separately because



(a) (b)

FIGURE 3.16: Execution time for estimation of Abilene traffic matrices.

of larger values compared to the rest. From Fig. 3.16, we can see that the proposed multi-view subspace learning technique based on CCA takes the least execution time compared to the other single-view learning techniques. In particular, CCA takes less than 0.26 seconds across all the experiments.

To conclude, we propose a multi-view learning technique for TM estimation using canonical correlation analysis. We defined the notion of multiple views of TM as those TMs estimated with inexpensive estimation techniques. We then showed how these views of a TM can be used in CCA for estimating more accurate TMs. Results on the Abilene dataset show that the proposed multi-view learning technique gives accurate estimates of TM, by reducing the temporal and spatial error by more than 80% compared to all the other state-of-the-art techniques. The bias and variance of the estimation error for top OD flows are very close to zero and hence the estimated TMs are robust for use in capacity planning and anomaly detection.

3.5 Summary

The first contribution of the thesis is focused on alleviating the limitations of existing decomposition techniques for traffic matrix estimation and proposes the use of two well-known decomposition techniques for TM estimation. Extensive experimentation is performed using publicly available real traffic matrices obtained

from the Abilene network. We have compared the proposed techniques with state-of-the-art techniques in terms of four evaluation measures, namely RTE, RSE, bias, and variance. We observe that proposed decomposition techniques outperforms other techniques in terms of the evaluation measures and provides a more accurate estimate of the traffic matrix. Besides, we do not make use of partial direct measurements for estimation like the third generation techniques, which is a costly affair to obtain. Thus, the proposed decomposition techniques provide an *improved estimation of Internet traffic data* with the least estimation error.

Chapter 4

Reconstruction of Internet Traffic Data

This chapter presents the second contribution of the thesis which proposes the use of matrix-CUR decomposition and relative error-bound tensor-CUR decomposition for reconstruction of missing values in Internet traffic data.

4.1 Introduction

Measuring Internet traffic volume data between Origin-Destination (OD) pairs (*say* in bytes/sec) is of paramount importance in several network engineering tasks such as, capacity planning, load balancing, capacity provisioning, anomaly detection, and fault-tolerance [60]. Though the knowledge of traffic data is very important, measuring the same at a large number of vantage points, at finer time scales is impractical and often a costly affair. In practice, the traffic data is represented as a matrix, commonly termed traffic matrix (TM), where the rows indicate time intervals, the columns indicate the pairs of nodes (points of presence *aka* PoPs), and $(i, j)^{th}$ entry represents the volume of data between the j^{th} node pair during the i^{th} time interval [60]. To reduce the cost of collecting TM data, random samples of traffic data are collected by the monitoring system across the network links

from which the complete data is estimated [85, 6, 86]. However, the data collection can be affected by network congestion and temporary failures of nodes/links. Due to the importance of accurate TMs, recovery of missing values and accurate reconstruction of TMs have been widely studied in the literature [9].

Compressive Sensing (CS) is a generic methodology for the reconstruction of missing values, that takes advantage of the presence of a certain type of structure in the data. Roughan *et al.* [9] identified some limitations in borrowing existing CS techniques, normally applied in other domains, to estimate the missing elements in traffic volume data. These are mainly due to large variations in the traffic volume, structured loss pattern in the data, and the low-rank nature of the traffic matrix.

At the heart of TM estimation and the other applications, that decompose the TMs are Singular Value Decomposition (SVD)-based matrix decomposition techniques and their variants. Though widely used, the representation of Internet traffic data as TMs has several fundamental issues. First, the traffic volume is measured at designated nodes (either PoPs or intermediate vantage points) and not as represented in a TM. Second, the measured traffic volume is at a certain time interval and thus, the data has an inherent multidimensional structure, not captured in a TM. Acar *et al.* [87] identified another fundamental problem with the matrix representation of traffic data, that the uniqueness of factor matrices cannot be guaranteed in the matrix decomposition methods. Finally, it is well-known that matrix-based algorithms cannot perform well in the recovery of missing traffic data when the data loss ratio is large, especially when the traffic data on several consecutive intervals is lost [60, 58].

To overcome these limitations, traffic volume data is modeled as a three-dimensional traffic tensor (TT) [87, 58]. Tensor representation would be a more natural representation of Internet traffic data as it would theoretically preserve the spatio-temporal properties better than stacked matrices, as well as capture the evolution of traffic during the time interval of measurement. For the Internet traffic data, the natural choice of dimensions for tensor representation are; ingress node, egress

node, and time-interval. Modeling traffic data using tensors has the following advantages: (i) uniqueness is guaranteed in the tensor decomposition methods [59] without imposing constraints such as orthogonality and independence, as done in case of matrix decomposition and (ii) the multi-dimensional structure of the traffic data is preserved without pairing two distinct dimensions (ingress and egress nodes) as a column of the TM.

Motivation: Just as in the case of TM estimation, it is also difficult to obtain accurate tensor data for Internet traffic due to missing measurements. Few works in the recent times have addressed the problem of accurate recovery of tensor data in various application domains, including compressive sensing of Internet traffic data and anomaly detection [88, 56]. In this work, we explore an alternate decomposition technique that helps accurate reconstruction of missing values in Internet traffic data using the tensor representation.

Various tensor decomposition techniques are used to reconstruct the missing values, that are predominantly based on CANDECOMP/PARAFAC (CP) decomposition [87, 17, 89] and Tucker decomposition [54]. These decomposition techniques obtain a low-rank approximation of tensors (input) by decomposing it in terms of three factor matrices (output). We discuss below two main limitations of the CP and Tucker decomposition techniques, that are widely used for reconstruction of missing values in the tensor. We introduce an alternate decomposition technique that avoids these limitations.

1. *Apriori knowledge of tensor rank*: The CP decomposition expresses the tensor as a sum of k rank-one tensors, where k denotes the rank of a tensor which acts as an important input parameter. However, determining the tensor rank is an NP-hard problem [18], and there is no straightforward algorithm to determine the rank. Similarly, Tucker decomposition decomposes n-dimensional tensor as a product of a core tensor and n factor matrices. Here *n-rank*, *i.e.*, the rank of matricized form of tensor, is an important input parameter for Tucker decomposition. However, determining suitable

values for *n-rank* apriori is also not feasible [90]. In both these decomposition techniques, knowing the rank apriori poses a fundamental limitation for their use in the reconstruction of missing values. Hence, all the recent works in [87, 56, 58] have the same limitation.

2. *Usage of alternating least squares (ALS) procedure*: Both CP and Tucker decomposition techniques employ the alternating least squares (ALS) procedure to solve the optimization framework required to obtain the factor matrices. The ALS procedure starts with random initialization of the three factor matrices. The optimization problem is then solved by designating one of them to be an optimization variable and the process is repeated for the other factor matrices, till convergence criteria is satisfied. In the ALS procedure, the factor matrices are sensitive to the method of initialization. Furthermore, the time of convergence depends on the matrix initialization because ALS decomposes the approximated tensor in each iteration till convergence. *See Section 4.5.3 for results that show this aspect.*

In addition, we also observed that the method used in [9] and the variants thereof, perform the global interpolation using decomposition technique followed by a local interpolation procedure. Extending such a procedure for tensor/matrix decomposition has a consequence that the reconstructed tensor/matrix is not exactly the one obtained from the factor matrices. This would also defeat the very purpose of tensor/matrix decomposition.

4.1.1 Main Contributions

In this work, we demonstrate the sensitivity of the ALS procedure in traffic tensor decomposition techniques, which has not been done in the literature till date. Since all the decomposition techniques proposed so far use the ALS procedure, we specifically design a tensor decomposition method that does not use the ALS procedure to obtain the factor matrices. The major contributions of this chapter are as follows:

- We study the sensitivity of factor matrix initialization in the ALS procedure on the error of reconstruction in traffic tensors and the computational time required for it.

- We introduce a new decomposition technique, relative-error bound tensor-CUR decomposition (TCUR-REB) for the reconstruction of missing values in Internet traffic tensor. TCUR-REB decomposition does not depend on the ALS procedure and does not require to know the rank of the tensor apriori. It also has a better bound on the low-rank approximation compared to the standard additive-error bound TCUR decomposition.

- We also propose to use matrix-CUR decomposition for the reconstruction of missing values in the traffic matrix. Matrix-CUR decomposition has an improved time complexity over SVD-based techniques.

- We reversed the order of combining global and local interpolation procedure. We first propose to use a local interpolation procedure, termed *zeroth-order interpolation* followed by a global interpolation procedure based on the proposed tensor/matrix decomposition technique. This ensures that the reconstructed traffic tensor/matrix is exactly the product of the factor matrices.

4.1.2 Summary of Experimental Results

We evaluated the proposed compressive sensing techniques on two publicly available Internet traffic datasets collected from GEANT [8] and Abilene [7] networks. We summarize the following from this evaluation:

1. For traffic tensor decomposition, the factor matrix initialization in the ALS procedure causes a large variation in the error of reconstructing the missing values and the convergence time for different loss rates.

2. TCUR-REB outperforms the state-of-the-art tensor decomposition techniques in terms of the error in the reconstruction of missing values and also has the

least computation time among the other techniques and hence it is *computationally inexpensive*.

3. Matrix-CUR decomposition outperforms the state-of-the-art matrix decomposition techniques in terms of the error in the reconstruction of missing values and computational time involved.

4. Changing the order of interpolation leads to zero error in the reconstructed tensor from its factors as opposed to the order proposed in [9, 17].

4.2 Related Work

In this section, we review literature that uses matrix or tensor representation for traffic data and highlight the limitations, thereby serving as motivation for the proposed contribution.

4.2.1 Compressive Sensing of Internet Traffic Matrix

Matrix decomposition techniques aim to find a low-rank approximation of traffic matrix to reconstruct missing values. Singular Value Decomposition (SVD) forms the basis for a variety of techniques employed in literature for compressive sensing of traffic matrices. As discussed in chapter 2, SVD decomposes TM $\mathbf{X}$ as $\mathbf{X} = \mathbf{ESV}^{\mathsf{T}}$. Sparsity Regularized SVD (SRSVD) [9] employs SVD on traffic matrix and obtains factorization of the form $\mathbf{X} = \bar{\mathbf{L}}\bar{\mathbf{R}}^{T}$, where $\bar{\mathbf{L}} = \mathbf{ES}^{1/2}$ and $\bar{\mathbf{R}} = \mathbf{VS}^{1/2}$, such that the product $\bar{\mathbf{L}}\bar{\mathbf{R}}$ exhibit low rank and satisfy traffic measurement equations. While the decomposed factor matrices in SRSVD can contain negative values, Nonnegative Matrix Factorization (NMF) [91] makes sure that decomposed factor matrices $\bar{\mathbf{L}}$ and $\bar{\mathbf{R}}$ contain nonnegative values only and obtains a similar factorization. Sparsity Regularized Matrix Factorization (SRMF) [9], on the other hand, besides employing SRSVD also takes into account the temporal and spatial properties of traffic matrices. It has been demonstrated that SRMF outperforms NMF, SRSVD, and SRSVD-base.

Recovery of the missing data from partial traffic measurements has also been carried out in [92, 93, 94, 95, 96]. Chen *et al.* [95] developed a general Compressive Sensing framework (LENS) which decompose a traffic matrix into four components: a Low-rank matrix, an Error matrix, a Noise matrix, and a Sparse anomaly matrix (hence the acronym LENS). This decomposition incorporates domain knowledge, such as temporal locality, spatial locality, and initial estimate.

Limitations: SVD-based matrix decomposition techniques employed for compressive sensing of traffic matrix make use of ALS procedures to obtain output factor matrices $\bar{\mathbf{L}}$ and $\bar{\mathbf{R}}$. ALS-based procedures are computationally expensive. This is because it performs the matrix decomposition at each iteration of the optimization. Thus, there is a need for a matrix-decomposition that avoids the use of ALS procedures and is computationally inexpensive.

4.2.2 Compressive Sensing of Internet Traffic Tensor

Tensor decomposition techniques have gained widespread attention over the last decade for various applications including recommender systems [51], image analysis [52], computer vision [53], and intelligent transportation systems [54, 55]. More recently, tensors of various dimensions have been utilized to represent Internet traffic data for anomaly detection [56], [57] and compressive sensing [58]. Below we present the techniques employed in literature for compressive sensing of Internet traffic tensors followed by their limitations.

CP decomposition: The work in [87] proposed a method based on CP weighted optimization (CPWOPT) to reconstruct the missing values in a tensor. Asif *et al.* [55] have also used CPWOPT for compressive sensing of traffic tensor of transportation data from a large road network. Zhou *et al.* [17] have employed Spatio-Temporal Tensor Completion (STTC) for the reconstruction of missing values in traffic tensor, that relies on CP decomposition along with considering the spatio-temporal constraints. The central approach in these works is to minimize the difference between the reconstructed traffic tensor and the original traffic tensor

with the constraints that factor matrices are of low norm. They demonstrated that STTC achieves promising accuracy over matrix-based methods and CPWOPT, for a wide range of loss patterns with varying loss rates.

Limitations: To obtain a low-rank approximation, the CP decomposition requires the knowledge of tensor rank apriori. Determining the tensor rank is challenging and it is an NP-hard problem [18]. Furthermore, CP decomposition based approaches employ ALS procedure to obtain the factor matrices, which is sensitive to initialization of factor matrices.

Tucker decomposition: Tan *et al.* [54] proposed a Tucker decomposition based approach to estimate the missing values in traffic tensor of the transportation information system data. Authors of [58] proposed a Sequential Tensor Completion (STC) algorithm based on incremental SVD so that the tensor can be decomposed for the current traffic data using the tensor decomposition result of the previous traffic data. This is an incremental algorithm, whereas the approach taken in our work performs tensor completion using the decomposition result of the same tensor data. The low rank (rank-k) approximation of a matrix using PCA is obtained by retaining few (k) components of the factor matrices. However, in the case of Tucker decomposition which is viewed as a higher-order generalization of PCA, one cannot obtain a rank-k approximation of a tensor by retaining k components in the factor matrices [90].

Limitations: The n-rank is an important input parameter for Tucker decomposition. Determining a suitable truncation value of n-rank apriori is again hard [90]. To obtain the factor matrices, Tucker decomposition based approaches also employ ALS procedure which is sensitive to initialization of factor matrices, as we would show in Section 4.5.3.

Tensor-CUR decomposition: Tensor-CUR decomposition is used in [97, 98] to reconstruct the missing values in a tensor. In [97], the tensor $\boldsymbol{\mathcal{C}}$ is constructed by selecting c frontal slices with highest Frobenius norm and the matrix $\mathbf{R}$ is constructed by selecting r tubes with the highest Euclidean norm. As a result of this selection strategy, it has been demonstrated in [97] that given a tensor

$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, a preferred mode $\alpha \in \{1, 2, 3\}$, a rank parameter k, an error parameter ϵ, and a failure probability $\delta \in (0, 1)$, the following condition holds

$$\|\boldsymbol{\mathcal{X}} - \tilde{\boldsymbol{\mathcal{X}}}\|_F \leq \|\mathbf{X}_{[\alpha]} - (\mathbf{X}_{[\alpha]})_k\|_F + \epsilon\|\boldsymbol{\mathcal{X}}\|_F \tag{4.1}$$

with probability at least 1 - δ, where $\tilde{\boldsymbol{\mathcal{X}}}$ denotes the tensor reconstructed using tensor-CUR decomposition, $(\mathbf{X}_{[\alpha]})_k$ denotes the best rank-k approximation of mode-α matricized form of $\boldsymbol{\mathcal{X}}$ obtained using Higher-Order SVD. The left-hand side (L.H.S.) of Eq. (4.1) signifies the error in reconstruction using tensor-CUR decomposition, which is computed by the Frobenius norm of difference between original tensor ($\boldsymbol{\mathcal{X}}$) and reconstructed tensor ($\tilde{\boldsymbol{\mathcal{X}}}$), and the right-hand side (R.H.S.) denotes the error incurred using HOSVD plus an additive factor. Eq. (4.1) simply states that the error incurred in reconstruction using tensor-CUR decomposition is always less than or equal to the error incurred in reconstruction using best rank-k approximation of HOSVD plus an additive factor. This method is called additive-error bound tensor-CUR (TCUR-AEB) decomposition. Interested readers may refer to a recent survey presented in [99] which details various state-of-the-art tensor decomposition methods applied to data mining.

Limitations: Bounds of this form are generally weaker because the additive error term is usually large in practice. It is suitable either for large-sized data or for a small percentage of missing data or when limited resources are available [13]. However, if the input tensor is exactly low rank, the left-hand side will become zero and the first term on the right-hand side will also be zero, while the additive-error term will remain non-zero. Thus, additive-error bound fails to reflect the case of perfect reconstruction in case the input tensor is exactly low-rank. Another limitation of the additive-error bound tensor-CUR decomposition is the need to have access to the complete tensor for the accurate decomposition of the given tensor.

Other Applications of Internet Traffic Tensor:

- Anomaly detection: Sun *et al.* [88] have used the tensor representation for

Internet application traffic data using four dimensions; ingress node, egress node, socket port number, and time-interval. They proposed dynamic and streaming tensor analysis for anomaly detection to identify anomalous traffic from normal traffic. Kim *et al.* [100] have arranged the traffic tensor as a three-dimensional traffic tensor using the dimensions; time-interval, ingress node, and egress node. They used a higher-order analogue of PCA for network anomaly detection. It has been shown that tensor decomposition techniques outperform the matrix decomposition techniques in terms of the percentage of false negatives and false positives. In [56], the authors used truncated HOSVD to locate the anomalies in noisy tensor data. The approach taken there is, however, an incremental one, where the historical data is used to reconstruct the data in current tensor.

- Low-rank approximation: Kasai *et al.* [57] have arranged the traffic tensor as a three-dimensional traffic tensor and utilized CP decomposition to obtain a low-rank approximation of traffic tensor, which in turn is used to model the normal traffic and separate anomalies.

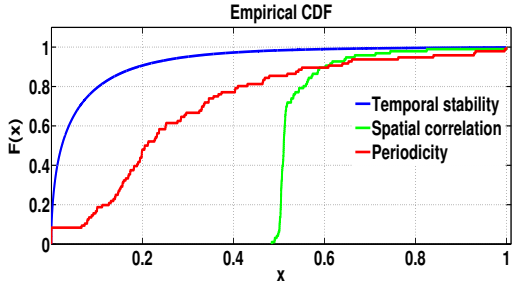
In summary, only the recent works in [87, 58], demonstrated the power of tensor decomposition techniques for compressive sensing of Internet traffic tensor data. It is very clear that tensor representation and tensor decomposition techniques have a promising future in the context of the reconstruction of missing values in Internet traffic data. However, all the works till date use either CP or Tucker decomposition, both of which use the ALS procedure, that is highly sensitive to the selection of initial factor matrices. These decomposition techniques also require apriori knowledge of tensor rank. We propose a compressive sensing technique that uses an alternate decomposition technique, termed TCUR-REB, that avoids all the limitations mentioned earlier. It also gives a stronger bound compared to the well-known TCUR-AEB decomposition.

4.3 Choice of Traffic Tensor Representation

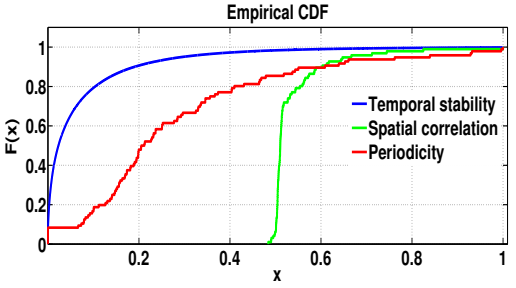
Though there is no unique way to represent the Internet traffic data using tensors, most of the works modeled the tensor using the three dimensions; ingress node, egress node, and time-intervals (which are the natural dimensions used to collect the data). However, the authors of [58, 56] have argued that such a representation does not capture the temporal stability, spatial correlation, and periodicity, fairly well. Hence, they constructed the tensor using the dimensions; ingress-egress node pairs, time-intervals, and days. In this section, we first demonstrate that either of these two representations for the tensor captures the similarity features fairly well, to avoid the controversy associated with the dimensions chosen for the tensor representation.

The work in [58, 56], measured temporal stability by taking the normalized difference of traffic volume during successive time-intervals across a given ingress-egress node pair. If the difference is low, traffic flows are considered to be more stable. The spatial correlation between two pairs of nodes corresponds to the correlation between the time-series of traffic flows. If the correlation is high, the traffic is considered to be spatially correlated. Periodicity is measured by taking the difference of traffic volume during the same time-intervals of successive days across a given pair of nodes. If the difference is low, the traffic is considered to be periodic.

Fig. 4.1 shows the cumulative distribution frequency (CDF) of the three similarity features using two different representations of GEANT traffic data (similar representation was used for the Abilene dataset also). We refer *Representation 1* as $(N \times N \times T)$, where N denotes the number of ingress/egress nodes and T denotes the number of time-intervals and *Representation 2* as $(O \times T \times D)$, where O, T and D denote the number of ingress-egress node pairs, time-intervals, and days, respectively. It can be clearly seen from the two plots, that they are exactly identical, irrespective of the representation. Additionally, we observe the following for GEANT dataset:



(a) CDF of similarity features using representation 1: ingress node, egress node, and time-intervals.



(b) CDF of similarity features using representation 2: ingress-egress node pairs, time-intervals, and days.

FIGURE 4.1: Cumulative Distribution Frequency (CDF) of the three similarity features.

1. For more than 90% of time-intervals, the normalized difference is smaller than 0.2 during successive time-intervals across all the node pairs.

2. More than 97% of the node pairs exhibit a spatial correlation of more than 0.5.

3. More than 60% of the time-intervals exhibit a normalized difference of less than 0.3 during successive days.

Similarly, for the Abilene traffic tensor (plots for which are not shown here), we observe the following:

1. For more than 75% of time-intervals, the normalized difference is smaller than 0.1 during successive time-intervals across all the node pairs.

2. All the node pairs (i.e., 100% node pairs) exhibit a spatial correlation of more than 0.5.

3. More than 55% of the time-intervals exhibit normalized difference of less than 0.3 during successive days.

From the above observations we note that $(N \times N \times T)$ retains temporal stability, spatial correlation, and periodicity. Hence, without loss of generality, we use the representation of the form $(N \times N \times T)$, for the tensor in our work. Incidentally, it is the natural way in which the data is collected by the network operators.

4.4 Proposed Decomposition Techniques

In this section, we present the TCUR-REB for the reconstruction of missing values in traffic tensor to alleviate the limitations concerning apriori knowledge of tensor rank and use of ALS-based procedures. In addition, we present the matrix-CUR decomposition for the reconstruction of missing values in the traffic matrix to alleviate the limitation concerning the use of ALS-based procedures.

4.4.1 Relative-error Bound Tensor-CUR (TCUR-REB) Decomposition

In this chapter, we propose a tensor-CUR decomposition technique that has an improved relative error bound on the reconstruction compared to the additive error bound proposed in [97]. We explain the four steps involved in the decomposition (also illustrated in the bounding box, for quick reference). The entire process is summarized in Algorithm 3.

$$\overbrace{\mathbf{X}_{[3]} \in \mathbb{R}^{I_3 \times I_1 I_2} \leftarrow \text{Mode-3 matricization of } \boldsymbol{\mathcal{X}}}^{\textbf{Step 1}}$$

$$\tilde{\boldsymbol{\mathcal{X}}} = \underbrace{\boldsymbol{\mathcal{C}}}_{\textbf{Step 2}} \times_3 \overbrace{\mathbf{U}}^{\textbf{Step 4}} \times_3 \underbrace{\mathbf{R}}_{\textbf{Step 3}} . \tag{4.2}$$

Step 1: *Mode-3 Matricization*: Three dimensional tensor $\boldsymbol{\mathcal{X}}$ should first be transformed into a matrix form, *i.e.*, two dimensional tensor, for obtaining $\boldsymbol{\mathcal{C}}$ and $\mathbf{R}$. For this, the traffic tensor should be subject to mode-3 matricization. Here, $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is represented as $\mathbf{X}_{[3]}$ of size $(I_3 \times I_1 I_2)$, as shown in line 1 of Algorithm 3. The frontal slices of tensor $\boldsymbol{\mathcal{X}}$ are the rows of $\mathbf{X}_{[3]}$ and tubes of tensor $\boldsymbol{\mathcal{X}}$ are the columns of $\mathbf{X}_{[3]}$. In our application domain,

the rows and columns of $\mathbf{X}_{[3]}$ denote the time-intervals and node pairs, respectively. As evident from Eq. (4.2), the next two important steps are to obtain the tensor $\boldsymbol{\mathcal{C}}$ and matrix $\mathbf{R}$.

Step 2: *Selection of Frontal Slices*: After the matricization step, the rows of the matrix $\mathbf{X}_{[3]}$ are the frontal slices of the tensor. Therefore, selection of frontal slices of tensor $\boldsymbol{\mathcal{X}}$ is simply the selection of rows of $\mathbf{X}_{[3]}$, as shown in line 2 of Algorithm 3. This is done by employing column selection strategy (COLUMNSELECT) using Algorithm 2 [13] (presented in chapter 2) with the following as input: $\mathbf{X}_{[3]}^{\mathsf{T}}$, rank parameter k, number of frontal slices that one wishes to retain c' and an error parameter ϵ. Algorithm 2 selects c original columns of input matrix based on normalized statistical leverage score such that, the number of columns selected is less than c', *i.e.*, $\mathbb{E}[c] \leq c'$. A randomized approach is used to select the columns based on the computed normalized leverage score for each column. A number is drawn between 0 and 1 from a uniform distribution. If this number is greater than the normalized leverage score of a column, that particular column is selected. The algorithm COLUMNSELECT gives c ($c \ll I_3$) original time-intervals with high leverage scores, that are rearranged as a three-dimensional tensor $\boldsymbol{\mathcal{C}}$ of size $(I_1 \times I_2 \times c)$ in line 3 of Algorithm 3.

Step 3: *Selection of Tubes*: After the matricization step, the columns of the matrix $\mathbf{X}_{[3]}$ contain the tubes of the tensor. Therefore, the selection of tubes of tensor $\boldsymbol{\mathcal{X}}$ is simply the selection of columns of $\mathbf{X}_{[3]}$, shown in line 4 of Algorithm 3. This is done by employing COLUMNSELECT with the input of $\mathbf{X}_{[3]}$, the rank parameter k, number of tubes that one wishes to retain r' and an error parameter ϵ. This step returns the indices of r node pairs with the highest leverage scores. These ingress-egress node pairs form the matrix $\mathbf{R}$ of size $(I_3 \times r)$ with the tubes of original tensor $\boldsymbol{\mathcal{X}}$ indexed by selected node pairs.

Step 4: *Obtaining matrix* $\mathbf{U}$: Finally, given $\mathbf{C}_{[3]}$, $\mathbf{X}_{[3]}$ and $\mathbf{R}$, the matrix $\mathbf{U}$ is obtained as

$$\mathbf{U} = \mathbf{R}^{+}\mathbf{X}_{[3]}\mathbf{C}_{[3]}^{+}$$

ALGORITHM 3: RELATIVE-ERROR BOUND TENSORCUR (TCUR-REB)

Input: $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, rank parameter k, number of frontal slices c' that one wish to select from $\boldsymbol{\mathcal{X}}$, number of tubes r' that one wish to select from $\boldsymbol{\mathcal{X}}$, and error parameter ϵ.

Output: $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{I_1 \times I_2 \times c}$ s.t. $\mathbb{E}[c] \leq c'$, $\mathbf{R} \in \mathbb{R}^{I_3 \times r}$ s.t. $\mathbb{E}[r] \leq r'$, and $\mathbf{U} \in \mathbb{R}^{r \times c}$.

$\mathbf{X}_{[3]} \in \mathbb{R}^{I_3 \times I_1 I_2} \leftarrow$ Mode-3 matricization of $\boldsymbol{\mathcal{X}}$

$\mathbf{C} \in \mathbb{R}^{I_1 I_2 \times c} \leftarrow$ COLUMNSELECT($\mathbf{X}_{[3]}^{\mathsf{T}}, k, c', \epsilon$)

$\boldsymbol{\mathcal{C}} \in \mathbb{R}^{I_1 \times I_2 \times c} \leftarrow \mathbf{C} \in \mathbb{R}^{I_1 I_2 \times c}$

$\mathbf{R} \in \mathbb{R}^{I_3 \times r} \leftarrow$ COLUMNSELECT($\mathbf{X}_{[3]}, k, r', \epsilon$)

$\mathbf{U} \leftarrow \mathbf{R}^{+}\mathbf{X}_{[3]}\mathbf{C}_{[3]}^{+}$;

return $\boldsymbol{\mathcal{C}}$, $\boldsymbol{R}$, *and* $\boldsymbol{U}$

It was shown in [13] that selection strategy based on normalized statistical leverage scores results in relative-error bound on the reconstruction. According to [13], matrix-CUR decomposes an input matrix $\mathbf{X}$ as a product of $\mathbf{C}$, $\mathbf{U}$ and $\mathbf{R}$ where $\mathbf{C}$ and $\mathbf{R}$ have few original columns and rows of $\mathbf{X}$, respectively. $\mathbf{U}$ is constructed in such a way that the following holds with a probability of at least 0.98.

$$\|\mathbf{X} - \mathbf{CUR}\|_F \leq (2+\epsilon)\|\mathbf{X} - \mathbf{X}_k\|_F \tag{4.3}$$

where $\mathbf{X}_k$ is the best rank-k approximation of $\mathbf{X}$ obtained using PCA.

By extending the same principle, we state in the following theorem, that the selection of frontal slices and tubes according to statistical leverage scores results in the relative-error bound on reconstruction error.

Theorem 4.4.1. Given a tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$, a rank parameter k, a preferred mode α (mode-3), an error parameter ϵ, and failure probability $\delta \in (0,1)$, construct $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{I_1 \times I_2 \times c}$ using COLUMNSELECT on $\mathbf{X}_{[3]}^{\mathsf{T}}$, $\mathbf{R} \in \mathbb{R}^{I_3 \times r}$ using COLUMNSELECT on $\mathbf{X}_{[3]}$, and $\mathbf{U} \in \mathbb{R}^{r \times c}$ as $\mathbf{U} = \mathbf{R}^{+}\mathbf{X}_{[3]}\mathbf{C}_{[3]}^{+}$, where $.^{+}$ denotes a Moore-Penrose generalized inverse of the matrix. Then, the following holds with a probability of at least 0.98

$$\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_3 \mathbf{U} \times_3 \mathbf{R}\|_F \leq (2+\epsilon)\|\mathbf{X}_{[3]} - (\mathbf{X}_{[3]})_k\|_F \tag{4.4}$$

where $(\mathbf{X}_{[3]})_k$ denotes the best rank-k approximation of mode-3 matricized form of $\boldsymbol{\mathcal{X}}$ obtained using HOSVD.

Proof 4.4.1. Matricization of tensor $\boldsymbol{\mathcal{X}}$ along any mode does not change the value of its Frobenius norm, hence the left hand side of Eq. (4.4) can be rewritten as

$$\|\boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{C}} \times_3 \mathbf{U} \times_3 \mathbf{R}\|_F = \|\mathbf{X}_{[3]} - (\boldsymbol{\mathcal{C}} \times_3 \mathbf{U} \times_3 \mathbf{R})_{[3]}\|_F. \tag{4.5}$$

which in turn can be expressed as

$$\|\mathbf{X}_{[3]} - (\boldsymbol{\mathcal{C}} \times_3 \mathbf{U} \times_3 \mathbf{R})_{[3]}\|_F \leq (2+\epsilon)\|\mathbf{X}_{[3]} - (\mathbf{X}_{[3]})_k\|_F. \tag{4.6}$$

The resulting Eq. (4.6) is in coherence with Eq. (4.3), the proof for which is as follows. It was shown in [48] that selection of columns using statistical leverage scores satisfies

$$\|\mathbf{X} - \mathbf{P}_c\mathbf{X}\|_F \leq (1+\epsilon/2)\|\mathbf{X} - \mathbf{X}_k\|_F \tag{4.7}$$

with probability at least 0.99, where $\mathbf{P}_c = \mathbf{C}\mathbf{C}^+$ denotes a projection matrix onto the column space of $\mathbf{C}$. CUR decomposition is limited to selection of columns only in some application and obtains decomposition as $\mathbf{X} \approx \mathbf{P}_c\mathbf{X}$. In other applications, CUR decomposition obtains decomposition in terms of columns and rows simultaneously and obtains a relative-error bound as given in Eq. (4.3).

Since, $\mathbf{U} = \mathbf{C}^+\mathbf{X}\mathbf{R}^+$, the left-hand side of Eq. (4.3) can be written as

$$\|\mathbf{X} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F = \|\mathbf{X} - \mathbf{C}\mathbf{C}^+\mathbf{X}\mathbf{R}^+\mathbf{R}\|_F.$$

After addition and subtraction of $\mathbf{C}\mathbf{C}^+$ and application of triangle inequality for the Frobenius norm, the L.H.S becomes

$$\begin{aligned}
\|\mathbf{X} - \mathbf{C}\mathbf{U}\mathbf{R}\|_F &\leq \|\mathbf{X} - \mathbf{C}\mathbf{C}^+\mathbf{X}\|_F + \|\mathbf{C}\mathbf{C}^+\mathbf{X} - \mathbf{C}\mathbf{C}^+\mathbf{X}\mathbf{R}^+\mathbf{R}\|_F \\
&\leq \|\mathbf{X} - \mathbf{C}\mathbf{C}^+\mathbf{X}\|_F + \|\mathbf{X} - \mathbf{X}\mathbf{R}^+\mathbf{R}\|_F \\
&= \|\mathbf{X} - \mathbf{P}_c\mathbf{X}\|_F + \|\mathbf{X} - \mathbf{X}\mathbf{P}_r\|_F
\end{aligned} \tag{4.8}$$

The second inequality in Eq. (4.8) holds because the projection matrix $\mathbf{C}\mathbf{C}^{+}$ does not change the Frobenius norm. $\mathbf{P}_r = \mathbf{R}^{+}\mathbf{R}$ is a projection matrix onto row space of $\mathbf{R}$. Matrix-CUR decomposition [13] involves two applications of COLUMNSELECT resulting in Eq. (4.3). TCUR-REB also involves two applications of COLUMNSELECT given by Eq. (4.7), first for selection of frontal slices and second for selection of tubes which results in Eq. (4.4). □

Note that in Eq. (4.4) there is no additive error term which generally is weaker than the relative-error bound. If $\boldsymbol{\mathcal{X}}$ is an exact low-rank tensor, then Eq. (4.4) implies perfect reconstruction, but the additive error term in Eq. (4.1) remains non-zero. Next, we highlight some advantages of TCUR-REB decomposition over HOSVD-based techniques for the decomposition of traffic tensors. Note that, both CP decomposition and Tucker decomposition are variants of HOSVD.

1. TCUR-REB achieves a relative-error bound (as given in Eq. (4.4)), which simply states that the Frobenius norm of the error in reconstruction is always less than or equal to $(2+\epsilon)$ times the error incurred in reconstruction using HOSVD based on rank-k approximation.

2. Note that the input rank parameter in Algorithm 3 does not imply that the output of the tensor-CUR decomposition, *i.e.*, the resulting tensor $\boldsymbol{\mathcal{C}} \times_3 \mathbf{U} \times_3 \mathbf{R}$, is a rank-$k$ approximation of the input traffic tensor $\boldsymbol{\mathcal{X}}$ because determining the tensor rank is an NP-hard problem. It rather implies that the error in reconstruction of original input traffic tensor $\boldsymbol{\mathcal{X}}$ using TCUR-REB is always less than or equal to the error in reconstruction using the rank-k approximation of HOSVD-based techniques.

3. The tensor $\boldsymbol{\mathcal{C}}$ and the matrix $\mathbf{R}$ can replace the factor matrices of CP decomposition and Tucker decomposition, with an added advantage of interpretability, because the original dimensions of the input tensor are retained as a subset.

4. The construction of output factors in TCUR-REB decomposition, *i.e.*, $\boldsymbol{\mathcal{C}}$, $\mathbf{U}$, and $\mathbf{R}$, does not involve the sensitive ALS procedure.

Tensor-CUR decomposition needs to have access to complete tensor data. This poses a limitation in its use as a decomposition technique for compressive sensing of traffic tensor. To overcome this limitation, we propose to fill the missing values with past available measurements before subjecting to tensor-CUR decomposition. Interpolation using a neighboring value is also termed *zeroth-order interpolation*. By filling in the missing values at a given time-interval with past values, we take into account the temporal properties of traffic tensor. The advantage of zeroth-order interpolation is that it is computationally inexpensive.

4.4.2 Matrix-CUR Decomposition

For compressive sensing of traffic matrices, we propose to use matrix-CUR decomposition in place of SVD-based techniques which involve ALS procedures. Matrix-CUR decomposition has an improved time complexity over SVD-based techniques. Here also, we first perform local interpolation procedure using *zeroth-order interpolation* and then employ global interpolation using matrix-CUR decomposition. In addition, we also employ k-nearest neighbour for better reconstruction as done in [9].

4.5 Experimental Results

We use two publicly available Internet traffic volume datasets collected from GEANT network [8] and Abilene network [7]. The dataset of Geant network consists of 17 traffic matrices labeled as **G01**, **G02**, ..., **G17**, with each containing traffic volume flow across 529 PoP pairs measured during one-week period. We represent the data as a three-dimensional traffic tensor using the dimensions; ingress node, egress node, and time interval. The characteristics of the dataset are summarized in Table 4.1.

TABLE 4.1: Datasets

Network	Time-interval	# IE pairs	Size of TT	Size of TM
GEANT (**G01**)	15 minutes	529	$23 \times 23 \times 671$	671×529
Abilene (**X10**)	5 minutes	121	$11 \times 11 \times 2016$	2016×121

The rank parameter chosen for the experimentation is two. Before discussing the results, we discuss how the missing values in the data are synthesized and also define the metrics used for evaluation.

4.5.1 Modeling Missing Values

The work in [101] identified various sources of error in collecting the traffic volume data and concluded that the missing values could be unstructured or structured. In [9, 17], the reasons for structured loss patterns were attributed to instrumentation errors and unreliable transport protocol usage. We consider the following loss patterns in traffic data, identical to [9, 17].

1. *Pure Random Loss*: This is simulated by randomly selecting a pre-determined percentage of entries of traffic matrix/tensor and making their value to be zero. This is an unstructured loss pattern.

2. xx*Time Random Loss*: This structured loss pattern, is simulated by randomly selecting xx% of time intervals and marking the traffic volumes across a specified percentage of node pairs to be zero for these time intervals.

3. xx*Element Random Loss*: This structured loss is simulated by randomly selecting xx% of the node pairs and marking the traffic volume during a specified percentage of time intervals to be zero for the selected node pairs.

4. *Row Random Loss*: This is simulated by randomly selecting an egress node and marking the traffic volume leaving it to be zero for a certain percentage of time intervals.

5. *Column Random Loss*: This is simulated similar to the previous case, but by randomly selecting an ingress node.

For each of these loss patterns, a tensor ($\boldsymbol{\mathcal{M}}$) of the required size is constructed with values initialized to unity. Then the values to be dropped according to the desired loss pattern are made zero. After this, the element-wise multiplication of $\boldsymbol{\mathcal{M}}$ is done with the original traffic tensor $\boldsymbol{\mathcal{X}}$. The resulting tensor is then subjected to compressive sensing techniques to reconstruct the missing values, to obtain the interpolated traffic tensor $\tilde{\boldsymbol{\mathcal{X}}}$.

In the case of reconstruction of missing values in the traffic matrix, the procedure for constructing matrix $\mathbf{M}$ for each of the loss patterns is identical to the construction of tensor $\boldsymbol{\mathcal{M}}$ explained above. The traffic matrix $\mathbf{X}$ is subject to element-wise multiplication with $\mathbf{M}$. The resulting matrix is subject to compressive sensing techniques which yields interpolated traffic matrix $\tilde{\mathbf{X}}$.

4.5.2 Evaluation Measures

We use the following three measures to evaluate the accuracy of reconstruction [102].

1. **Normalized Mean Absolute Error (NMAE)**: It denotes the average of absolute errors in the reconstruction of missing values. For traffic matrix,

$$NMAE = \frac{\sum_{i,j:M(i,j)=0} |X(i,j) - \tilde{X}(i,j)|}{\sum_{i,j:M(i,j)=0} |X(i,j)|}$$

   where $\tilde{X}(i,j)$ is the $(i,j)^{th}$ element of the final interpolated traffic matrix. For traffic tensor, NMAE is computed as follows:

$$NMAE = \frac{\sum_{i,j,k:\mathcal{M}(i,j,k)=0} |\mathcal{X}(i,j,k) - \tilde{\mathcal{X}}(i,j,k)|}{\sum_{i,j,k:\mathcal{M}(i,j,k)=0} |\mathcal{X}(i,j,k)|}$$

where $\tilde{\mathcal{X}}(i,j,k)$ is the $(i,j,k)^{th}$ element of the final interpolated traffic tensor.

2. **Normalized Mean Squared Error (NMSE)**: It denotes the average of squared errors in the reconstruction of missing values. For traffic matrix, NMSE is computed as follows:

$$NMSE = \frac{\frac{1}{|\Omega|}\sum_{i,j:M(i,j)=0}|X(i,j)-\tilde{X}(i,j)|^2}{\frac{1}{|T|}\sum_{i,j}|X(i,j)|^2}$$

where $|\Omega|$ is the total number of missing entries and T denotes the total number of time intervals. For traffic tensor, NMSE is given as follows:

$$NMSE = \frac{\frac{1}{|\Omega|}\sum_{i,j,k:\mathcal{M}(i,j,k)=0}|\mathcal{X}(i,j,k)-\tilde{\mathcal{X}}(i,j,k)|^2}{\frac{1}{|T|}\sum_{i,j,k}|\mathcal{X}(i,j,k)|^2}.$$

3. **Normalized Root Mean Squared Error (NRMSE)**: It denotes the standard deviation of the errors in the reconstruction of missing values. For traffic matrix, NRMSE is computed as follows:

$$NRMSE = \frac{\sqrt{\frac{1}{|\Omega|}\sum_{i,j:M(i,j)=0}|\mathcal{X}(i,j)-\tilde{X}(i,j)|^2}}{\sqrt{\frac{1}{|T|}\sum_{i,j}|\mathcal{X}(i,j)|^2}}$$

For traffic tensor, NRMSE is computed as follows:

$$NRMSE = \frac{\sqrt{\frac{1}{|\Omega|}\sum_{i,j,k:\mathcal{M}(i,j,k)=0}|\mathcal{X}(i,j,k)-\tilde{\mathcal{X}}(i,j,k)|^2}}{\sqrt{\frac{1}{|T|}\sum_{i,j,k}|\mathcal{X}(i,j,k)|^2}}$$

The performance of these techniques is evaluated in terms of NMAE, NMSE, and NRMSE. We take into account both non-structured as well as structured loss patterns with loss rates ranging from 1% to 98%. First, we present the experimental

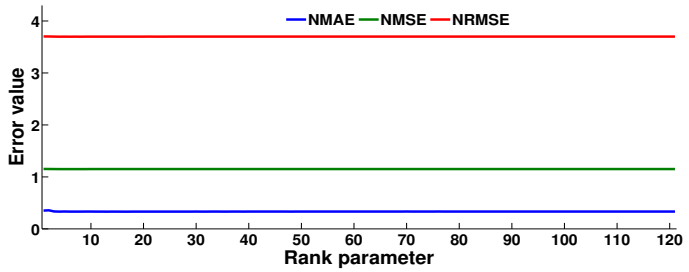


FIGURE 4.2: Sensitivity of TCUR-REB to the rank parameter using Abilene traffic tensor.

results for compressive sensing of traffic tensor using tensor decomposition techniques. Next, we present the experimental results for compressive sensing of the traffic matrix using matrix decomposition techniques.

4.5.3 Compressive Sensing of Traffic Tensor

Sensitivity of TCUR-REB to Rank Parameter:

We studied the sensitivity of the proposed technique, TCUR-REB, to the choice of input rank parameter using various error measures. We used Abilene dataset for this experiment. For the loss rate of 95%, the input rank parameter is varied from 2 to 121 in steps of unity. Figure 4.2 shows the NMAE, NMSE, and NRMSE for different values of the rank parameter. As it can be seen from the figure, the variation in the different error measures is close to 0. Experiments conducted with other loss rates also show the same trend. We also observed no sensitivity to the rank parameter with the GEANT dataset. As the rank parameter increases, the number of selected columns increases, but there is no variation in the error, indicating that the proposed algorithm is not sensitive to the choice of rank parameter. Henceforth, we present all the results for the rank parameter equal to 2.

Sensitivity of ALS-based Procedure:

To demonstrate the performance of the proposed technique, *i.e.*, zeroth-order interpolation followed by (**TCUR-REB**), we compare with the following state-of-the-art tensor compressive sensing techniques.

1. **CP-ALS** : Incorporates the CP decomposition and ALS procedure for compressive sensing [103].

2. **TUK-ALS**: Incorporates Tucker decomposition and ALS procedure [103].

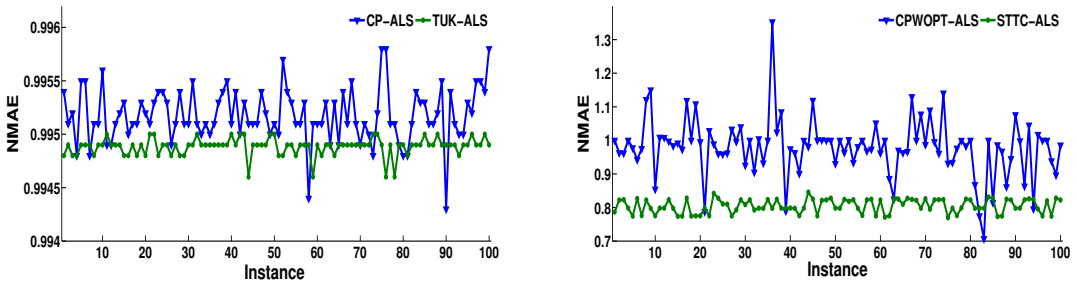
3. **CPWOPT-ALS** : Uses a first-order optimization approach to solve the weighted ALS procedure [87].

4. **STTC-ALS**: Exploits the spatio-temporal properties and CP decomposition with the ALS procedure [17].

5. **TCUR-AEB**: Uses Frobenius and Euclidean norm criteria for selection of frontal slices and tubes of tensor [97].

Before presenting the results, we demonstrate the sensitivity of the ALS procedure with respect to the initialization of factor matrices. Experiments are conducted for different loss rates and the observations are similar. We present the results for a loss rate of 95% and use four ALS-based decomposition techniques; CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS for the reconstruction of the tensor. The rank parameter is fixed as 2. For each technique, we conducted 100 runs with a random initialization of factor matrices. We compare the NMAE, NMSE, NRMSE and computational time for all the techniques to show that the techniques based on ALS are sensitive to the initialization of factor matrices.

Fig. 4.3 shows the NMAE obtained with different ALS-based methods using the tensors constructed with GEANT data. The x-axis in the plots indicates the runs and the y-axis shows the error in reconstruction of missing values in terms of NMAE. We can see that the NMAE values vary across different runs of the experiment, but this is lower for CP-ALS and TUK-ALS, compared to CPWOPT-ALS and STTC-ALS. Larger values of NMAE for CP-ALS and TUK-ALS indicate a high error in the reconstruction of missing values. The variation ratios of NMAE (maximum over the minimum) for CP-ALS and TUK-ALS are 1.0015 and 1.0004, respectively. On the other hand, CPWOPT-ALS has a high variation ratio of 1.913 with respect to NMAE. STTC-ALS has a variation ratio of 1.100.



(a) NMAE (b) NMAE

FIGURE 4.3: Error in reconstruction for 95% pure random loss in GEANT data for CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS.

We observed a similar trend for NMSE and NRMSE values as well. We also computed the confidence intervals and observe that CP-ALS and TUK-ALS have smaller confidence intervals as compared to CPWOPT-ALS and STTC-ALS. For example, the confidence intervals of NMAE with CP-ALS and TUK-ALS are {0.9951,0.9952 } and {0.9949,0.9949}, respectively, while for CPWOPT-ALS and STTC-ALS, it is {0.9626,0.9971 } and {0.8009,0.8090}, respectively.

Experiments with Abilene dataset also showed that for most of the runs on Abilene traffic data, CP decomposition and Tucker decomposition result in identical values of NMAE, NMSE, and NRMSE with least variation. However, the values of NMAE, NMSE, and NRMSE are high suggesting a large error in the reconstruction. On the other hand, CPWOPT-ALS and STTC-ALS show lower NMAE, NMSE, and NRMSE values. However, the variation in the values is high across the runs. For CPWOPT-ALS, the NMAE lies in the range of {0.74 to 1.04 }, the NMSE lies in the range {0.51 to 1.38}, and the NRMSE lies in the range {2.49 to 4.11}. For STTC-ALS, the NMAE varies between 0.72 and 0.85, the NMSE varies between 0.99 and 1.13, and the NRMSE varies between 3.49 and 3.72.

These experiments on two different traffic datasets show that STTC-ALS gives the least error in the reconstruction of missing values, while CP-ALS and TUK-ALS give the largest error. However, what is more important to notice is that ALS-based procedures converge to different reconstructed tensors with different initializations of factor matrices. There is a large variation in the NMAE, NMSE, and NRMSE values across the runs, with CPWOPT-ALS exhibiting the largest

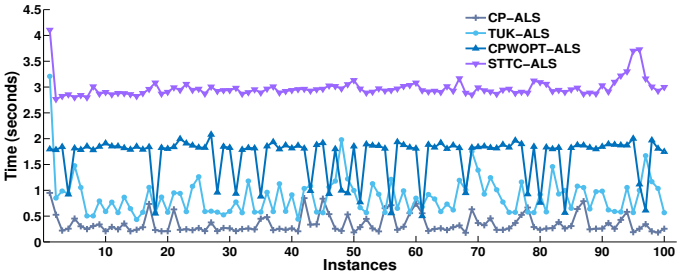


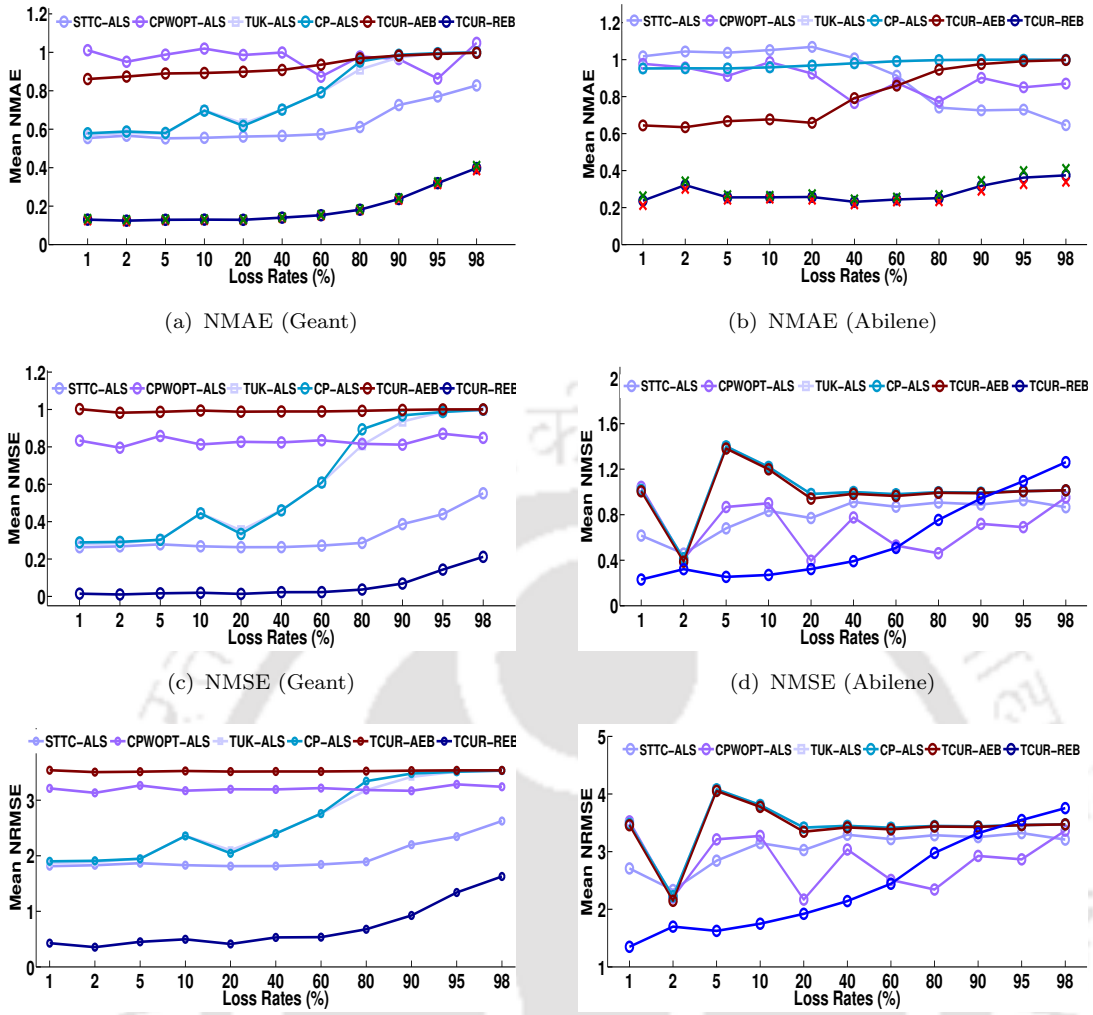
FIGURE 4.4: Computational time taken for compressive sensing of GEANT data using different ALS-based methods.

variation, thereby suggesting that the ALS-based methods are highly sensitive to initialization of factor matrices.

We also computed the computational time incurred in running different ALS-based decomposition methods for the reconstruction of missing values in GEANT and Abilene traffic tensor data. For GEANT traffic tensor, different ALS-based techniques exhibit different convergence times across the runs as shown in Fig. 4.4. In fact, the variation factor (maximum divided by the minimum) for CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS are 5.3508, 7.3964, 4.1412, and 1.4850, respectively. For Abilene traffic tensor, CP-ALS exhibits faster convergence and is not much sensitive to the initialization of factor matrices. The rest of the methods show a wide variation in convergence times. The largest variation is seen with the STTC-ALS method. We found STTC-ALS had the least NMAE, NMSE, and NRMSE values, but had the largest variation in its convergence times across the runs. On the other hand, CP-ALS had a low variation in the convergence times but had high NMAE, NMSE, and NRMSE values. The variation factor should be minimal considering the fact that the same optimization formulation is solved with different initializations of the factor matrices. It is precisely to avoid the sensitivity of ALS-based procedures, we proposed the use of TCUR-REB decomposition for compressive sensing of Internet traffic data.

Pure Random Loss Pattern:

First, we present the performance of each technique for unstructured loss pattern. We varied the loss rate from 1% to 98% and run each experiment to reconstruct the data 100 times. Fig. 4.5 shows the error in reconstruction of missing values in



(a) NMAE (Geant)

(b) NMAE (Abilene)

(c) NMSE (Geant)

(d) NMSE (Abilene)

(e) NRMSE (Geant)

(f) NRMSE (Abilene)

FIGURE 4.5: Error in reconstruction of missing values with Pure Random Loss in Geant and Abilene traffic tensor with loss rates from 1% to 98%.

terms of the mean values of NMAE, NMSE, and NRMSE using different tensor decomposition techniques on Geant and Abilene traffic tensors, respectively. In the plots, the x-axis shows the loss rate variation in percentage and the y-axis shows the error in reconstruction.

For GEANT traffic data, we observe that the proposed technique (TCUR-REB) gives the least error in the reconstruction of missing values compared to other techniques across all the loss rates. Table 4.2 shows the error in reconstruction with 98% of the values dropped at random from the GEANT traffic tensor. The 95% confidence intervals for mean NMAE with 98% pure random loss were found

to be smaller for TCUR-REB compared to the other techniques. In particular, TCUR-REB has a confidence interval of {0.3912,0.4089}; while TCUR-AEB, CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS have confidence interval of {0.9976,0.9978}, {0.9982,0.9993}, {0.9979,0.9990}, and {0.9830,1.0602}, and {0.8179,0.8501}, respectively. The results were similar with NMSE and NRMSE values as well.

TABLE 4.2: Reconstruction error for 98% pure random loss of GEANT TT.

Technique	Mean NMAE	Mean NMSE	Mean NRMSE
TCUR-REB	0.39	0.21	1.62
TCUR-AEB	0.99	1.00	3.54
CP-ALS	0.99	0.99	3.53
TUK-ALS	0.99	0.99	3.53
CPWOPT-ALS	1.05	0.84	3.24
STTC-ALS	0.83	0.55	2.62

For the Abilene traffic tensor, TCUR-REB and TCUR-AEB perform better than CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS. The TCUR-REB technique also outperforms the tensor-CUR decomposition [97] with additive-error bound (TCUR-AEB). Table 4.3 shows the error in reconstruction with 98% of the values dropped at random from the Abilene traffic. The 95% confidence intervals for mean NMAE with 98% pure random loss were found to be the least for TCUR-REB compared to the other techniques. In particular, TCUR-REB had a confidence interval of {0.3395,0.3911}, while TCUR-AEB, CP-ALS, TUK-ALS, CPWOPT-ALS, and STTC-ALS had confidence intervals of {0.9975,0.9979}, {0.9994,0.9999}, {0.9994,1.0000}, and {0.8330,1.0067}, and {0.6240,0.6702}, respectively. Similar results were observed with respect to the NMSE and NRMSE values as well.

Fig. 4.5(a) and Fig. 4.5(b) also show the standard deviation of reconstruction error in terms of NMAE for GEANT and Abilene data sets, respectively. To maintain the clarity in the figures, we show the standard deviation only for NMAE, using red and green cross (not prominent on the plots), where the red cross indicates the (mean value - standard deviation value) and the green cross indicates the (mean

value + standard deviation value). TCUR-REB exhibited less variation for most of the cases and this was no different for experiments on the GEANT dataset.

TABLE 4.3: Reconstruction error for 98% pure random loss of Abilene TT.

Technique	Mean NMAE	Mean NMSE	Mean NRMSE
TCUR-REB	0.37	1.26	3.75
TCUR-AEB	0.99	1.01	3.47
CP-ALS	0.99	1.01	3.47
TUK-ALS	0.99	1.01	3.47
CPWOPT-ALS	0.87	0.95	3.36
STTC-ALS	0.64	0.87	3.21

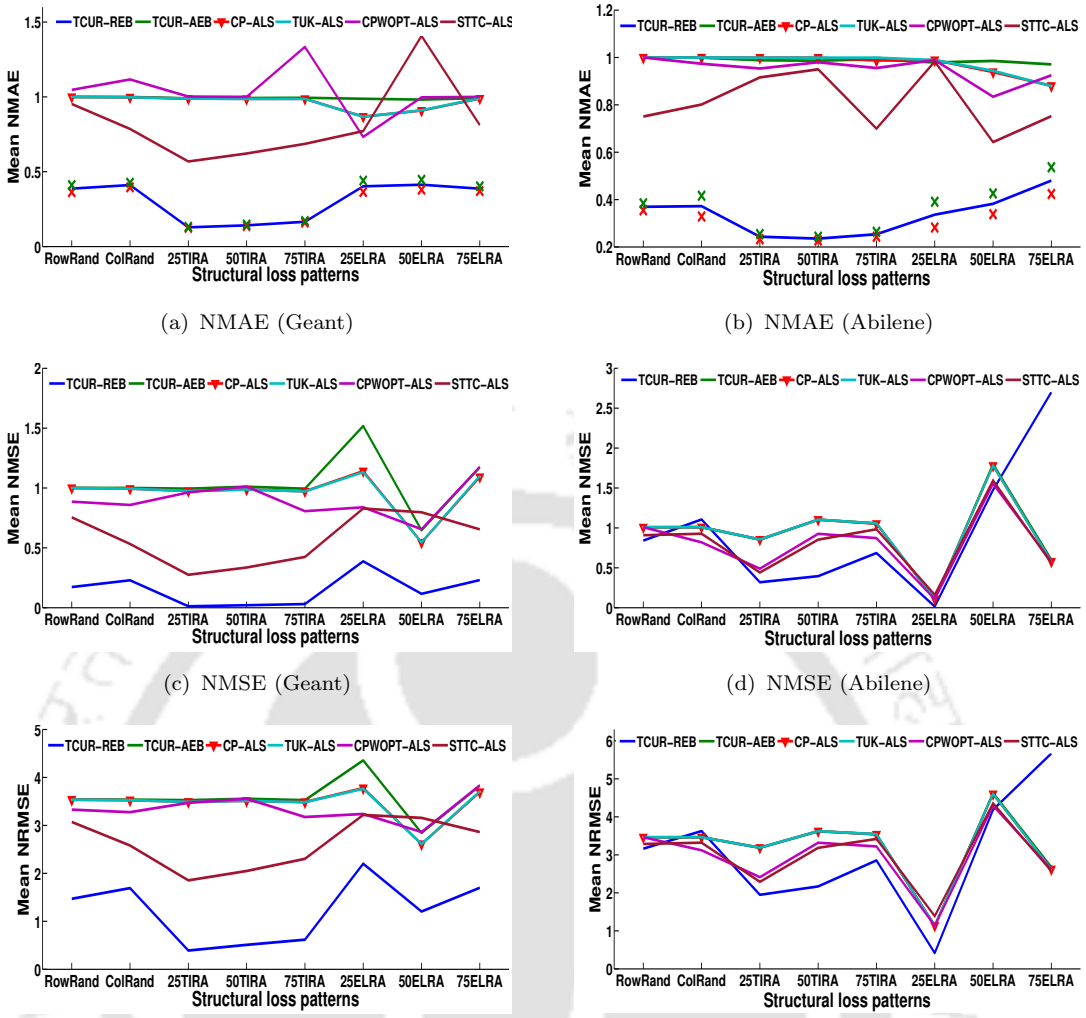
Structured Loss Pattern:

For the structured loss patterns as well, we conducted experiments with the loss rates ranging from 1% to 98% and the observations are similar. Here, we present the results only for the worst-case loss rate of 98%. Fig. 4.6 shows the error in the reconstruction of missing values with six decomposition techniques for GEANT and Abilene traffic tensor, respectively. In each of these plots, the x-axis indicates different structural loss patterns; row random (RowRand), column random (ColRand), 25 time random (25TIRA), 50 time random (50TIRA), 75 time random (75TIRA), 25 element random (25ELRA), 50 element random (50ELRA), and 75 element random (75ELRA). The y-axis indicates the error in terms of the mean values of NMAE, NMSE, and NRMSE.

Results demonstrate that the proposed technique gives the least error in the reconstruction of missing values. Table 4.4 and Table 4.5 show the error for GEANT and Abilene traffic tensor with 98% row random loss pattern, respectively. In Fig. 4.6(a) and Fig. 4.6(b), we also show the standard deviation in NMAE. To maintain the clarity of the figures, we show the standard deviation only for NMAE using red and green cross marks. For most of the cases, least standard deviation was observed for CP-ALS and TUK-ALS methods.

Temporal Stability Analysis of Volumetric Data:

We also used the traffic tensors constructed using different methods to conduct



(a) NMAE (Geant)

(b) NMAE (Abilene)

(c) NMSE (Geant)

(d) NMSE (Abilene)

(e) NRMSE (Geant)

(f) NRMSE (Abilene)

FIGURE 4.6: Error in reconstruction of missing values for different structured loss patterns with 98% loss rate in Geant and Abilene TT.

TABLE 4.4: Error for GEANT tensor with 98% RowRand loss.

Technique	Mean NMAE	Mean NMSE	Mean NRMSE
TCUR-REB	0.39	0.17	1.47
TCUR-AEB	0.99	1.00	3.53
CP-ALS	0.99	1.00	3.53
TUK-ALS	0.99	1.00	3.53
CPWOPT-ALS	1.05	0.88	3.32
STTC-ALS	0.95	0.75	3.07

temporal tracking analysis of volumetric traffic between OD pairs. The temporal stability feature as given in [58, 56] is used for this purpose. CP-ALS and

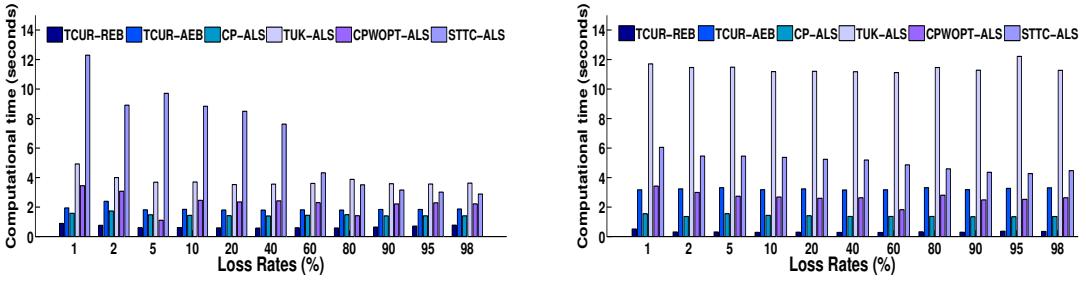
TABLE 4.5: Error for Abilene tensor with 98% RowRand loss.

Technique	Mean NMAE	Mean NMSE	Mean NRMSE
TCUR-REB	0.48	0.84	3.16
TCUR-AEB	0.99	1.00	3.46
CP-ALS	0.99	1.01	3.46
TUK-ALS	0.99	1.01	3.46
CPWOPT-ALS	1.00	0.01	3.46
STTC-ALS	0.75	0.91	3.28

TUK-ALS exhibit a low normalized difference between successive time-intervals, when compared to the proposed technique. For 95% pure random loss in GEANT data, CP-ALS and TUK-ALS had 92.71% and 90.55% of the time-intervals with a normalized difference less than 0.1 during successive time-intervals, whereas the proposed method had 83.20% of time intervals. We conclude that the traffic tensors obtained using CP-ALS and TUK-ALS are temporally more stable as compared to the proposed method. The reason for this is that CP decomposition and Tucker decomposition are exact decomposition techniques, while tensor-CUR decomposition gives an approximation to the input tensor.

Computational Time:

Finally, we also measured the computational time required for the reconstruction of missing values. Experiments were run on a 1.10 GHz Windows machine using Matlab. Fig. 4.7 shows the computational time for compressive sensing on GEANT and Abilene traffic data using different tensor decomposition techniques. In each of these plots, the x-axis indicates the loss rate variation and the y-axis denotes the computational time in seconds. For Geant traffic tensor, it can be seen that with 98% pure random missing values TCUR-REB takes 0.77 secs, which is the least time as compared to TCUR-AEB (1.86 secs), CP-ALS (1.40 secs), TUK-ALS (3.62 secs), CPWOPT-ALS (2.21 secs), and STTC-ALS (2.88 secs). For Abilene traffic tensor, TCUR-REB takes only 0.34 secs compared to TCUR-AEB (3.30 secs), CP-ALS (1.36 secs), TUK-ALS (11.26 secs), CPWOPT-ALS (2.63 secs), and STTC-ALS (4.46 secs) with 98% pure random missing values. The computational



(a) Geant (b) Abilene

FIGURE 4.7: Computational time taken for compressive sensing of GEANT and Abilene traffic tensor.

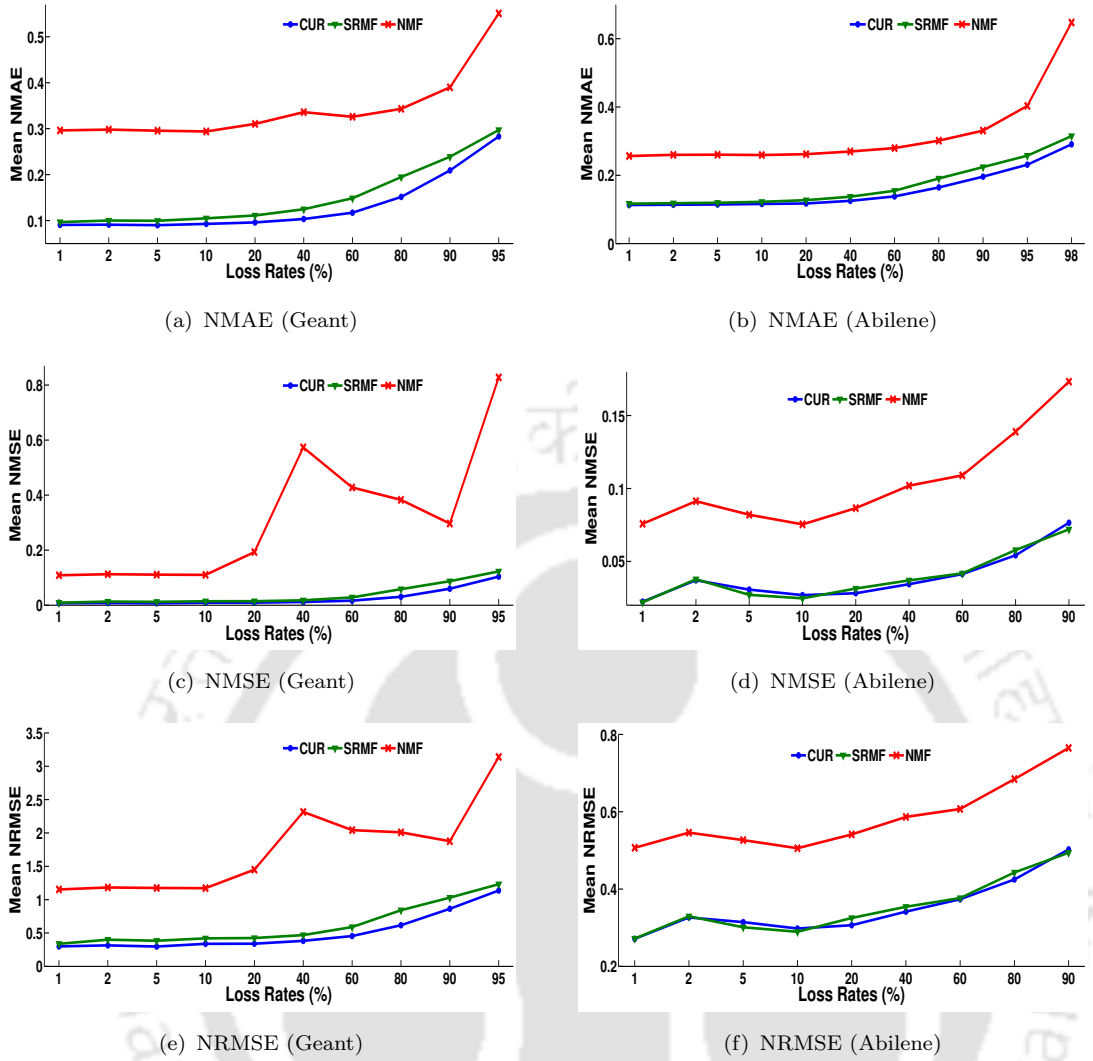
time is reduced with tensor-CUR decomposition mainly due to avoiding the ALS procedure, commonly used in all the other decomposition techniques.

As mentioned earlier, the compressive sensing technique in [17] first performs global interpolation using a decomposition technique followed by local interpolation. This procedure causes the reconstructed tensor not to be an exact product of the factor matrices. We first perform local interpolation, termed zeroth-order interpolation followed by global interpolation using TCUR-REB decomposition. By this, the reconstructed traffic tensor is ensured to be an exact product of the factor matrices. We compute the reconstruction error between the final reconstructed tensor and the product of factor matrices for the Abilene data. With 80% of missing values at random, the reconstruction error for (STTC + KNN) is 39.78%, while it is 0% with the proposed technique. Though not reported here, we find the error to be zero irrespective of the loss rates with the modified order of interpolation techniques.

4.5.4 Compressive Sensing of Traffic Matrix

Pure Random Loss:

Fig. 4.8 shows the error in reconstruction of missing values for pure random loss using different matrix decomposition techniques on Geant and Abilene traffic matrix. For Geant TM, it is observed that matrix-CUR decomposition outperforms



(a) NMAE (Geant)

(b) NMAE (Abilene)

(c) NMSE (Geant)

(d) NMSE (Abilene)

(e) NRMSE (Geant)

(f) NRMSE (Abilene)

FIGURE 4.8: Error in reconstruction of missing values with Pure Random Loss in Geant and Abilene traffic matrix with different loss rates.

other matrix decomposition techniques in terms of exhibiting least error in reconstruction (in terms of NMAE, NMSE, NRMSE values) for loss rates ranging from 1% to 95%. For 98% pure random loss of Geant TM, SRMF is shown to outperform matrix-CUR decomposition. In particular, the NMAE, NMSE and NMRSE value for matrix-CUR decomposition is 0.39, 0.19, and, 1.56 respectively; while for SRMF, it is 0.37, 0.18, and, 1.49 respectively. Note that this difference is not significant. On the other hand, for Abilene TM, the proposed matrix decomposition technique exhibits least NMAE value across all loss rates ranging from 1% to 98%, as compared to SRMF and NMF. For NMSE and NRMSE values as well, the proposed decomposition technique exhibits least value as compared to SRMF

and NMF for Geant TM. We present the error in reconstruction for 95% and 98% pure random loss of Abilene TM in Table 4.6.

TABLE 4.6: Reconstruction error for 95% and 98% pure random loss of Abilene TM.

Technique	NMSE	NRMSE	NMSE	NRMSE
	95%	95%	98%	98%
CUR	0.08	0.53	0.12	0.63
SRMF	0.09	0.54	0.13	0.66
NMF	0.66	1.29	3.30	3.11

Structured Loss Pattern:

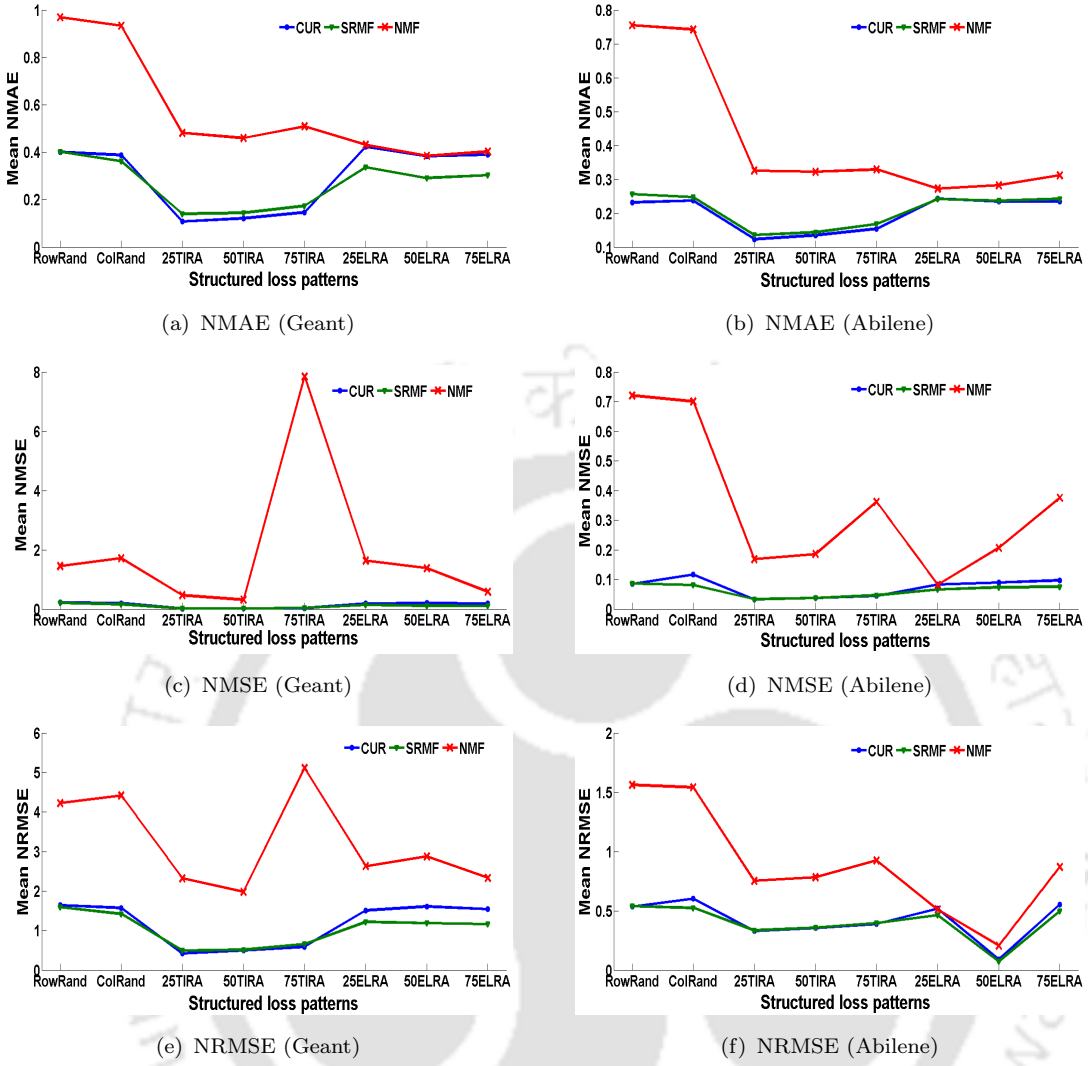
Fig. 4.9 shows the error in the reconstruction of missing values with structured loss patterns using three matrix decomposition techniques for GEANT and Abilene traffic matrix. In each of these plots, the x-axis indicates different structural loss patterns and the y-axis indicates the error in terms of the mean values of NMAE, NMSE, and NRMSE. Results demonstrate that the proposed matrix decomposition technique gives the least error in the reconstruction of missing values for row random and time random loss pattern. However, for column random and element random loss pattern, SRMF outperforms CUR in terms of exhibiting least error.

Computational Time:

Fig. 4.10 shows that proposed matrix decomposition exhibits least computational time for compressive sensing of Geant and Abilene traffic matrix as compared to SRMF and NMF. The reduction in computational time achieved with matrix-CUR decomposition is mainly due to avoiding the ALS procedure, commonly used in all the other decomposition techniques.

4.6 Summary

In this chapter, we introduced relative-error bound tensor-CUR (TCUR-REB) decomposition for the reconstruction of missing values in tensor data. For an



(a) NMAE (Geant)

(b) NMAE (Abilene)

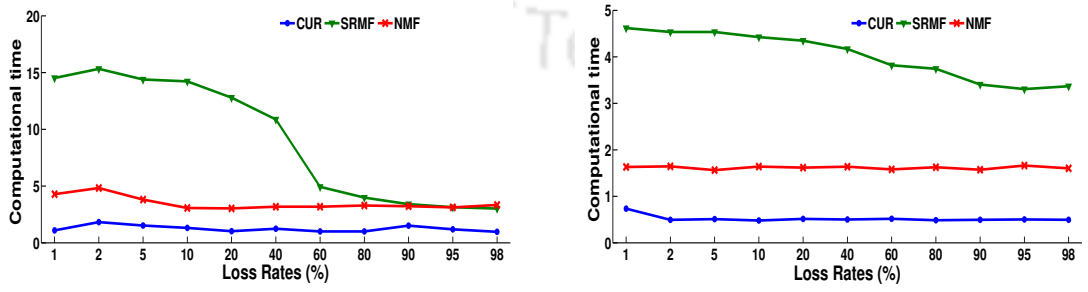
(c) NMSE (Geant)

(d) NMSE (Abilene)

(e) NRMSE (Geant)

(f) NRMSE (Abilene)

FIGURE 4.9: Error in reconstruction of missing values for different structured loss patterns with 98% loss rate in Geant TM and 95% loss rate in Abilene TM.



(a) Geant

(b) Abilene

FIGURE 4.10: Computational time taken for compressive sensing of GEANT and Abilene traffic matrix.

application domain, we chose compressive sensing of Internet traffic data. The proposed tensor-CUR decomposition avoids fundamental limitations of other tensor decomposition techniques namely, knowledge of rank apriori and sensitivity to initialization of factor matrices due to the ALS procedure. The proposed technique also exhibits an improvement over the popular TCUR-AEB decomposition technique in terms of obtaining a relative-error bound on low-rank approximation as compared to a weaker additive-error bound. We also revisit the order of global and local interpolation procedure so that the final reconstructed tensor is a product of factor matrices. In addition, we also propose the use of matrix-CUR decomposition for compressive sensing of traffic matrices. Matrix-CUR decomposition also does not make use of computationally expensive ALS procedures to obtain factor matrices. Extensive experimentation on two publicly available datasets shows the efficacy of proposed tensor and matrix decomposition techniques over other state-of-the-art tensor and matrix decomposition techniques, in terms of exhibiting least error in reconstruction of missing values and taking the least computational time.

Chapter 5

Analysis of Internet Traffic Data

In this chapter, we present the third contribution of the thesis which performs two distinct analysis of Internet traffic data, namely structural analysis and volume anomaly analysis, where the former deals with the identification of different structural patterns in the traffic matrix and the latter deals with the identification of time-intervals and ingress-egress node pairs associated with the volume anomalies.

5.1 Introduction

The traffic matrix is the key input to a multitude of network operations and management tasks including network monitoring and generation of synthetic traffic matrices. These tasks require a deeper analysis of components and structures of traffic matrices, which falls in the realm of structural analysis. The identification of structural patterns of traffic matrix is important in the current scenario when there is a scarcity of publicly available datasets. In this scenario, these identified structural patterns are of great help to network researchers for the generation of synthetic traffic matrices which are used for carrying out various research-related activities. The objective of structural analysis is to decompose a traffic matrix as a sum of its constituents or structural patterns, which are later classified into different categories based on their behavior. This systematic decomposition sheds

light on the intrinsic structure of the traffic matrix, and consequently on the behavior of the network as a whole. In fact, by categorizing the structural patterns in this fashion, one can gain significant insights into the whole-network properties of Internet traffic data.

The identified structural patterns are of help in the generation of synthetic traffic matrices which are used for testing purposes in network-wide applications. For example, the synthetic traffic matrices generated using identified spikes patterns are used for volume anomaly analysis where *a volume anomaly* refers to a large and sudden positive or negative changes in the traffic volume flow.

These changes are typically caused by unexpected events such as outages coming from network equipment failures, and network attacks like Denial-of-Service (DoS) attacks. With the rapid growth in IP traffic flow [1], analysis of these volume anomalies is crucial for network operators to correctly manage large and unexpected congestion problems at the backbone network caused by traffic volume anomalies. Analysis of volume anomalies consists of the identification of time intervals during which network experiences the anomaly, as well as the PoP pairs which triggered it.

The key challenge involved in structural analysis and volume anomaly analysis is the high-dimensional multi-variate structure of the traffic matrix. PCA and its variants have been extensively employed in literature to obtain a low-dimensional representation of traffic matrix which facilitates the analysis of traffic matrix. The low dimensional representation of the traffic matrix obtained after PCA, namely eigenflows, are considered as the structural patterns of the traffic matrix. This is because each eigenflow captures a particular source of temporal variability (*structural patterns*) common to all OD flows [10]. Each OD flow is then expressed as a weighted sum of these structural patterns or eigenflows. These structural patterns are later classified into three categories: periodic, spikes and noise. However, Wang *et al.* [19] demonstrated the inefficiency of categorization approach for structural analysis in terms of completeness, *i.e.*, each pattern has to be classified into *at least one category*, and orthogonality, *i.e.*, each pattern *must not be classified into more*

than one category at the same time. This results in a high value of Unclassified Energy Rate (UER). They have proposed a robust variant of PCA, namely Robust Principal Component Analysis (RPCA), for structural analysis of the traffic matrix. The RPCA problem is to decompose a given matrix as a sum of a low-rank matrix (corresponding to the periodic pattern), a sparse matrix (corresponding to spikes pattern), and a random matrix (corresponding to noise pattern). However, this decomposition fails in case if the sparse matrix is also a low-rank matrix or if the low-rank matrix is also a sparse matrix. This results in confusion between the identification of periodic and spikes patterns. Although RPCA has resolved the issue concerning the completeness, however, the problem of orthogonality still persists. In addition, PCA suffers from two fundamental limitations stemmed in its assumption as stated in chapter 1, *i.e.*, *assumption of continuous random variables* and *lack of interpretability*. As a consequence of the lack of interpretability, one can not identify the PoP pairs responsible for a particular structural pattern.

For anomaly analysis, PCA was first exercised by Lakhina *et al.* [11]. Lakhina *et al.* [11] separates the principal component space, *i.e.*, the set of right singular vectors, into normal and anomalous subspaces, and hence referred to as PCA-subspace method. The first k PCs comprise the normal subspace and remaining PCs comprise anomalous subspace. The PCA-subspace method identifies a time-interval as anomalous if the magnitude of projection onto the anomalous subspace exceeds an associated PCA Q-statistic threshold. After Lakhinas studies, various other authors [20, 21, 22] have used PCA-subspace method.

Three limitations are identified for the PCA-subspace method for volume anomaly analysis. The first one is associated with the lack of interpretability of obtained factor matrices. As a consequence, identification of PoP pairs which triggered the volume anomaly becomes fundamentally difficult because of the inability of recovering original input space. The second limitation is associated with the choice of the number of PCs in the normal subspace. Ringberg *et al.* [23] have demonstrated that the false positive rate is very sensitive to small differences in the number of PCs in the normal subspace. The third limitation is because of the fact that a

large anomaly may inadvertently pollute the normal subspace as a result of which it goes undetected [23].

5.1.1 Main Contributions

To overcome these limitations, the third contribution of the thesis proposes the use of two techniques, namely Correspondence Analysis (CA) and matrix-CUR decomposition, for structural analysis and volume anomaly analysis. Since the output of both techniques can be interpreted in terms of original columns and rows of traffic matrix, both techniques retain the original input space. In addition, both techniques exhibit independence from the underlying distributional assumption of traffic matrices. The advantage of both techniques affirms them to be more apt for *structural analysis and volume anomaly analysis of traffic matrices* as compared to PCA-based approaches. The main contributions of this chapter are:

1. *Structural analysis*: We employ CA and matrix-CUR decomposition techniques to perform structural analysis. Both of these techniques address the above-identified limitations. In addition, we show that matrix-CUR decomposition addresses the problems specific to the classification of structural patterns namely completeness and orthogonality are retained.

2. *Volume anomaly analysis*: We propose the use of CA and matrix-CUR decomposition for volume anomaly analysis of the traffic matrix. All the identified limitations in volume anomaly analysis are addressed by each of the employed techniques.

5.1.2 Summary of Experimental Results

We make use of Abilene and Geant traffic matrices for structural analysis. For volume anomaly analysis, we present the results on Abilene traffic matrices and synthetic traffic matrices. Below we present the summary of experimental results.

1. *Structural analysis*:

   - For the majority of the Geant traffic matrices, CA outperforms PCA in terms of completeness, orthogonality, and UER.
   - For the majority of the Abilene TMs, CA outperforms PCA in terms of completeness and UER; but lags in terms of orthogonality.
   - Matrix-CUR decomposition, on the other hand, outperforms PCA and CA in terms of orthogonality and completeness for all the matrices.

2. *Volume anomaly analysis*:

   - Matrix-CUR decomposition and CA identify identical volume anomalies during Wednesday and Thursday and flowing across Chicago-Los Angeles and Los Angeles-Chicago PoP pair for an Abilene traffic matrix.
   - For synthetic traffic matrices, matrix-CUR decomposition is effective with a high detection rate and low false-positive rate for volume anomaly analysis as compared to the PCA-subspace method. CA, on the other hand, is a qualitative approach, hence can't be used in this comparison.

5.2 Related Work

In this section, we review the work present in the literature for structural analysis and volume anomaly analysis of traffic matrix and highlight the limitations, thereby serving as motivation for the proposed contribution.

5.2.1 Structural Analysis

PCA was first performed by Lakhina *et al.* [10] for the structural analysis of $\mathbf{X} \in \mathbb{R}^{T \times P}$ as given in Eq. (2.1). It has been observed in [10] that TM exhibits a low intrinsic dimensionality, *i.e.*, most of the variation (energy) present in TM

can be well approximated by 5 to 10 PCs. Eigenflows $\{\mathbf{e}_i\}_{i=1}^{P}$, which represents a particular temporal pattern common in all OD flows, are computed corresponding to each PC as follows:

$$\mathbf{e}_i = \frac{\mathbf{X}\mathbf{v}_i}{\sigma_i} \tag{5.1}$$

for i = 1, 2, ...,P. Since, contribution to the overall energy by v_1 is maximum, eigenflow e_1 corresponding to v_1 captures the strongest temporal trend common to all OD flows, e_2 captures the next strongest, and so on. These eigenflows are considered as the structural patterns of traffic matrix. OD flows are then expressed as a weighted linear combination of these structural patterns. The weight associated with an eigenflow denotes the extent to which that temporal pattern (eigenflow) is present in an OD flow. As per Lakhina's observation, most of the OD flows are found to be composed of few eigenflows. Later, these eigenflows are classified into three categories - deterministic (d), spikes (s) and noise (n) eigenflows corresponding to periodic, spikes, and noise structural patterns, respectively. This categorization of eigenflows allows having a significant insight into the structure and properties of network traffic.

Despite providing meaningful insights for structural analysis of TM, approach in [10] suffers from few drawbacks. A significant number of eigenflows remain unclassified. Lakhina *et al.* [10] claimed that top eigenflows, *i.e.*, eigenflows corresponding to top PCs, satisfy the criteria of d-eigenflows. On the contrary, it has been found that top eigenflows may correspond to indeterminate[1] and non-determinate[2] eigenflows. This results in a high value of Unclassified Energy Rate (UER), which is defined as the percentage of energy captured by the principal component vector that corresponds to either indeterminate or non-determinate eigenflow. The energy captured by the i^{th} principal component is given by the square root of its corresponding singular value (σ_i). Hence, unclassified energy rate (UER) is expressed as

$$\text{UER} = \frac{\sum_{i \in UEID} \lambda_i}{\sum_{i=1}^{P} \lambda_i} \tag{5.2}$$

[1]Eigenflows that satisfy more than one category.
[2]Eigenflows that does not satisfy any category.

where $\lambda_i = \sqrt{\sigma_i}$ and the union of unclassified eigenflow ID (UEID) is the total number of unclassified eigenflows, *i.e.*, indeterminate and non-determinate eigenflows.

Wang *et al.* [19] stated that in the case when TM is polluted with large volume anomalies, the eigenflows classification may result in increased indeterminate and non-determinate structures. To deal with such traffic matrices, a robust variant of PCA, *i.e.*, Robust Principal Component Analysis (RPCA), has been used in [19]. Because of ambiguity in eigenflow categorization, *i.e.*, presence of indeterminate and non-determinate eigenflows, they have decomposed the traffic matrix $\mathbf{X}$ into three sub-matrices as follows:

$$\mathbf{X} = \mathbf{X_a} + \mathbf{X_e} + \mathbf{X_n} \tag{5.3}$$

where $\mathbf{X}_a$ represents a deterministic traffic matrix, $\mathbf{X}_e$ represents anomaly traffic matrix, and $\mathbf{X}_n$ represents the noise traffic matrix. $\mathbf{X}_a, \mathbf{X}_e$, and $\mathbf{X}_n$ contain periodic, spikes, and noise structural patterns, respectively. The decomposition described in Eq. (5.3) fails in case if the sparse matrix is also a low-rank matrix or if the low-rank matrix is also a sparse matrix. This results in confusion between the identification of deterministic TM and anomaly TM. This approach has resolved the issue concerning the completeness, however, the problem of orthogonality still persists.

Limitations: Despite the simplicity, the PCA-based approach suffers from two fundamental limitations. First, *lack of interpretability of obtained structural patterns*. Unfortunately, being linear combinations of up to all the OD flows, these structural patterns are notoriously difficult to interpret in terms of the original PoP pairs generating it. The second limitation is associated with *the assumption of continuous random variables*, which results in excessive skewness and kurtosis of the obtained factor matrices [12].

5.2.2 Volume Anomaly Analysis

Lakhina *et al.* have extended their work in [11] and proposed the PCA-subspace method for detection, identification, and quantification of large volume anomalies across the network. The PCA-subspace method involves the separation of the space occupied by PCs into two disjoint subspaces: normal subspace S (consisting of few top PCs) and abnormal subspace $\tilde{S}$ (consisting of remaining PCs). The separation procedure involves the projection of original traffic data onto PCs in order from the top. As soon as the projection exceeds a threshold (*say* a 3σ deviation from the mean), that PC and all subsequent PCs are assigned to anomalous subspace. All previous PCs are assigned to normal subspace. Original traffic data is projected onto normal and anomalous subspaces thereby obtaining the modeled and the residual traffic, respectively. It has been assumed in [11] that anomalies reside in abnormal subspace and a volume anomaly will result in a large change to residual traffic. Squared prediction error of residual traffic is computed at each time step and is flagged as an anomalous time-interval if found to exceed the detection threshold. The anomalous flows are identified using the following strategy: associate j^{th} flow as anomalous if j^{th} column of residual traffic has the maximum contribution during anomalous time-intervals.

Huang *et al.* [22] proposed a new framework for anomaly analysis that makes use of distributed tracking combined with PCA-subspace method, and considered the tradeoff between detection accuracy and communication overhead. Rubinstein *et al.* [20] analysed poisoning and defense techniques for the PCA-subspace detector in backbone networks. Other works utilizing a PCA-based approach on SNMP measurements include [21, 9]. However, SNMP data is highly susceptible to error due to the following factors: missing link observations, incorrect link observations, and inaccurate polling times. Hence, works in [104, 105] have applied the PCA-subspace method on OD flow data which is a rich source of information.

Limitations: Three limitations are identified when the PCA-subspace method is employed for volume anomaly analysis.

Interpretability Identification of PoP pairs which triggered the volume anomaly becomes fundamentally difficult.

Number of PCs Ringberg *et al.* [23] have demonstrated that the false positive rate is very sensitive to small differences in the number of principal components in the normal subspace.

Polluted normal subspace When a large volume anomaly inadvertently pollutes the normal subspace, the anomaly goes undetected [23].

5.3 Proposed Techniques

In this section, we present the use of two techniques namely CA and matrix-CUR decomposition for structural analysis and volume anomaly analysis of traffic matrices which alleviates the above-identified limitations.

5.3.1 Correspondence Analysis

We apply CA on TM $\mathbf{X}$ by using the algorithm 1. The SVD of the matrix of standardized residuals $\mathbf{S}_r$ results in three factor matrices namely the matrix $\mathbf{J}$ containing the left singular vectors, the diagonal matrix Σ containing singular values, and $\mathbf{K}$ containing right singular vectors. The obtained factor matrices are used for performing both structural analysis and volume anomaly analysis. We first present the procedure for performing structural analysis followed by the procedure performed for volume anomaly analysis.

Structural Analysis:

We designate the obtained left singular vectors as the structural patterns of $\mathbf{X}$. Each OD flow is then expressed as a weighted linear combination of these structural patterns where the weights are defined by the rows of the matrix of right singular vectors. These derived set of structural patterns are classified into one of the following three categories:

1. **Deterministic ('d') structural patterns**, which exhibit a diurnal pattern in network traffic as well as the difference between weekdays and weekends traffic.

2. **Spike ('s') structural patterns**, which captures a massive upsurge/down surge in traffic flow volume.

3. **Noise ('n') structural patterns**, which captures the remaining random variation exhibiting relatively time-invariant properties across all OD flows.

The identical procedure is used for categorizing deterministic patterns and spikes patterns as presented in [10]. Noise patterns are categorized using the Chi-square goodness of fit test. It returns a test decision for the null hypothesis whether the data in the input vector comes from a normal distribution. The test decision is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise. If the test decision is 0, then the structural pattern satisfies the criteria for n-eigenflow. The performance of this categorization procedure is evaluated on the basis of Unclassified Energy Rate (UER) which is given in Eq. (5.2).

Volume Anomaly Analysis:

We make use of the biplots obtained using CA. The first two columns in **F** (principal coordinates of rows) are plotted as dots. These dots denote the rows of **X** namely the time intervals. Similarly, the first two columns in **G** (principal coordinates of columns) are plotted as lines. These lines denote columns of **X** namely PoP pairs. The length of a line in the biplot is directly proportional to the standard deviation of the corresponding column of **X**. Similarly, the distance of a dot from the origin is directly proportional to standard deviation of the corresponding row of **X**. Rows or columns of **X** having entries exceeding largely from their mean are plotted far from origin as compared to others in the biplot. This feature of CA biplot will assist in volume anomaly analysis as it will place large volume anomalous time intervals and PoP pairs far from the origin as compared to normal time intervals and PoP pairs. The association between a row and a column of **X** can be understood through biplot in the following manner: If i^{th} time interval and j^{th}

PoP pair exhibit volume anomaly, then the i^{th} time interval and j^{th} PoP pair will be close to each other in the biplot. In other words, the angle between these two vectors will be small.

5.3.2 Matrix-CUR Decomposition

Both PCA and CA obtain P structural patterns. Examining each of these and classifying them is an involved task. To address this issue, matrix-CUR decomposition is employed. Applying matrix-CUR on **X** results in three matrices namely **C, U** and **R**. We designate the columns of **C** as structural patterns which consists of c columns, where $c << P$. As the resulting structural patterns are very small in number, matrix-CUR alleviates the overhead in examining and classifying each of the P structural patterns. Traffic flow across each PoP pair is then expressed as a weighted sum of these *few* structural patterns which are later classified into three categories.

The construction of **C** involves computation of normalized statistical leverage score for all columns of **X**. This score depends on the Euclidean norm of rows of the matrix containing top k right singular vectors obtained after PCA of **X**. The probability of selection of i^{th} column is given as:

$$p_i = min\Big\{1, c * \frac{1}{k}\sum_{j=1}^{k}(v_{ij})^2\Big\} \tag{5.4}$$

where v_{ij} denote the $(i,j)^{th}$ element of the matrix of right singular vectors and $p_i \in [0,1]$. Eq. (5.4) states p_i is directly proportional to the standard deviation of i^{th} column of **X**. This imply that the columns of **X** having high standard deviation have high probability to be part of **C**. Along the same line, the rows of **X** having high standard deviation have high probability to be part of **R**. With the construction of **C** and **R**, matrix-CUR decomposition separates the columns and rows of TM having high standard deviation from the remaining ones. From the matrix **C** and **R**, we designate the columns of **C** as anomalous PoP pairs,

if $p_j = 1$. In a similar fashion, we designate the rows of $\mathbf{R}$ as anomalous time interval, if $p_j = 1$.

5.4 Experimental Results

We make use of Abilene and Geant traffic matrices for structural analysis. For volume anomaly analysis, we present the results on Abilene traffic matrices and synthetic traffic matrices. To obtain ground-truth information, we utilize three traffic matrices which are generated synthetically using the procedures discussed in [24] followed by injecting anomalies. We consider the rank parameter equal to 2 in all the experimental results.

5.4.1 Sructural Analysis

Fig. 5.1 shows the plot of results obtained after structural analysis of Geant and Abilene traffic matrices.

1. From the plot of indeterminate structural patterns, we observe that for all the Geant traffic matrices CA captures a lesser number of indeterminate structural patterns (which is desired) as compared to PCA. For Abilene TM, PCA outperforms CA for the majority (19 out of 24) of the TMs. When employed matrix-CUR decomposition on Abilene TMs, the number of indeterminate patterns are the least compared to PCA and CA. A similar result is obtained on Geant data set (except for 4^{th} week and 8^{th} week TMs).

   This experiment demonstrates that both CA and matrix-CUR are able to classify more number of structural patterns into *only one of three categories* with greater accuracy, thereby achieving more orthogonality in the categorization procedure.

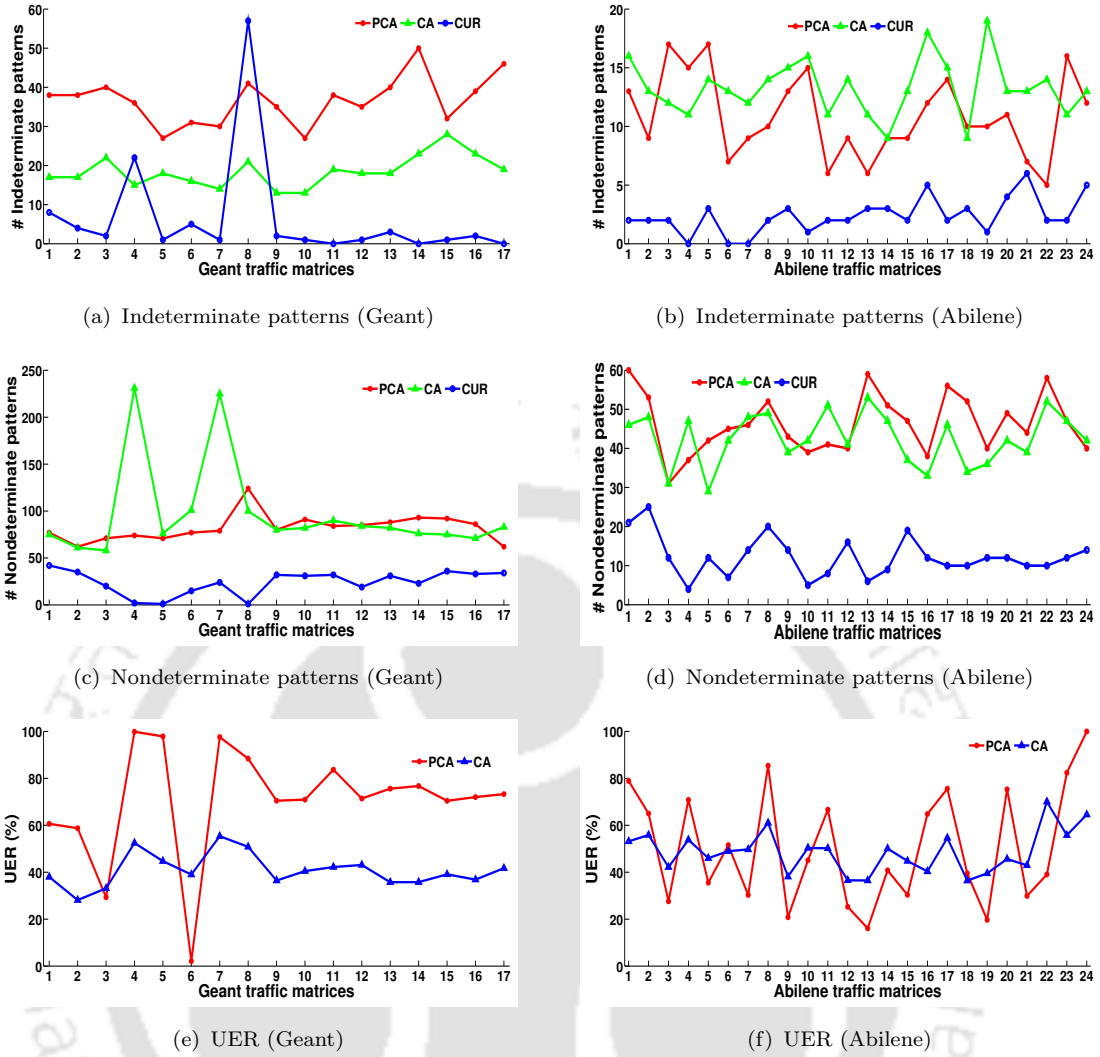


FIGURE 5.1: Structural analysis of Geant and Abilene traffic matrices.

2. From the plot of nondeterminate structural patterns, we clearly observe that matrix-CUR obtains the least number of unclassified structural patterns which is the desired result. Whereas CA closely competes with PCA in terms of the number of nondeterminate structural patters (with majority times obtaining low numbers). This result indicates that CA and matrix-CUR decomposition techniques are able to categorize more number of structural patterns into *any of the three categories* as compared to PCA, thereby attaining more completeness in categorization procedure.

3. From the unclassified energy rate plot, we observe that CA obtains low UER percentage compared to PCA for the Geant dataset (except 6^{th} week

TM). For the Abilene dataset, the average UER percentage using CA is significantly low compared to PCA (lesser UER percentage in 12 out of 24 TMs). This implies that the structural patterns categorized using CA capture more energy as compared to those categorized using PCA for the majority of the TMs.

5.4.2 Volume Anomaly Analysis

Correspondence Analysis (CA):

Fig. 5.2 shows the biplot of Abilene traffic matrix **X10** with red lines denoting PoP pairs and blue dots denoting time intervals. The label for time-intervals is coded in a string of length of six characters. The first character of the string denotes the day of the week of the TM. The rest of the characters in the string denote the ending time interval of measurement. For example, the label 523:05 is interpreted as 5^{th} day of traffic measured during 23:00 to 23:05 time interval. The ending time interval 23:05 is suffixed with the 5^{th} day to form the string 523:05. The first day of the week starts with Saturday as present in the Abilene dataset. Therefore, the day of the week mapping is as follows: {1:Saturday}, {2:Sunday}, {3:Monday}, {4:Tuesday}, {5:Wednesday}, {6:Thursday}, {7:Friday}. The PoP pairs are represented as a string of length of nine characters. The first four characters denote the source PoP code. The last four characters of the string denote the destination PoP code. For example, `CHIN-LOSA` stands for traffic originating at Chicago with destination to Los Angeles. In order to avoid cluttering of the PoP pair labels, we have labeled only two PoP pairs namely, Chicago-Los Angeles (CHIN-LOSA) in the bottom left and Los Angeles-Chicago (LOSA-CHIN) in the bottom right. We observe that these two PoP pairs corresponding to (CHIN-LOSA) and (LOSA-CHIN) lie far from the origin as compared to remaining PoP pairs, *i.e.*, OD flows corresponding to these PoP pairs exhibit the highest standard deviation. In other words, these PoP pairs are experiencing a large volume of anomalous behavior. The time intervals which tend to lie along these vectors (PoP pairs) correspond to the anomalous time intervals, as they lie far from the origin as compared to

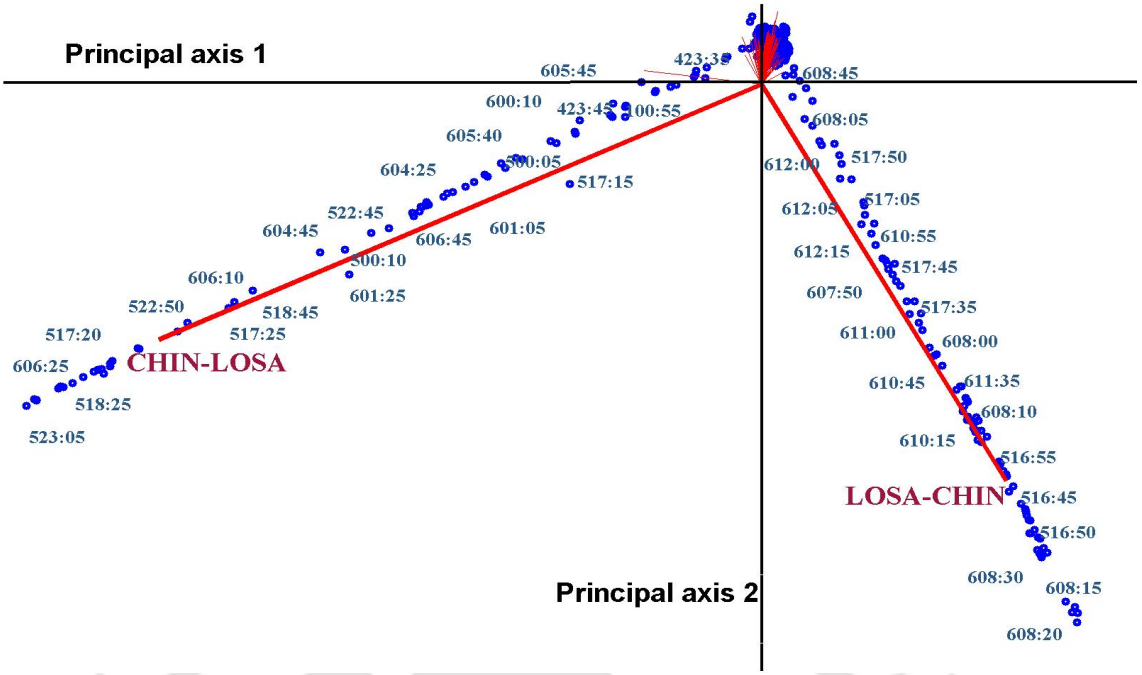
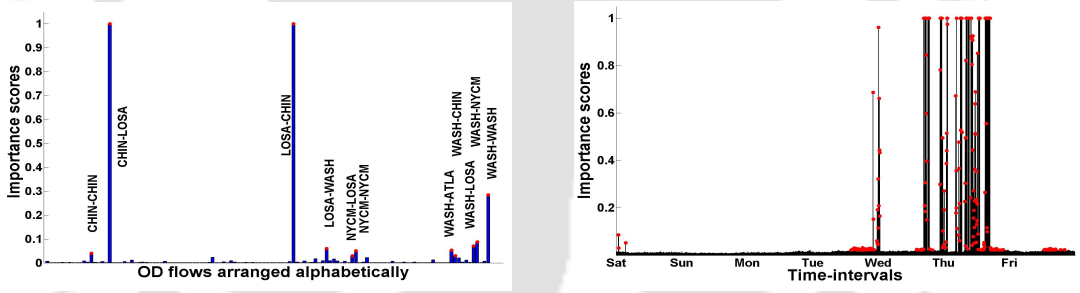


FIGURE 5.2: Biplot of Abilene traffic matrix **X10** using CA



(a) Importance scores of PoP pairs of Abilene traffic matrix

(b) Importance scores of time intervals of Abilene traffic matrix

FIGURE 5.3: Volume anomaly analysis of Abilene traffic matrix.

the remaining time intervals. Anomalies along the CHIN-LOSA PoP pair are observed on Wednesday during ending time intervals (23:05, 18:25, 17:20, 17:25, 18:45, ...) hours and Thursday ending time intervals (06:25, 06:10, 01:25,04:45, ...) hours. Similarly, anomalies along LOSA-CHIN PoP pair are observed on Wednesday during ending time intervals (16:50, 16:45, 16:55, 17:35, 17:45, ...) hours and Thursday ending time intervals (08:20, 08:15, 08:30, 10:15, 08:10, ...) hours.

Matrix-CUR Decomposition:

Fig. 5.3 shows the normalized statistical leverage scores of all columns and rows of

the Abilene traffic matrix. The red dots denote the columns (rows) of original TM which are selected to form the columns (rows) of **C** (**R**). A total of 11 columns and 304 rows are selected to form the matrix **C** and **R**, which are denoted by the red dots in the Fig. 5.3. From Fig. 5.3, we observe that PoP pairs Chicago-Los Angeles (CHIN-LOSA) and Los Angeles-Chicago (LOSA-CHIN) have a normalized statistical leverage score of unity. This implies that OD flows corresponding to these two PoP pairs exhibit high standard deviation, and thus can be regarded as large volume anomalous PoP pairs. We observe anomalies on Wednesday during ending time intervals (23:05, 18:25, 17:20, 17:25, 18:45, 16:50, 16:45, 16:55, 17:35, 17:45, ...) hours and on Thursday ending time intervals (06:25, 06:10, 01:25,04:45, (08:20, 08:15, 08:30, 10:15, 08:10, ...) hours.

Experimentation shows that the anomalous time intervals identified by PCA, CA and matrix-CUR decomposition are almost similar. However, the takeaway point in CA and matrix-CUR decomposition is the identification of PoP pair and time intervals which triggered the anomaly, which is an inherently challenging problem with PCA. Note that the observation obtained using matrix-CUR decomposition and CA go hand in hand. PoP pairs exhibiting anomalous behavior in CA biplot form the matrix **C** in matrix-CUR decomposition. The time intervals identified as anomalous one in CA biplot form the matrix **R** in matrix-CUR decomposition.

In the considered datasets for experimentation, the labeled ground truth information of volume anomalies is not available. In order to compute the false positive rate and anomaly detection rate, we generate traffic matrices synthetically. Methodically inject anomalies between 1 to 50. We adopt three models for generating synthetic traffic matrices using the procedures given in [24]. These are (i) simple TM model, (ii) Modulated Gravity Model (MGM), and (iii) Non-stationary Conditionally Independent Model (NCIM).

We demonstrate the effectiveness of the matrix-CUR decomposition technique for volume anomaly analysis in terms of the total number of detection and false positive rate. Figs. 5.4 and 5.5 show the plot of the total number of detections and false-positive rate. We observe that (i) Matrix-CUR shows superior performance

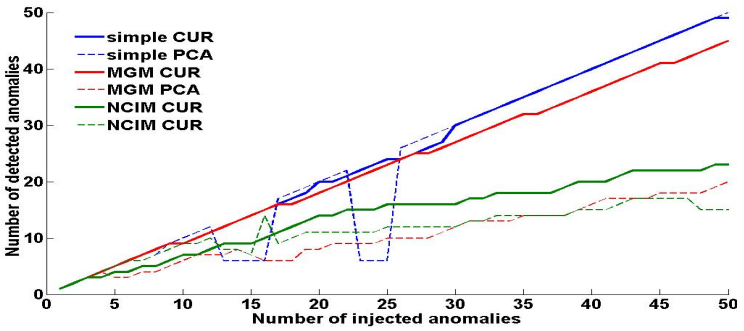


FIGURE 5.4: Total number of detection for injected anomalies of size 100

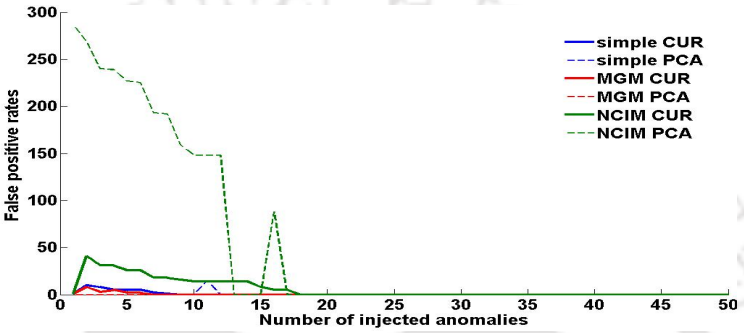


FIGURE 5.5: False positive rate for injected anomalies of size 100

than PCA in terms of the total number of detected anomalies for all three models, except for a few instants only for the simple TM model. (ii) Matrix-CUR decomposition exhibits a low false-positive rate for all three models, except for the MGM and NCIM model for a low number of injected anomalies.

CA, on the other hand, is a qualitative approach and hence can't be compared with matrix-CUR and PCA in terms of the total number of detections and false-positive rates. Experimentation shows that anomalous time intervals and PoP pairs identified using CA are in coherence with those obtained using matrix-CUR decomposition. Overall, our results show that matrix-CUR decomposition is effective with a high detection rate and low false-positive rate for volume anomaly analysis as compared to the PCA-subspace method. Also, it is quite accurate in identifying the time intervals and the underlying PoP pairs and which triggered the volume anomaly.

5.5 Summary

The use of CA and matrix-CUR decompositions have alleviated the most difficult limitation mentioned by [23] in their work on sensitivity of PCA in volume anomaly detection namely interpretation of the obtained subspace. Both CA and matrix-CUR decomposition obtain identical anomalies and identical time intervals during which anomalies occurred. In structural analysis, both CA and matrix-CUR decomposition are able to reduce the number of indeterminate and non-determinate eigenflows and UER percentage. Results show that both these techniques achieve *improved analysis of Internet traffic data* in terms of their corresponding evaluation measures.

Chapter 6

Conclusion and Future Work

This research work is motivated towards alleviation of identified limitations of the existing state-of-the-art decomposition techniques for *estimation, reconstruction, and analysis of Internet traffic data*. We proposed the use of matrix-CUR decomposition to alleviate the rank parameter sensitivity of PCA-based approaches for TM estimation. We proposed to use a multi-view subspace learning technique based on CCA to avoid the overfitting because of the naive use of multiple sources of information for TM estimation. Note that both the proposed techniques do not make use of multiple expensive sources of information for TM estimation as done in recent third-generation techniques. For the reconstruction of missing values, we proposed the use of matrix-CUR decomposition and relative-error bound tensor-CUR decomposition. Note that both the proposed techniques avoid the use of computationally expensive ALS-based procedures. In addition, TCUR-REB does not require the apriori knowledge of tensor rank. For analysis of Internet traffic data, we proposed the use of matrix-CUR decomposition and Correspondence Analysis (CA). Both of these techniques address the limitations of PCA-based approaches, namely (i) lack of interpretability and (ii) assumption of continuous random variables. Extensive experimentation is performed on real and synthetic traffic data to demonstrate the efficacy of the proposed techniques over state-of-the-art techniques.

We present the outcome of the experimental observations for each of the application tasks.

- **Estimation of Internet Traffic Data**

  1. Matrix-CUR decomposition is observed to exhibit less sensitivity to the rank parameter k as compared to PCA-based approaches.
  2. Matrix-CUR decomposition outperforms the state-of-the-art techniques in terms of obtaining low estimation error.
  3. The proposed multi-view learning technique reduces the estimation error by more than 80% compared to the single-view learning technique.

- **Reconstruction of Internet Traffic Data**

  1. Large variation in the error of reconstructing missing values is observed when ALS procedure is employed.
  2. For all loss models and all loss rates, TCUR-REB outperforms the state-of-the-art tensor decomposition techniques in terms of the error in the reconstruction of missing values.
  3. For all loss models and all loss rates, matrix-CUR decomposition outperforms the state-of-the-art matrix decomposition techniques in terms of the error in the reconstruction of missing values.

- **Analysis of Internet Traffic Data**

  1. On majority of the traffic matrices, CA and matrix-CUR decomposition outperform PCA in terms of completeness and orthogonality of structural pattern categorization. In addition, the obtained structural patterns retain the interpretability.
  2. Matrix-CUR decomposition and CA not only identifies the time interval associated with a volume anomaly but also the PoP pairs which triggered the volume anomaly.

3. Experimentation using synthetic TMs shows that the matrix-CUR decomposition is effective with a high detection rate and low false-positive rate for volume anomaly detection as compared to PCA-subspace method. CA, on the other hand, is a qualitative approach, hence can't be used in this comparison.

6.1 Scope for Future Work

The contributions of this thesis can be extended in a number of ways. Some of these possible future research directions are listed below:

- *Multi-view subspace learning for volume anomaly analysis*: It has been cited in literature that the performance of volume anomaly analysis depends on the input source of information. *For instance*, a given anomaly might get detected in OD flow level data but might go undetected in link-level data. Making use of multiple sources of information for volume analysis using a multi-view subspace learning technique will enhance the performance and provide an *improved volume anomaly analysis performance*.

- *Multi-view subspace learning for the reconstruction of missing values*: Missing value problem is inherently present in multiple sources of traffic information. Making use of multiple sources of information using multi-view subspace learning will help in the improved reconstruction of missing values in Internet traffic data.

- *Relative-error bound Tensor-CUR (TCUR-REB) for volume anomaly analysis*: In the future, we will explore the application of this tensor decomposition technique to other applications such as anomaly detection.

Bibliography

[1] "Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper," Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html, Accessed: 2019-02-26 [Online].

[2] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: Methodology and experience," *IEEE/ACM Transactions on Networking (ToN)*, vol. 9, no. 3, pp. 265–280, 2001.

[3] "Cisco Netflow," Available: https://en.wikipedia.org/wiki/NetFlow, Accessed: 2019-02-26 [Online].

[4] "Sampled Netflow," Available: https://sflow.org/, Accessed: 2019-02-26 [Online].

[5] "iperf-The Ultimate Speed Test Tool for TCP, UDP and SCTP," Available: https://iperf.fr/, Accessed: 2019-02-26 [Online].

[6] 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.

[7] "Abilene dataset, http://www.maths.adelaide.edu.au/matthew.roughan/stuff/abilene.tar.gz,."

[8] 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.

[9] Y. Zhang, M. Roughan, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices," in *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4. ACM, 2009, pp. 267–278.

[10] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *ACM SIGMETRICS Performance evaluation review*, vol. 32, no. 1. ACM, 2004, pp. 61–72.

[11] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM computer communication review*, vol. 34, no. 4. ACM, 2004, pp. 219–230.

[12] S. Kolenikov, G. Angeles *et al.*, "The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices," *Chapel Hill: Carolina Population Center, University of North Carolina*, pp. 1–59, 2004.

[13] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[14] 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.

[15] Q. Zhao, Z. Ge, J. Wang, and J. Xu, "Robust traffic matrix estimation with imperfect information: Making use of multiple data sources," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 34, no. 1. ACM, 2006, pp. 133–144.

[16] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[17] H. Zhou, D. Zhang, K. Xie, and Y. Chen, "Spatio-temporal tensor completion for imputing missing internet traffic data," in *2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC)*. IEEE, 2015, pp. 1–7.

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

[19] Z. Wang, K. Hu, K. Xu, B. Yin, and X. Dong, "Structural analysis of network traffic matrix via relaxed principal component pursuit," *Computer Networks*, vol. 56, no. 7, pp. 2049–2067, 2012.

[20] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. ACM, 2009, pp. 1–14.

[21] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*. USENIX Association, 2005, pp. 30–30.

[22] L. Huang, X. Nguyen, M. Garofalakis, M. I. Jordan, A. Joseph, and N. Taft, "In-network PCA and anomaly detection," in *Advances in Neural Information Processing Systems*, 2007, pp. 617–624.

[23] H. Ringberg, A. Soule, J. Rexford, and C. Diot, "Sensitivity of PCA for traffic anomaly detection," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 35, no. 1. ACM, 2007, pp. 109–120.

[24] P. Tune and M. Roughan, "Spatiotemporal traffic matrix synthesis," in *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4. ACM, 2015, pp. 579–592.

[25] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[26] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[27] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[28] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 29.

[29] R. A. Fisher, "The precision of discriminant functions," *Annals of Eugenics*, vol. 10, no. 1, pp. 422–429, 1940.

[30] W. Hatheway, "Contingency-table analysis of rain forest vegetation," in *International Symposium on Stat Ecol New Haven 1969*, 1971.

[31] M. O. Hill, "Reciprocal averaging: an eigenvector method of ordination," *The Journal of Ecology*, pp. 237–249, 1973.

[32] ——, "Correspondence analysis: a neglected multivariate method," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 23, no. 3, pp. 340–354, 1974.

[33] L. Orlóci, "Multivariate analysis in vegetation research. Dr. W," *W. Junk bv, Publishers, The Hague, 451p*, 1978.

[34] S. Nishisato and S. Nishisato, "Analysis of categorical data: Dual scaling and its applications," 1980.

[35] J. Meulman, *Homogeneity analysis of incomplete data*. DSWO Press, 1982, vol. 1.

[36] H. O. Hirschfeld, "A connection between correlation and contingency," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, no. 4. Cambridge University Press, 1935, pp. 520–524.

[37] J.-Y. Bansard, D. Rebholz-Schuhmann, G. Cameron, D. Clark, E. Van Mulligen, F. Beltrame, E. D. H. Barbolla, F. Martin-Sanchez, L. Milanesi, I. Tollis *et al.*, "Medical informatics and bioinformatics: a bibliometric study," *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 3, pp. 237–243, 2007.

[38] K. Fellenberg, N. C. Hauser, B. Brors, A. Neutzner, J. D. Hoheisel, and M. Vingron, "Correspondence analysis applied to microarray data," *Proceedings of the National Academy of Sciences*, vol. 98, no. 19, pp. 10 781–10 786, 2001.

[39] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular biology of the cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[40] E. J. Beh, "Simple correspondence analysis: a bibliographic review," *International Statistical Review*, vol. 72, no. 2, pp. 257–284, 2004.

[41] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Springer, 1971, pp. 134–151.

[42] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, vol. 58, no. 3, pp. 453–467, 1971.

[43] N. Mitrovic, M. T. Asif, U. Rasheed, J. Dauwels, and P. Jaillet, "CUR decomposition for compression and compressed sensing of large-scale traffic data," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, 2013, pp. 1475–1480.

[44] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations," *Journal of the ACM (JACM)*, vol. 51, no. 6, pp. 1025–1041, 2004.

[45] R. Penrose, "A generalized inverse for matrices," in *Mathematical proceedings of the Cambridge philosophical society*, vol. 51, no. 3. Cambridge University Press, 1955, pp. 406–413.

[46] P. Drineas, R. Kannan, and M. W. Mahoney, "Fast Monte Carlo algorithms for matrices III: Computing a compressed approximate matrix decomposition," *SIAM Journal on Computing*, vol. 36, no. 1, pp. 184–206, 2006.

[47] P. Paschou, M. W. Mahoney, A. Javed, J. R. Kidd, A. J. Pakstis, S. Gu, K. K. Kidd, and P. Drineas, "Intra-and interpopulation genotype reconstruction from tagging SNPs," *Genome Research*, vol. 17, no. 1, pp. 96–107, 2007.

[48] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.

[49] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*. Springer, 1992, pp. 162–190.

[50] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[51] W. W. Sun, J. Lu, H. Liu, and G. Cheng, "Provable sparse tensor decomposition," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 3, pp. 899–916, 2017.

[52] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *2010 IEEE International Conference on Image Processing*. IEEE, 2010, pp. 517–520.

[53] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[54] 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.

[55] M. T. Asif, N. Mitrovic, J. Dauwels, and P. Jaillet, "Matrix and tensor based methods for missing data estimation in large traffic networks," *IEEE Transactions on intelligent transportation systems*, vol. 17, no. 7, pp. 1816–1825, 2016.

[56] K. Xie, X. Li, X. Wang, G. Xie, J. Wen, J. Cao, and D. Zhang, "Fast tensor factorization for accurate internet anomaly detection," *IEEE/ACM Transactions on Networking (TON)*, vol. 25, no. 6, pp. 3794–3807, 2017.

[57] H. Kasai, W. Kellerer, and M. Kleinsteuber, "Network volume anomaly detection and identification in large-scale networks based on online time-structured traffic tensor tracking," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 636–650, 2016.

[58] K. Xie, L. Wang, X. Wang, G. Xie, J. Wen, G. Zhang, J. Cao, D. Zhang, K. Xie, X. Wang *et al.*, "Accurate recovery of internet traffic data: A sequential tensor completion approach," *IEEE/ACM Transactions on Networking (TON)*, vol. 26, no. 2, pp. 793–806, 2018.

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

[60] P. Tune, M. Roughan, H. Haddadi, and O. Bonaventure, "Internet traffic matrices: A primer," *Recent Advances in Networking*, vol. 1, pp. 1–56, 2013.

[61] C. Tebaldi and M. West, "Bayesian inference on network traffic using link count data," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 557–573, 1998.

[62] J. Cao, D. Davis, S. Vander Wiel, and B. Yu, "Time-varying network tomography: router link data," *Journal of the American statistical association*, vol. 95, no. 452, pp. 1063–1075, 2000.

[63] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang, "Experience in measuring internet backbone traffic variability: Models metrics, measurements and meaning," in *Teletraffic Science and Engineering*. Elsevier, 2003, vol. 5, pp. 379–388.

[64] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM, 2003, pp. 301–312.

[65] A. Soule, A. Nucci, R. Cruz, E. Leonardi, and N. Taft, "How to identify and estimate the largest traffic matrix elements in a dynamic environment," *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 1, pp. 73–84, 2004.

[66] L. Tan and X. Wang, "A novel method to estimate IP traffic matrix," *IEEE Communications Letters*, vol. 11, no. 11, pp. 907–909, 2007.

[67] M. Malboubi, "Optimal-coherent network inference (OCNI): Principles and applications," *IEEE Transactions on Network and Service Management*, vol. 15, no. 2, pp. 811–824, 2018.

[68] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[69] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 129–136.

[70] W. Liu, X. Yang, D. Tao, J. Cheng, and Y. Tang, "Multiview dimension reduction via Hessian multiset canonical correlations," *Information Fusion*, vol. 41, pp. 119–128, 2018.

[71] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 251–260.

[72] Y.-H. Yuan, Y. Li, X.-B. Shen, Q.-S. Sun, and J.-L. Yang, "Laplacian multiset canonical correlations for multiview feature extraction and image recognition," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 731–755, 2017.

[73] D. M. Witten, R. Tibshirani, and T. Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[74] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 462–471.

[75] Y. Niu and H. Tian, "Study on a new model for network traffic matrix estimation," in *2014 Sixth International Symposium on Parallel Architectures, Algorithms and Programming*. IEEE, 2014, pp. 152–154.

[76] A. Gunnar, M. Johansson, and T. Telkamp, "Traffic matrix estimation on a large IP backbone: a comparison on real data," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 149–160.

[77] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast accurate computation of large-scale IP traffic matrices from link loads," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 31, no. 1. ACM, 2003, pp. 206–217.

[78] K. Papagiannaki, N. Taft, and A. Lakhina, "A distributed approach to measure IP traffic matrices," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 161–174.

[79] E. Zhao and L. Tan, "A PCA based optimization approach for IP traffic matrix estimation," *Journal of Network and Computer Applications*, vol. 57, pp. 12–20, 2015.

[80] D. Jiang, X. Wang, L. Guo, H. Ni, and Z. Chen, "Accurate estimation of large-scale IP traffic matrix," *AEU-International Journal of Electronics and Communications*, vol. 65, no. 1, pp. 75–86, 2011.

[81] H. Zhou, L. Tan, Q. Zeng, and C. Wu, "Traffic matrix estimation: A neural network approach with extended input and expectation maximization iteration," *Journal of Network and Computer Applications*, vol. 60, pp. 220–232, 2016.

[82] L. Nie, D. Jiang, L. Guo, and S. Yu, "Traffic matrix prediction and estimation based on deep learning in large-scale IP backbone networks," *Journal of Network and Computer Applications*, vol. 76, pp. 16–22, 2016.

[83] L. Nie, Y. Li, and X. Kong, "Spatio-temporal network traffic estimation and anomaly detection based on convolutional neural network in vehicular ad-hoc networks," *IEEE Access*, vol. 6, pp. 40 168–40 176, 2018.

[84] Y. Tian, W. Chen, and C.-T. Lea, "An SDN-based traffic matrix estimation framework," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1435–1445, 2018.

[85] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM, 2003, pp. 301–312.

[86] G. Michau, N. Pustelnik, P. Borgnat, P. Abry, A. Nantes, A. Bhaskar, and E. Chung, "A primal-dual algorithm for link dependent origin destination matrix estimation," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 1, pp. 104–113, 2017.

[87] 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.

[88] J. Sun, D. Tao, and C. Faloutsos, "Beyond streams and graphs: dynamic tensor analysis," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 374–383.

[89] H. Zhou, D. Zhang, K. Xie, and Y. Chen, "Robust spatio-temporal tensor recovery for internet traffic data," in *2016 IEEE Trustcom/BigDataSE/ISPA*. IEEE, 2016, pp. 1404–1411.

[90] N. Hao, M. E. Kilmer, K. Braman, and R. C. Hoover, "Facial recognition using tensor-tensor decompositions," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 437–463, 2013.

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

[92] G. Gürsun and M. Crovella, "On traffic matrix completion in the internet," in *Proceedings of the 2012 ACM conference on Internet measurement conference*. ACM, 2012, pp. 399–412.

[93] M. Mardani and G. B. Giannakis, "Robust network traffic estimation via sparsity and low rank," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4529–4533.

[94] R. Du, C. Chen, B. Yang, and X. Guan, "Vanet based traffic estimation: A matrix completion approach," in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 30–35.

[95] Y.-C. Chen, L. Qiu, Y. Zhang, G. Xue, and Z. Hu, "Robust network compressive sensing," in *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 2014, pp. 545–556.

[96] 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*. IEEE, 2015, pp. 2443–2451.

[97] M. W. Mahoney, M. Maggioni, and P. Drineas, "Tensor-CUR decompositions for tensor-based data," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 957–987, 2008.

[98] L. Wang, K. Xie, T. Semong, and H. Zhou, "Missing data recovery based on tensor-CUR decomposition," *IEEE Access*, vol. 6, pp. 532–544, 2018.

[99] E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, "Tensors for data mining and data fusion: Models, applications, and scalable algorithms," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 2, p. 16, 2017.

[100] H. Kim, S. Lee, X. Ma, and C. Wang, "Higher-order PCA for anomaly detection in large-scale networks," in *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2009, pp. 85–88.

[101] M. Roughan, "A case study of the accuracy of SNMP measurements," *Journal of Electrical and Computer Engineering*, vol. 2010, p. 33, 2010.

[102] H. Zhou, D. Zhang, and K. Xie, "Accurate traffic matrix completion based on multi-gaussian models," *Computer Communications*, vol. 102, pp. 165–176, 2017.

[103] B. W. Bader, T. G. Kolda *et al.*, "Matlab tensor toolbox version 2.5," *Available online, January*, vol. 7, 2012.

[104] A. Lakhina, M. Crovella, and C. Diot, "Characterization of network-wide anomalies in traffic flows," in *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. ACM, 2004, pp. 201–206.

[105] ——, "Mining anomalies using traffic feature distributions," in *ACM SIGCOMM computer communication review*, vol. 35, no. 4. ACM, 2005, pp. 217–228.

Publications Related to Thesis

Journal(s):

Accepted:

- **Awnish Kumar**, Sandeep Vidyapu, Vijaya V. Saradhi, and T. Venkatesh. "A Multi-view Subspace Learning Approach to Internet Traffic Matrix Estimation", *IEEE Transactions on Network and Service Management 17*, no. 2 (2020): 1282-1293.

Under Review:

- **Awnish Kumar**, Vijaya V. Saradhi, and T. Venkatesh. "Compressive Sensing of Internet Traffic Data using Relative-error Bound Tensor-CUR Decomposition", *Journal of Network and Service Management*, 2021.

Conferences:

- **Awnish Kumar**, V. Vijaya Saradhi, and T. Venkatesh. "Role of Correspondence Analysis in Network Traffic Flow Analysis", *Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE)*. ACM, 2014.
- **Awnish Kumar**, Vijaya V. Saradhi, and T. Venkatesh. "Network-wide Volume Anomaly Detection using Alternate Matrix Decomposition Techniques" *IEEE ANTS 2017*.
- **Awnish Kumar**, Vijaya V. Saradhi, and T. Venkatesh. "Compressive Sensing of Internet Traffic Matrices using CUR Decomposition", *Proceedings of the 19th International Conference on Distributed Computing and Networking*. ACM, 2018.

Workshops:

- **Awnish Kumar**, Vijaya V. Saradhi, and T. Venkatesh. "Interpretable Structural Analysis of Traffic Matrix", *ICML 2017 Time Series Workshop*, Sydney, Australia, 2017.