This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30045549

Reproduced with the kind permission of the copyright owner.

Copyright: 2009, Curtin University of Technology
Scalable Network-Wide Anomaly Detection Using Compressed Data

D. S. Pham, B. Saha, M. Lazarescu, and S. Venkatesh

Abstract

Detecting network traffic volume anomalies in real time is a key problem as it enables measures to be taken to prevent network congestion which severely affects the end users. Several techniques based on principal component analysis (PCA) have been outlined in the past which detect volume anomalies as outliers in the residual subspace. However, these methods are not scalable to networks with a large number of links. We address this scalability issue with a new approach inspired from the recently developed compressed sensing (CS) theory. This theory induces a universal information sampling scheme right at the network sensory level to reduce the data overhead. Specifically, we address exploit the compressibility characteristics of the network data and describe a framework for anomaly detection in the compressed domain. Our main theoretical contribution is a detailed theoretical analysis of the new approach which obtains the probabilistic bounds on the principal eigenvalues of the compressed data. Subsequently, we prove that volume anomaly detection using compressed data can achieve equivalent performance as it does using the original uncompressed and reduces the computational cost significantly. The experimental results on both the Abiliene and synthetic datasets support our theoretical findings and demonstrate the advantages of the new approach over the existing methods.

Updated Feb 9, 2009

The authors are with the Institute for Multisensor and Content Analysis (IMPCA), Curtin University of Technology, PO Box U1987, WA 6845 (email: dspham@ieee.org).
I. Introduction

Anomaly detection in wide-area network is a crucial and challenging task nowadays. A volume anomaly is a significant change in the underlying network traffic, which could result in network congestion and affects the end users [16]. To address the problem of anomaly detection, a framework is needed to enable the detection of anomalies in *real* time in order to allow the prompt modification of the border gateway protocol (BGP) routing table. Given the high volume of traffic and the complexity of the network, the process of detecting anomalies using the raw data is both computationally expensive and requires significant storage resources. In this paper we present an approach aimed at detecting *volume* anomalies in wide area network while addressing three key aspects of the problem: (i) the high dimensionality of the data, (ii) the limited storage resources and (iii) effect of the noise on the data.

A typical network is monitored by local nodes (e.g. routers) where each node sends a time-varying data stream to the central location for processing. The central processing unit takes the decision after processing all globally collected streams. In a $N$-link network, link data measures the traffic flows in each edge link. Most Internet service providers (ISPs) use simple network management protocols (SNMP’s) to collect the average traffic volumes transmitted between two point-of-presence (PoP’s) in a network. The problem is that the data stream aggregated from $N$ links of the network requires to be processed in real time. In many real world cases the size of the network is *very large* both in terms of the number of nodes and links. For example, AT&T’s has 10,214 routers and 12,500 links [21] which is directly connected customer access routers. Another example is the skitter tool of CAIDA [2] which collected data from a IP graph of 629,647 nodes and 1,230,572 links.

However, most current techniques to process the data from network links are not scalable to large networks as the ones mentioned above, and are only suited for small networks. For example, Lakhina et al. [16] use principal component analysis (PCA) to detect anomalies in the *residual subspace* of the original data. While encouraging results have been shown for small-size problems, this PCA approach suffers from two disadvantages. First, it is not scalable to network size due to the computation of the large covariance matrix required to obtain the projection onto the residual subspace. Specifically, the eigenvalue decomposition complexity is $O(N^3)$ and requires a memory storage of $O(N^2)$. Second, it requires to span the aggregation window of size greater than $N$ and thus fine-grain time resolution (less than $N$) anomaly detection is poor.

We propose a novel framework for anomaly detection in the compressed domain to tackle the scalability issue in large-sized networks. Our work has been motivated by the principles of a recently proposed
information sampling theory called Compressed Sensing (CS) [3], [7]. This theory can be used to reduce the dimensionality of network data considerably. Our theoretical contribution is a result on the approximate preservation of the principal subspace in the compressed domain which implies that techniques developed in the original domain can be readily used in the compressed domain. Thus, anomaly detection can be analogous to the uncompressed case, but with the advantage of a lower number of measurements. In addition, the lower number of measurements also allows for anomaly detection at a finer time resolution. Importantly, the computational complexity is sub-linear with the number of links \( N \).

In the CS framework, the \( N \)-dimensional data is captured directly via a non-adaptive and simple linear projection such that the number of measurements \( M \) is only proportional to the intrinsic information level in the signal. The intrinsic information level is reflected in its sparsity \( K \), implying that the signal has only few non-zero coefficients in some basis.

There are two supporting observations for considering the CS framework in network data processing. First, it has been previously shown by [16] that high dimensional network flows can be represented by few intrinsic dimensions. Hence it is reasonable to assume that the network data is approximately sparse and only a small number of non-adaptive measurements is needed to retain information about the main traffic. Second, the restricted isometry property (RIP) (see Appendix) helps in retaining the geometry of the data structure in low dimensional subspace. We note that while the CS theory is primarily developed for reconstruction of sparse signals from compressive samples, our aim is to detect volume anomalies and it would be an advantageous to work directly with the compressed data. The CS theory plays a key role as information-preserving compression techniques in our framework. We will be using a PCA technique in combination with RIP to detect anomalies in the compressed domain. The residual subspace still retains the noise-like characteristics which is sufficient for anomaly detection. Thus, by working in the CS domain, the rank of the new covariance matrix is \( M \) and the computational complexity is reduced to \( \mathcal{O}(M^3) \) where \( M \ll N \).

We evaluate our algorithm on real traffic traces collected from the Abilene network\(^1\) over four weeks and synthetic data simulated following the property of typical networks. Our experiment verifies that on the real dataset the proposed method using compressed data achieves equivalent performance with a detection rate of more than 94%. For synthetic data, our experiments shows that the PCA technique performs even better in compressed domain than uncompressed domain for high dimensional data. For example, the detection rate and false positive rates are 98.4% and 2% respectively for a data of 1000

---

\(^1\)www.abilene.iu.edu
links and 2000 snapshots in compressed domain, compared to detection rate 98.2% and false positive rate 9% in the uncompressed domain. Most importantly, the proposed method requires less memory storage and can be as 100 times faster than the PCA method using the raw data.

The paper is organized as follows. In Section 2 we discuss related prior work. Section 3 describes the problem in detail and provides background information on CS and network anomaly detection. Section 4 explains our proposed method and its analysis. Section 5 describes the data sets, experimental setup and results while the conclusions are presented in Section 6.

II. RELATED WORK

Network anomaly detection is critical in managing IP networks. The first step is feature extraction to separate major traffic for better diagnosis of anomalies. Lakhina et al. [16] characterize network anomalies and use PCA to find a projection to the residual subspace for anomaly detection. This work however uses the original data and is only suitable for small-sized networks. Ling et al. [13] propose a decentralized version of the PCA approach and design a distributed protocol for reducing the communication overheads between the nodes and the central unit. Whilst it reduces the communication cost, it still needs to solve an eigenvalue problem of a perturbation matrix which scales significantly with the network-size. Zhang et al. [25] has also used PCA based subspace method for spatial anomaly detection. He has also provided a comparative performance study of four different models (e.g. ARIMA, Fourier, Wavelet and PCA) for temporal anomaly detection. Donoho [24] has introduced estimating traffic matrix from link count data by exploiting the maximum sparsity of the network. Zhang [25] used \( l_1 \) norm for minimizing the sparsity constraint of OD (origin-destination) flow matrix and used subspace based techniques directly on OD flows for identifying anomaly. The drawback of above algorithms are that they never addressed the scalability issues of the network. Li et al. [17] propose a sketch subspace method which uses a set of randomly chosen universal hash functions for dimensionality reduction and the subspace based method [16] for anomaly detection. However, the drawback of the approach is that it relies on heuristics when selecting the hash functions. An alternative approach has been described in [15] which uses random projections (RP). However, this method cannot capture the intrinsic structure of the underlying data as outlined by [16].

In the second step of making statistical decision, the \( Q \)-statistic has been used in PCA techniques [13], [16], [17] due to its simplicity and robustness in practice, particularly for network data. Other generic techniques have been used in other applications such as SVM for EEG data analysis [9] and support vector regression for on-line anomaly detection of time series data [19].
It is important to note that there are other approximate SVD routines to compute the principal subspace such as incomplete Cholesky [18] or Lanczos [11]. However, their complexity is still depends on the network size $N$ and our goal is not about numerical SVD routines but rather proving that it is possible to work on compressed data for the approximate performance, which is much less dependent of the network size. These approximate SVD techniques can also be readily applied to the compressed data to even further reduce computational cost.

III. BACKGROUND AND PROBLEM IDENTIFICATION

A. PCA and Network Anomaly Detection

Lakhina et al. [16] observed that volume anomalies are rare events in high-level traffic aggregation and ‘hidden’ in normal traffic. However, as most of the normal traffic is found in a low-dimensional subspace, PCA can be used to separate the residual subspace (which reflects the local fluctuations) from the principal traffic (which reflects the long-term trend) so that anomalies can be more easily detected. $L$ snapshots of the traffic from $N$ links $\mathbf{x}_i \in \mathbb{R}^N$, $i = 1, \ldots, L$ are collected in an aggregated matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$ from which the sample covariance matrix can be estimated $\Sigma_\mathbf{x} = (1/L)\mathbf{X}\mathbf{X}^T$ after $\mathbf{X}$ is centralized. Given that $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ are the principal eigenvectors of $\Sigma_\mathbf{x}$ corresponding to the largest $K$ eigenvalues $\lambda_1, \ldots, \lambda_K$, the projection onto the residual subspace is $\mathbf{P} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)$. Thus, for any observed data $\mathbf{x}$, its projection into the residual subspace is $\mathbf{z} = \mathbf{P}\mathbf{x} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{x}$. If $\mathbf{z}$ follows a multivariate normal distributed, the squared prediction error (SPE) statistic is given as

$$ t_{SPE} = \|\mathbf{z}\|^2 = \|(\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{x}\|^2 $$

and follows a noncentral chi-square distribution under the null hypothesis that the data is ‘normal’. Hence, rejection of the null hypothesis can be based on whether $t_{SPE}$ exceeds a certain threshold corresponding to a desired false alarm rate $\beta$. In [16], the $Q$-statistic is used to compute the threshold, which involves the largest $K$ eigenvalues, i.e. $Q_\beta = Q_\beta(\lambda_1, \ldots, \lambda_K)$ (see (4)). An anomaly is detected when $t_{SPE} > Q_\beta$.

B. Compressed Sensing

In compressed sensing, the original signal is directly recorded in a compressed format and the number of stored samples is proportional to the information level in original signal. Let us assume that $\mathbf{x} \in \mathbb{R}^N$ admits a linear representation by a set of orthonormal basis functions $\Psi$ with coefficients $\alpha$, i.e. $\mathbf{x} = \Phi\alpha$. Two cases of interest are i) Sparse signal: the signal $\mathbf{x}$ is said to be $K$-sparse if only $K$ entries of $\alpha$ are nonzero; and ii) Compressible signal: the magnitudes of the coefficients $\alpha$, when ordered, follow

February 10, 2009 DRAFT
an exponential decay [5]. When $x$ is sparse or compressible, CS theory [5], [7], [12] has proved that it is possible to ‘sense’ $x$ via a simple, non-adaptive and linear projection $y = \Phi x$. The sensing matrix $\Phi \in \mathbb{R}^{M \times N}$ has a significantly smaller number of rows than columns, i.e., $M \ll N$, meaning that the dimension of $y$ is considerably smaller than $x$. Importantly, it is possible to perfectly recover $x$ from $y$ under suitable conditions on the sensing matrix $\Phi$. This implies that all the salient information about $x$ is captured in $y$, making CS an universal dimensionality reduction technique. When classification is needed instead of recovery, the use of CS is clearly an advantage as the dimensionality of the problem is reduced (in practice, $M = O(K \log N) \ll N$, and $K$ is a reasonable upper bound on the sparsity). As the network data is shown to be compressible, we are motivated to use CS to address the scalability issue in the original PCA approach. In the following sections we describe our proposed method that employs both CS and PCA for the network anomaly detection problem.

Remarks: There is a possibility of considering other "dimensionality reduction" techniques. Since most dimensionality reduction techniques are adaptive, computationally expensive and likely lossy. In contrast to new CS techniques is non adaptive, simple and universal and it allows the reconstruction of original data if it is required and most importantly it can be done at sensing stage. We also note that CS should not be confused with random projection (RP). Random projection (RP) is used in CS as the measuremen matrix because they satisfy the restricted isometry property with high probability, but RP is not the only choice. Recent progress shows that the CS measurement matrix can be deterministically constructed [6]. Furthermore, CS theory [7] is concerned with a deep mathematical question on the inverse problem. RP is only concerned with length preservation of the projection.
IV. PROPOSED APPROACH AND ANALYSIS

Our proposed approach\(^2\) is illustrated in Fig. 1. A sensing matrix \( \Phi \) is used to obtain a compressed version of the traffic from every node \( y = \Phi x \). PCA is then performed on \( y \) and anomaly detection is performed in the derived residual subspace.

A. Design of the sensing matrix \( \Phi \)

Let signal \( x \) be sparse in the basis \( \mathbf{I} \). To effectively capture the information about an \( K \)-sparse signal \( x \), CS theory requires each \( 2K \) columns of \( \Phi \) should behave like an orthogonal system. Equivalently, this requires the mutual coherence of the overcomplete system \( \Phi = [\phi_1, \ldots, \phi_N] \), which is defined as

\[
\mu(\Phi) = \max_{i \neq j} |\langle \phi_j, \phi_k \rangle|,
\]

be as small as possible. The problem of designing an overcomplete system \( \Phi \) with a small mutual coherence is known in information theory as frame design [22]. For a real-valued matrix \( \Phi \), the lower bound on the mutual coherence is known as the Welch bound [22]

\[
\mu(\Phi) \geq \sqrt{(N - M)/(M(N - 1))}.
\]

In practice, the Welch bound may not exist for every pair \((M, N)\) and even if it exists, designing \( \Phi \) which meets the Welch bound still remains a difficult task. The challenge is to maintain at the same time a small number of measurements \( M \) and low mutual coherence. CS theory [5] overcomes this difficulty by exploiting the fact that random matrices satisfy the RIP condition (see Appendix) with high probability. Examples of the random matrices include random Gaussian, random Bernoulli, and random partial Fourier matrices. However, this is only an asymptotic result, i.e. suitable for problems where \( N \) is large. For problems of smaller size, a particular realization might not achieve small mutual coherence. Thus, in order to obtain a good sensing matrix, we start with a random Gaussian matrix and then apply the recently proposed algorithm by Elad [8]. This algorithm exploits the fact that the mutual coherence of \( \Phi \), with each column normalized to unit norm, is the maximum magnitude of the off-diagonal elements of the Gram matrix \( G = \Phi^T \Phi \) where the Gram matrix has the rank \( M \). Hence, by iteratively shrinking the entries of the Gram matrix, forcing its rank to \( M \), and taking square root, a smaller mutual coherence

\(^2\)The diagram shown in the paper is only symbolic and we suggest that CS projection could be done either at the central node or within the network. If done at the central node, there are "database-friendly" projection matrices of mostly 0's (with \( p=2/3 \)) and few 1's (with \( p=1/6 \)), meaning that it can be efficiently implemented. If done within the network like the randomized gossip algorithm [20], [1], the extra bandwidth is needed only locally and hence still feasible.
Fig. 2. Plot of eigenvalues

for $\Phi$ with a specified rank $M$ is achieved. Even though the algorithm could be sensitive to the parameter setting and its convergence is yet to be studied, we found in practice that this method can improve the mutual coherence quite considerably.

In practice, the actual signal $x$ might not be sparse in the basis $I$ but in some $\Psi$. In this case, CS theory requires $\mu(\Phi\Psi)$ to be small instead. If $\Phi_o$ is the optimal sensing matrix for the basis $I$ then the optimal matrix $\Phi$ for the basis $\Psi$ is found from $\Phi = \Phi_o\Psi^{-1}$, assuming that $\Psi$ is invertible.

B. Anomaly detection with CS data

The PCA approach is now applied to the compressed data $Y = [y_1, \ldots, y_L]$ where $y = \Phi x$. Suppose that the compressed data is centralized, the projection matrix is found from the eigenvalue decomposition of the CS covariance matrix $\Sigma_y = (1/L)YY^T$. As our analysis shows subsequently, the number of principal eigenvectors remains the same in the CS domain and that the SPE statistic in the CS domain also approximately follows the non-central $\chi^2$ distribution. Hence, it is sensible to use $Q$-statistic to set
the threshold for anomaly detection. It is based on $(1 - \beta)$ confidence level of standard normal distribution where $\beta$ is the desired false alarm rate:

$$Q_{\beta} = \theta_1 \left[ c_{\beta} \sqrt{2\theta_2 h_0^2} \theta_1 + \frac{\theta_2 h_0(h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}}. \quad (4)$$

where, $h_0 = 1 - \frac{2r \theta_i}{\theta_0}$, $\theta_i = \sum_{j=K+1}^{N} \xi_j^i$ for $i = 1, 2, 3$, $c_{\beta} = (1 - \beta)$ percentile in a standard normal distribution and $Q_{\beta}$, and $\xi_j, i = 1, \ldots, M$ are the eigenvalues of $\Sigma_y$.

C. Temporal analysis of PCA in CS domain

In practice, the projection onto the residual and principal subspaces are obtained from the eigenvalue decomposition of the sample covariance matrix. When the data is sufficiently large, it follows from asymptotic theory that the subspaces computed approach the true one. However, it is more likely in
practice that the covariance is computed over a window of finite length $L$. The finite sample effect states that the quality of the covariance matrix depends on how large $L$ compared with the dimension of the data $N$. When $L$ is relatively small, the empirical distribution of the eigenvalues, and hence eigenvectors, can be severely distorted, i.e. large eigenvalues tend to be too large and small eigenvalues tend to be too small. When $L$ is smaller than the dimension of the data $N$, the empirical covariance matrix even becomes rank-deficient. Even though one might still be able to compute a few principal eigenvalues and eigenvectors, this rank deficiency implies larger error on the empirical values. On the other hand, if the data is embedded to a much lower dimension $M$ via an information-preserving projection (such as the CS projection), the empirical covariance matrix in the compressed domain may not suffer from this rank deficiency problem as long as the window length $L > M$. As a consequent, the empirical distribution of the eigenvalues and eigenvectors may be less distorted, which means that the computation of the principal subspace will have a smaller error. For the problem of anomaly detection in wide area network, this means fine grain time resolution can be achieved in CS domain. This will be demonstrated subsequently in the experimental section.

D. Analysis

In this section, we provide a detailed analysis to enable a better understanding of the original PCA method and offer a justification as to why CS allows anomaly detection in the compressed domain.

First, we consider a model for the network data as follows

$$\mathbf{x} = \mathbf{s} + \mathbf{n}$$

where $\mathbf{x} \in \mathbb{R}^N$ is the snapshot of network traffic over $N$ links. It consists of two parts: $\mathbf{s}$ characterizes the long-term structure of the traffic and $\mathbf{n}$ represents the locally temporal variation which occurs in all links. Previous studies have found that $\mathbf{s}$ lies in a small dimensional space while $\mathbf{n}$ has a noise-like behavior. Hence, it is reasonable to assume that $\mathbf{s}$ is sparse in some basis $\Psi$ and $\mathbf{n}$ are iid Gaussian noise with variance $\sigma^2$. Further, we can assume without loss of generality that

$$\mathbf{s} = \Psi \alpha_s = \begin{bmatrix} \Psi_s \\ \Psi_n \end{bmatrix} \begin{bmatrix} \alpha_s \\ 0 \end{bmatrix} = \Psi_s \alpha_s$$

(6)

where $\alpha_s \in \mathbb{R}^K$ represents the $K$ nonzero entries of the coefficients. The covariance matrix of $\mathbf{x}$ is

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} \Psi_s & \Psi_n \end{bmatrix} \begin{bmatrix} \Lambda_K + \sigma^2 I_K \\ \sigma^2 I_{N-K} \end{bmatrix} \begin{bmatrix} \Psi_s^T \\ \Psi_n^T \end{bmatrix}$$

(7)
where $\Lambda_K = \text{diag}\{\alpha_{s,i}^2\} \in \mathbb{R}^{K \times K}$. Each term $\lambda_i = E[\alpha_{s,i}^2], i = 1, \ldots, K$ represents the power distribution in the principal subspace. It is clear that $\Psi_s$ is also the eigenvector corresponding to the $K$-principal subspace. The principal eigenvalues of $\Sigma_x$ are $\lambda_1 + \sigma^2, \ldots, \lambda_K + \sigma^2, \sigma^2, \ldots, \sigma^2$. In the PCA method, when there is sufficient data, the sample principal eigenvectors tend to $\Psi_s$ hence the data is projected to the residual subspace using the projection matrix $P = I - \Psi_s \Psi_s^T$ resulting in

$$z = Px = (I - \Psi_s \Psi_s^T)(\Psi_s \alpha_s + n) = Pn.$$ (8)

As $n \sim \mathcal{N}(0, \sigma^2 I)$ we have $z \sim \mathcal{N}(0, \sigma^2 PP^T)$. As $z$ follows the multivariate normal distribution, we know that $t = \|z\|^2_2$ follows a noncentral $\chi^2$ distribution. For anomaly detection in the residual subspace, the tail behavior of this distribution is important in setting a suitable threshold for a certain desired false alarm rate. Previous works use a normal approximation for the tail of this noncentral $\chi^2$ distribution. For our analysis, we concentrate on the energy aspect of this residual subspace. The following result is intermediate.

**Lemma 1:** $E[t] = (N - K)\sigma^2$. The lemma states that the mean of the statistic used for anomaly detection is equal to the power of the noise spread in the residual subspace as can be readily seen from (7). Volume anomalies are thus detected through the changes in the total power of the residual subspace.

Next, we characterize the subspaces in the compressed domain. We show that under CS assumptions, the principal subspace is approximately preserved while the residual subspace, though not preserved, still has a noise-like behavior. This suggests that in the CS domain, subtracting the principal subspace for anomaly detection is equivalent to that in the original domain, which is our argument for doing anomaly detection in the CS domain.

We make the following assumptions. Firstly, we assume that the noise is small compared with the principal network traffic, i.e. $\lambda \gg \sigma^2$, which is consistent with real data. Secondly, we assume\(^3\) the sensing matrix $\Phi$ contains iid Gaussian entries with variance of $1/N$. Finally, we assume that $M = \mathcal{O}(K \log N)$\(^7\), which is the standard setting in CS. The basis for our argument is the following result.

**Theorem 1:** For a network with a sufficiently large number of links $N$, the principal subspace is approximately preserved in a sense that with probability $1 - \delta$, (where $\delta > 0$) the change in principal

\(^3\)We note that this is different from the standard CS choice of $1/M$. This is just to simplify our arguments in subsequent analysis only in a sense to prove the eigenvalues remain the same. Effectively, $\Phi$ is scaled down by a factor $\sqrt{N/M}$ but anomaly detection is obviously invariant to scaling so it introduces no technical difficulty.
eigenvalues are bounded by
\[ |\lambda_i - \xi_i| \leq \sqrt{2}\lambda_1 (3\sqrt{\frac{K}{M}} + \sqrt{\frac{M}{N}} + 3\sqrt{\frac{2\ln \frac{1}{\delta}}{M}}) \] (9)
for \( i = 1, \ldots, K \), where \( \lambda_1 \) is the largest eigenvalue of \( \Sigma \).

Remarks: A similar result on the bound of eigenvalues due to random projection is given in [23, Section 8.2]. However, it contains some parameters which are unclear. Furthermore, their result is not probabilistic which is the nature of random projections. Lastly, Lemma 8.4 in [23] only provides the upper bound, whilst our result provides both upper and lower bound using the theory of invariant subspaces.

The above result suggests that the power of the principal subspace is approximately preserved in the compressed domain. Consequently, the total power of the residual subspace is also approximately preserved, though the actual residual subspace itself is not preserved under the CS projection. The small variation in the principal subspace thus subsequently translates to a small change in the performance of anomaly detection using compressed data as compared to original uncompressed data. The following result quantifies this small change in terms of one important aspect of anomaly detection, the false alarm (FA) rate.

Theorem 2: In an identical condition described above, the change in false alarm rate is very small and is bounded by (with probability at least \( 1 - \delta \), where \( \delta > 0 \))
\[ \Delta FA \leq O\left(\sqrt{\frac{M}{N}} + \frac{2\ln \frac{1}{\delta}}{M}\right) \] (10)

Remarks: We note particularly that in general a random matrix used as a projection does not preserve the principal subspace as the bounds on the deviation of eigenvalues can be too large to be useful. However, in a case of CS where the dimension of the principal subspace is small (i.e. sparse signals) and CS parameters are chosen properly, then the principal subspace is preserved. Furthermore, it is interesting note that in such cases any CS projection can approximately preserve the principal subspace.

V. Experimental Results

The performance of our approach was evaluated using two experiments involving real and synthetic datasets.

A. Abilene Data

The aim of the first experiment was to determine the volume anomaly detection of the method using a real world dataset. The Abilene dataset\(^4\) consists of the readings collected from 41 network links over

\(^4\)www.abilene.iu.edu
TABLE I

ANOMALY DETECTION PERFORMANCE ON SYNTHETIC DATA.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Time (secs)</th>
<th>AUC</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>PCA</td>
<td>CS-PCA</td>
</tr>
<tr>
<td>100</td>
<td>0.023</td>
<td>0.993</td>
<td>0.997</td>
</tr>
<tr>
<td>500</td>
<td>0.430</td>
<td>0.996</td>
<td>0.991</td>
</tr>
<tr>
<td>1000</td>
<td>3.364</td>
<td>0.982</td>
<td>0.984</td>
</tr>
<tr>
<td>2000</td>
<td>20.932</td>
<td>0.986</td>
<td>0.979</td>
</tr>
</tbody>
</table>

a period of several months. The OD trace contains the measurement from each link for every 10-second interval. We use a subset of the data which covers a period of 2 weeks (1008 measurements per week). The majority of the data reflects normal network conditions with only 6 real anomalies (verified manually) in the original dataset. In addition, we injected 45 synthetic anomalies of different magnitude following the procedure described in [16]. Throughout this section, we will denote PCA computed on uncompressed data by PCA (Original) only and PCA computed on compressed data by CS-PCA (Compressed).

The trace from the first week was used as the training set while the trace from the second week was used as test data. For the Abilene network data we have $N = 41$, $M = 16$, $K = 6$, and the Welch bound is 0.1976. Using Elad’s algorithm [8], we achieve a mutual coherence of 0.36 from the initial coherence of 0.55. The threshold limit $Q_\beta$ was computed according to the $1 - \beta$ confidence level and $\beta$ was varied between 0 and 1. As $\beta$ controls the desired false positive rate (FPR), we recorded the FPRs for each of the $\beta$ values.

The results from the Abilene dataset experiment are summarized in Figure 4. The figure shows the receiver operating characteristics (ROC) plots of the PCA method applied on the uncompressed and compressed data. As it can be observed from the plot, the performance on the compressed data is very close to that on the uncompressed data. To further quantify this, we also compared the two ROC curves using (i) the area under the ROC curve (AUC) and (ii) equal error rate (EER) where the false positive being equal to false negative. An effective classifier should achieve an AUC close to 1 and ERR small. From the ROC curves, we determined that the AUC/EER values were 0.95/0.09 and 0.94/0.11 for the compressed and uncompressed data respectively. These figures support our claim that the proposed approach performs equally to PCA on the original domain.
Fig. 4. ROC curve for the Abilene network data

B. Synthetic Data

The aim of the second set of experiments was twofold: (i) to demonstrate the scalability of our approach with large $N$ and (ii) to show that it has a better time resolution property when processing the data stream in CS domain.

The synthetic data sets are generated following the equations (5) and (6). We conduct the experiment for different number of links $N$ with values 100, 500, 1000 and 2000. The number of readings is $L = 2000$. To optimize the CS projection as mentioned previously, we have selected DCT as a basis for $\Phi$ and the sparsity $K$ of the principal signal $s$ is 4. We added zero-mean Gaussian noise ($n$) with $\sigma = 0.01$. To simulate abnormal network conditions we injected 70 anomalies of different magnitude in the signal following the procedure mentioned in [16].

We selected the number of CS measurements according to $M \sim K \log N$ so that for the number of links being considered, the values of $M$ were 38, 118, 180, and 290 respectively. The sensing matrices ($\Phi$) were random Gaussian with a mutual coherence of 0.37, 0.37, 0.33 and 0.26 respectively.

**Scalability performance**: The results from the scalability analysis are shown Figures 5(b), 5(a) and Table I. Specifically, Figure 5(a) shows a plot of the eigenvalue distribution and Figure 5(b) shows a plot of the residual subspace for both PCA and CS-PCA. The results indicate the preservation of the residual subspace in the CS domain for high dimensional data. This a very encouraging result from the point of
Fig. 5. Synthetic data

(a) Eigenvalue plot for original and compressed data

(b) Observations in the residual subspace

Fig. 6. Trade-off between computational time and error rate

(a) Computational time

(b) Equal error rate

view of detecting anomaly in CS domain.

Table I compares the performance of PCA and CS-PCA in terms of actual computational time, AUC, and EER. As can be seen, our proposed method reduces the computational time by a factor ranging from 6 to as much as 200 when $N$ varies from 100 to 2000. Furthermore, the reduction in performance in terms of AUC and EER is small and for high dimensional data ($N = 1000$ and 2000), the PCA method in CS domain performs better than the PCA in original domain. The better performance is a result of better orthogonality of random vectors in high dimensional space.

We also study the trade-off between compression and performance. Loosely speaking, a smaller value of the CS dimension $M$ reduces the computational complexity at the cost of potentially lower performance.
due to the increase in the mutual coherence of the sensing matrix. In the CS literature, the value of $O(K \log N)$ has been frequently suggested. We show that this value is also about the same for the network anomaly detection problem. We consider $N = 2000$ and vary the value of $M$ between 100 to 1000 and measure the EER and computational time. The results are shown in Figs. 6(a) and 6(b). As can be seen, selecting $M$ in the range $250 - 300$ gives moderately low error rate at a large reduction in computational time. If $M$ is too low, the error rate becomes much larger. If $M$ is too large, the reduction in error rate is not very significant whilst the computational time increases somewhat quadratically.

Temporal accuracy: We tested the time resolution property of PCA method on the synthetic data of dimension ($N$) equal to 500 and length ($L$) equal to 2000. The stream was processed by a different length of moving window ($WL = 100, 200, 300, 500$) in both uncompressed and compressed domain. Figure 7(a) shows the plot between EER and WL obtained from PCA and CS-PCA. The plot shows that, CS-PCA performs better than the PCA method with a smaller window length. For example, when $WL = 100$, the EER for CS-PCA (i.e, 0.08) is better than the EER of PCA (i.e. 0.2) and the better performance of CS-PCA continues as long as the window length (WL) is less than the dimension of the original data ($N = 500$). As mentioned in earlier section IV-C, the better performance in CS comes due to better estimation of the covariance matrix in CS domain. Loosely speaking, if the number of training sample is less than the dimension of the sample, then the estimation of the covariance matrix is poor, results in poor performance in original domain. The finite sample effect is overcome in CS domain, where dimension of the samples is less than the window length and results in better covariance matrix estimation.
estimation and better performance than original domain. Hence, we can achieve a fine grain temporal resolution property of PCA technique in compressed domain instead of uncompressed domain. To further demonstrate this, we have also provided the ROC curves in Figure 7(b) by choosing different value of WL for PCA and CS-PCA.

VI. Conclusion

A new framework of network anomaly detection in the compressed domain has been presented. We have demonstrated that when the data is compressible, the CS approach allows the analysis and detection to be equivalently performed over the compressed version of the data. This is extremely important in making algorithms scalable to high-dimensional problems. We note that our framework presented in this paper can be extended to a distributed setting similar to [13] to further reduce the communication overhead between the nodes and the central unit.

In this work, we have proved that the principal subspace of PCA method is preserved via the CS projection whilst the residual subspace retains its noise-like behavior. This forms the basis for our proposed algorithm which has been validated using both real and synthetic data. The results indicate that the proposed approach is highly scalable without any significant degradation in performance.

Appendix 1

Restricted Isometric Property [4]: An $M \times N$ matrix $\Phi$ is said to have the $K$-RIP property with the constant $\delta_K$ if the following inequalities hold for all $K$-sparse vector $x$

$$(1 - \delta_K)\|x\|^2 \leq \|\Phi x\|^2 \leq (1 + \delta_K)\|x\|^2$$

APPENDIX 2

Proof of Theorem 1

Let $\Lambda_K = \text{diag}[\lambda_1 \ldots \lambda_K]$ where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_K$. Using the assumptions above, we have

$$\Sigma_x \approx [\Psi_s \Psi_n] \begin{bmatrix} \Lambda_K & 0 \\ 0 & \sigma^2 I_{N-K} \end{bmatrix} \begin{bmatrix} \Psi_s^T \\ \Psi_n^T \end{bmatrix},$$

$$\Sigma_y = \Phi \Sigma_x \Phi^T.$$  \hspace{1cm} (12)

We shall show that $\Sigma_y$ also has a matching principal subspace in a sense that the $K$ principal eigenvalues of $\Sigma_y$, which we denote by $\xi_1, \ldots, \xi_K$ are close to the $K$ principal eigenvalues of $\Sigma_x$, while the rest is small. To do so, we first define the residual matrix

$$E = \Sigma_x \Phi^T - \Phi^T \Sigma_y.$$  \hspace{1cm} (13)
Applying Theorem 8.1.11 in [11] on approximate invariant subspaces we have
\[ |\lambda_i - \xi_i| \leq \sqrt{2}(\|E\|_2/\sigma_M + \|I_M - \Phi\Phi^T\|_2\|\Sigma_x\|_2), \quad i = 1, \ldots, M, \] (15)
where \( \|\cdot\|_2 \) denotes the spectral norm and \( \sigma_M \) is the \( M \)th singular value of \( \Phi \). We need to show the two terms are small for the CS case. We note the following results on Gaussian random matrices for large \( N \) (see [10], [3]) that
\[ 1 - \sqrt{M/N} \leq \sigma_{\min}(\Phi) \leq \sigma_{\max}(\Phi) \leq 1 + \sqrt{M/N} \] (16)
holds with an overwhelming probability, where \( \sigma_{\min} \) and \( \sigma_{\max} \) denote the minimum and maximum singular values.

In our case, sensing matrix \( \Phi \) has dimension \( M \) by \( N \) (\( M \ll N \)) and the entries are i.i.d Gaussian with mean zero and standard deviation \( \frac{1}{\sqrt{N}} \). We can get the deviation bound on the eigenvalues of the matrix \( \Phi \) following equations (3.19) and (3.20) in [4] and for \( t > 0 \), we have
\[ \Pr(\sigma_{\max}(\Phi) > 1 + \sqrt{M/N} + t) \leq e^{-N\frac{t^2}{2}} \] (17)
\[ \Pr(\sigma_{\min}(\Phi)) < 1 - \sqrt{M/N} - t) \leq e^{-N\frac{t^2}{2}} \] (18)
The above equations can be modified as follows:
\[ \Pr(\sigma_{\max}(\Phi)) < 1 + \sqrt{M/N} + t) \geq 1 - e^{-N\frac{t^2}{2}} \] (19)
\[ \Pr(\sigma_{\min}(\Phi)) > 1 - \sqrt{M/N} - t) \geq 1 - e^{-N\frac{t^2}{2}} \] (20)
Let \( \delta = e^{-N\frac{t^2}{2}} \) or \( t = \sqrt{2\ln(1/\delta)/N} \), then we can say that the following equations hold with high probability \( 1 - \delta \) (where \( \delta > 0 \)),
\[ \sigma_{\max}(\Phi) \leq 1 + \sqrt{M/N} + \sqrt{2\ln(1/\delta)/N} \] (21)
\[ \sigma_{\min}(\Phi) \geq 1 - \sqrt{M/N} - \sqrt{2\ln(1/\delta)/N} \] (22)
Similarly,
\[ \sigma_{\max}(\Phi\Phi^T) \leq 1 + 2(\sqrt{M/N} + \sqrt{2\ln(1/\delta)/N}) \] (23)
\[ \sigma_{\min}(\Phi\Phi^T) \geq 1 - 2(\sqrt{M/N} + \sqrt{2\ln(1/\delta)/N}) \] (24)
hold with high probability at least \( 1 - \delta \). Hence,
\[ \|I_M - \Phi\Phi^T\|_2\|\Sigma_x\|_2 \leq 2\lambda_1(\sqrt{M/N} + \sqrt{2\ln(1/\delta)/N}) \] (25)
holds with probability \(1 - \delta\).

Next, we bound \(\|E\|_2\). To simplify the maths, we note from the approximate invariant subspace theory that \(\|E\|_2\) is dependent on whether the subspace induced by the random projection covers the principal subspace, i.e. their relative geometrical relationship. This relationship is in fact independent of the coordinates we choose. Suppose that there exists a rotational transformation \(f : \Psi \rightarrow I\) and under this rotational transformation \(\Phi \rightarrow \Phi'\). Another key observation is that \(\Phi\) is a Gaussian random matrix. As the randomness properties of \(\Phi\) are unchanged under such rotational transformation, it follows that \(\Phi'\) is also a Gaussian random matrix (see Lemma 2). Because randomness properties are only used in subsequent derivations and not the actual values of \(\Phi\), it is equivalent to set \(\Psi = I\) when bounding \(\|E\|_2\). We write \(\Phi = [\Phi_1 \ \Phi_2]\) where \(\Phi_1 \in \mathbb{R}^{M \times K}\) and \(\Phi_2 \in \mathbb{R}^{M \times (N-K)}\). After some straightforward manipulations and using \(\lambda \gg \sigma^2\) we have \(\|E\|_2 \approx \|F\|_2\) where

\[
F = \begin{bmatrix}
\Phi_1^T(I_M - \Phi_1 \Phi_1^T)\Lambda_K \\
-\Phi_1^T \Phi_1 \Lambda_K \Phi_1^T
\end{bmatrix}.
\] (26)

Invoking inequalities on matrix norm (see [11]), we have

\[
\|F\|_2 \leq \|\Phi_1^T(I_M - \Phi_1 \Phi_1^T)\|_2\|\Lambda_K\|_2 + \|\Phi_2^T \Phi_1 \Phi_1^T\|_2\|\Lambda_K\|_2
\] (27)

\[
\|F\|_2 \leq \lambda_1(\|\Phi_1^T(I_M - \Phi_1 \Phi_1^T)\|_2 + \|\Phi_2^T \Phi_1 \Phi_1^T\|_2).
\] (28)

We now bound the eigenvalues of \(\Phi_1\). However, to use the results (17) and (18), we need to scale \(\Phi_1\) by introducing \(\Phi'_1 = \frac{\sqrt{N}}{M} \Phi_1^T\). It is easy to verify that

\[
\sigma(\Phi_1) = \sigma(\Phi_1^T) = \sigma(\sqrt{M/N} \Phi'_1) = \sqrt{M/N} \sigma(\Phi'_1)
\] (29)

Similarly, following (17) and for \(t_1 > 0\)

\[
\Pr(\sigma_{\text{max}}(\Phi_1) < 1 + \sqrt{K/M} + t_1) \geq 1 - e^{-M\frac{t_1^2}{2}}
\] (30)

if \(\delta_1 = e^{-M\frac{t_1^2}{2}}\), then \(t_1 = \sqrt{2\ln(1/\delta_1)/M}\). Now, following (29) and (30), we have,

\[
\sigma_{\text{max}}(\Phi_1) \leq \sqrt{M/N}(1 + \sqrt{K/M} + \sqrt{2\ln(1/\delta_1)/M}))
\] (31)

holds with high probability \(1 - \delta_1\), where \(\delta_1 > 0\). Similarly,

\[
\sigma_{\text{min}}(\Phi_1 \Phi_1^T) \leq (M/N)(1 - 2(\sqrt{K/M} + \sqrt{2\ln(1/\delta_1)/M}))
\] (32)

holds with probability at least \(1 - \delta_1\). Next, we will get the bound for first right hand term of (28) as follows:

\[
\|\Phi_1^T(I_M - \Phi_1 \Phi_1^T)\|_2 \leq \|\Phi_1^T\|_2\|(I_M - \Phi_1 \Phi_1^T)\|_2
\] (33)
Now combining (31), (32) and (33), we can write
\[
\| \Phi_1^T (I_M - \Phi_1 \Phi_1^T) \|_2 \leq \sqrt{M/N} (1 + \sqrt{K/M} + \sqrt{2 \ln (1/\delta_1)/M})
\] (34)
holds with high probability $1 - \delta_1$. Next we compute bound for the second term of (28) as follows:
\[
\| \Phi_2^T \Phi_1 \Phi_1^T \|_2 \leq \| \Phi_2^T \|_2 \| \Phi_1 \Phi_1^T \|_2 \leq \| \Phi_2 \|_2 \| \Phi_1 \Phi_1^T \|_2
\] (35)

Now, from (21), (31) and (35), we can have
\[
\| \Phi_2^T \Phi_1 \Phi_1^T \|_2 \leq (1 + \sqrt{M/N} + \sqrt{2 \ln (1/\delta_1)/N}) ((M/N)
\[
\left[ 1 + 2(\sqrt{K/M} + \sqrt{2 \ln (1/\delta_1)/M}) \right].
\] (36)

After some straightforward simplification, we get
\[
\| \Phi_2^T \Phi_1 \Phi_1^T \|_2 \leq (M/N) (1 + \sqrt{2 \ln \frac{1}{\delta_1} \left( \frac{1}{\sqrt{N}} + \frac{2}{\sqrt{M}} \right)})
\] (38)
holds with high probability $1 - \delta_1$. Now following (28), (34), (38)
\[
\| F \|_2 \leq \lambda_1 \left( \sqrt{M/N} (1 + \sqrt{K/M} + \sqrt{2 \ln (1/\delta_1)/M}) +
\] (M/N) \left( 1 + \sqrt{2 \ln \frac{1}{\delta_1} \left( \frac{1}{\sqrt{N}} + \frac{2}{\sqrt{M}} \right)} \right)
\] (39)
\[
(40)
\]
Since $M \ll N$, we have
\[
\| F \|_2 \leq \lambda_1 \left[ \sqrt{M/N} (1 + \sqrt{K/M} + \sqrt{2 \ln (1/\delta_1)/M}) \right].
\] (41)

Again combining (15), (25), (41), we get
\[
| \lambda_i - \xi_i | \leq \sqrt{2} \lambda_1 \left[ \sqrt{M/N} (1 + \sqrt{K/M} + \sqrt{2 \ln (1/\delta_1)/M}) +
\] 2(\sqrt{M/N} + \sqrt{2 \ln (1/\delta_1)/N})
\] (42)
\[
(43)
\]
After some straightforward simplification and letting $\delta = \delta_1$, the following inequality holds with high probability $1 - \delta$,
\[
| \lambda_i - \xi_i | \leq \sqrt{2} \lambda_1 \left[ 3 \sqrt{M/N} + \sqrt{K/M} + 3 \sqrt{2 \ln (1/\delta_1)/N} \right]
\] (44)

**APPENDIX 2**

**PROOF OF THEOREM 2 (BOUND ON FALSE ALARM RATE)**

As we mentioned in the section IV-D, The residual statistics has normal distribution and the tail of the distribution is most important in setting a suitable threshold for certain false alarm rate. Since, according to the CS assumptions, the principal subspace is approximately preserved in CS domain and
the perturbation in principal eigenvalues are very small (see theorem 1). As a result of this, there is also a small deviations in false alarm rate. In figure 8, the continuous graph is showing the tail behavior in original domain and other two other graphs are perturbed tail behavior in CS domain. The deviation in the tail behavior also deviates the detection threshold and the false alarm rate respectively.

Following the section 3 in [14], the residual statistics $Q$ has normal distribution and it can be written (see A.7 in [14]) as,

$$\left(\frac{Q}{\theta_1}\right)^{h_0} \sim \mathcal{N}\left[1 + \frac{\theta_2 h_0(h_0 - 1)}{\theta_1^2}, \frac{2\theta_2 h_0^2}{\theta_1^2}\right].$$

(45)

Here, $\theta_1 = \sum_{i=K+1}^{N} \lambda_i$, $\theta_2 = \sum_{i=K+1}^{N} \lambda_i^2$, $\theta_3 = \sum_{i=K+1}^{N} \lambda_i^3$, $h_0 = 1 - \frac{2\theta_1}{\theta_2\sigma_2}$ and $K$ is the number of principal components. Let, $z \sim \mathcal{N}(\mu, \sigma)$, where $z = (\frac{Q}{\theta_1})^{h_0}$, $\mu = 1 + \frac{\theta_2 h_0(h_0 - 1)}{\theta_1^2}$ and $\sigma = \frac{2\theta_2 h_0^2}{\theta_1^2}$. So we can write $\frac{z - \mu}{\sigma} \sim \mathcal{N}(0,1)$. Now we have,

$$\frac{Z_{\beta} - \mu}{\sigma} \sim C_{\beta}$$

(46)

Here, $C_{\beta} = (1 - \beta)$ percentile of normal distribution with zero mean and unit variance, and $Z_{\beta}$ is the detection threshold limit. We denote the detection threshold in original domain by $Z_{\beta}^x$ and for CS domain, it is denoted by $Z_{\beta}^y$. Please note, any parameter with superscript/superscript $x$ refers to original data and with $y$ refers to CS data. As showing in the figure 8, $Z_{\beta}^x$ is the detection threshold in original domain.
and get perturbed by an amount of $\pm \delta z^x_\beta$ in CS domain. Hence change in false alarm would be the area under the curve between $z^x_\beta$ and $z^x_\beta \pm \delta z^x_\beta$ and which can be computed as follows:

$$\Delta \Pr(FA) = N(z^x_\beta) \delta z^x_\beta$$

(47)

where $N(z^x_\beta)$ is short for the pdf value at $z^x_\beta$ Hence, we need to compute the amount of deviation $\delta z^x_\beta$.

Now, we have from (46),

$$\frac{Z^x_\beta - \mu_x}{\sigma_x} = \frac{Z^y_\beta - \mu_y}{\sigma_y} = C_\beta.$$  

(48)

where,

$$\mu_x = 1 + \frac{\theta^x_2 (h^x_0 (h^x_0 - 1))}{(\theta^x_1)^2}, \quad \mu_y = 1 + \frac{\theta^y_2 (h^y_0 (h^y_0 - 1))}{(\theta^y_1)^2}, \quad \sigma_x = \frac{2\theta^x_2 (h^x_0)^2}{(\theta^x_1)^2}, \quad \sigma_y = \frac{2\theta^y_2 (h^y_0)^2}{(\theta^y_1)^2}$$

(49)

So,

$$Z^x_\beta = \mu_x + C_\beta \sigma_x.$$  

(50)

and

$$Z^y_\beta = \mu_y + C_\beta \sigma_y.$$  

(51)

From (50) and (51), we can write

$$Z^x_\beta - Z^y_\beta = (\mu_x - \mu_y) + C_\beta (\sigma_x - \sigma_y).$$

(52)

and therefore, following the triangle inequality, we get

$$|Z^x_\beta - Z^y_\beta| \leq |(\mu_x - \mu_y)| + C_\beta |(\sigma_x - \sigma_y)|.$$  

(53)

Denote

$$\delta_1 = \theta^x_1 - \theta^y_1,$$

$$\delta_2 = \theta^x_2 - \theta^y_2,$$

$$\delta_3 = \theta^x_3 - \theta^y_3,$$

$$\delta_4 = h^x_0 - h^y_0.$$  

(54)

(55)

(56)

(57)

Following [14], we can write (in case of original domain)

$$\theta^x_i = \text{tr}(\Sigma_x)^i - \sum_{j=1}^{K} (\lambda^x_j)^i \quad i = 1, 2, 3.$$  

(58)

Similarly we can write (in case of compressed domain),

$$\theta^y_i = \text{tr}(\Sigma_y)^i - \sum_{j=1}^{K} (\lambda^y_j)^i \quad i = 1, 2, 3.$$  

(59)
Now, from (58) and (59), we get
\[ \delta_1 = \theta_1^x - \theta_1^y = (\text{tr}(\Sigma_x) - \text{tr}(\Sigma_y)) + \sum_{j=1}^{K}((\lambda_j^x) - (\xi_j^y)). \] (60)

We now attempt to bound each RHS term of (59). First, for the trace terms we recall that
\[ \text{tr}(\Sigma_x) - \text{tr}(\Sigma_y) = \text{tr}(\Sigma_x) - \text{tr}(\Phi \Sigma_x \Phi^T) \] (61)

From the previous remark due to Lemma 2, we can assume \( \Sigma_x \) being an diagonal matrix, \( \Sigma_x = \text{diag}(\lambda_1, \ldots, \lambda_N) \). Let \( \phi_i \) be the \( i^{th} \) column of the matrix \( \Phi^T \), then we have
\[ \text{tr}(\Phi \Sigma_x \Phi^T) = \text{tr}(\sum_{i=1}^{M} \phi_i^T \Sigma_x \phi_i) = \text{tr}(\sum_{i=1}^{M} (\sum_{j=1}^{N} \phi_{ij}^2 \lambda_j)) = \text{tr}(\sum_{j=1}^{N} \lambda_j (\sum_{i=1}^{M} \phi_{ij}^2)) \] (62)

Since, each column the matrix \( \Phi \) is normalized to unity, then we can have,
\[ \text{tr}(\Phi \Sigma_x \Phi^T) = \text{tr}(\Sigma_x) \] (63)

Second, we bound the eigenvalues terms of (59) using Theorem 1, which implies that
\[ \lambda_j^x - \xi_j^y = \sqrt{2}\lambda_1^x (3\sqrt{\frac{M}{N}} + \sqrt{\frac{K}{N}} + 3\sqrt{\frac{2\ln \frac{1}{\delta}}{N}}) \] (64)

So, combining (60), (63) and (64), we get
\[ \delta_1 = \sum_{j=1}^{K}((\lambda_j^x) - (\xi_j^y)) = \sqrt{2}\lambda_1^x K(3\sqrt{\frac{M}{N}} + \sqrt{\frac{K}{N}} + 3\sqrt{\frac{2\ln \frac{1}{\delta}}{N}}) \] (65)

Since, \( K < M \ll N \), then we have \( M/N \gg K/N \), and
\[ \delta_1 \sim 3\sqrt{2}\lambda_1^x \mathcal{O}(K \left[ \sqrt{\frac{M}{N}} + \sqrt{\frac{2\ln \frac{1}{\delta}}{N}} \right]) \] (66)

Similarly for \( n = 2, 3 \) we can prove,
\[ \delta_2 \sim 6\sqrt{2}(\lambda_1^x)^2 \mathcal{O}(K \left[ \sqrt{\frac{M}{N}} + \sqrt{\frac{2\ln \frac{1}{\delta}}{N}} \right]) = 2\lambda_1^x \delta_1 \] (67)

and
\[ \delta_3 \sim 9\sqrt{2}(\lambda_1^x)^3 \mathcal{O}(K \left[ \sqrt{\frac{M}{N}} + \sqrt{\frac{2\ln \frac{1}{\delta}}{N}} \right]) = 3(\lambda_1^x)^2 \delta_1 \] (68)

Next, we bound each term of (53). From (49), we have
\[ \mu_x - \mu_y = \frac{\theta_2^x h_0^x (h_0^x - 1)}{(\theta_1^x)^2} - \frac{\theta_2^y h_0^y (h_0^y - 1)}{(\theta_1^y)^2} \] (69)
Denote \( f(x) = \frac{\theta_2 h_2(h_2-1)}{(\theta_1)^2} \) and \( f(y) = \frac{\theta_2 h_2(h_2-1)}{(\theta_1)^2} \), where, \( \text{x} = \frac{\theta_1^2}{\theta_2}, \text{y} = \frac{\theta_1^2}{\theta_3} \). From (69), we can write,

\[
|\mu_x - \mu_y| = |f(x) - f(y)| \leq \left| \frac{\partial f}{\partial \theta_1^2} \right| \Delta \theta_1^2 + \left| \frac{\partial f}{\partial \theta_2^2} \right| \Delta \theta_2^2 + \left| \frac{\partial f}{\partial \theta_3} \right| \Delta h_0^2
\]

Now, \( \Delta \theta_1^2 = \delta_1, \Delta \theta_2^2 = \delta_2 \) and \( \Delta \theta_3^2 = \delta_3 \). After straightforward manipulation we have,

\[
f(x) = \frac{\theta_2^2 \left( 1 - \frac{2 \theta_2^2}{3 \theta_3^2} \right) \left( -\frac{2 \theta_3^2}{3 (\theta_3^2)^2} \right)}{(\theta_1^2)^2}
\]

which can be simplified to,

\[
f(x) = -\frac{2}{3} \left[ \frac{\theta_2^2}{\theta_1^2 \theta_2^2} - \frac{2 \theta_2^2}{3 (\theta_3^2)^2} \right]
\]

\[
\frac{\partial f}{\partial \theta_1^2} = \frac{2}{3} \frac{\theta_2^2}{(\theta_2^2)^2 \theta_3}
\]

\[
\frac{\partial f}{\partial \theta_2^2} = -\frac{2}{3} \left[ \frac{\theta_3^2}{\theta_1^2 (\theta_3^2)^2} + \frac{2 \theta_2^2}{(\theta_2^4)^2} \right]
\]

and,

\[
\frac{\partial f}{\partial \theta_3^2} = -\frac{2}{3} \left[ \frac{1}{\theta_1^2 \theta_2^2} - \frac{4 \theta_3^2}{3 (\theta_3^2)^2} \right]
\]

So, from (70), we can write

\[
|\mu_x - \mu_y| \leq C_1 \delta_1
\]

where

\[
C_1 = \left[ \frac{2}{3} \frac{\theta_3^2}{(\theta_1)^2 \theta_2^2} - \frac{2}{3} \left( \frac{\theta_3^2}{\theta_3^2 (\theta_2)^2} \right) \right] \left( \frac{2 \theta_2^2}{3 (\theta_3^2)^2} \right) \left( \frac{1}{\theta_1^2 \theta_2^2} - \frac{4 \theta_3^2}{3 (\theta_3^2)^2} \right)
\]

In similar way, we can prove that,

\[
\sigma_x - \sigma_y \leq C_2 \delta_1
\]

where

\[
C_2 = \left[ -\frac{2}{3} \frac{\theta_2^2}{(\theta_1)^2} + \frac{2}{3} \left( \frac{1}{\theta_1^2} \right) \frac{2 \theta_3^2}{3 (\theta_2)^2} \right]
\]

Then combining (47), (53), (76) and (78), we have

\[
\Delta Pr(FA) \leq \lambda_1 O(K \left[ \sqrt{\frac{M}{N}} + \sqrt{\frac{2 \ln 1}{N}} \right])
\]

holds with probability at least \((1 - \delta)\).
APPENDIX 3

PRESERVATION OF GAUSSIANITY UNDER UNITARY TRANSFORMATION

The following lemma could have appeared in a standard statistical text. For completeness, the result and its proof is given to support the claim in the main theorem.

**Lemma 2:** Suppose that \( \Phi \in \mathbb{R}^{M \times N} \) is an iid random matrix whose entries follow a zero-mean Gaussian distribution with variance \( \sigma^2 \). Let \( U \in \mathbb{R}^{N \times N} \) be a unitary matrix. Then \( \Phi' = \Phi U \) is also an iid Gaussian random matrix with the same variance \( \sigma^2 \).

**Proof:** First we prove that \( \mathbb{E}[\phi'_{ij}] = \mathbb{E}[\phi_{ij}] = 0 \) and \( \text{Var}[\phi'_{ij}] = \sigma^2 \). We start from

\[
\phi'_{ij} = \sum_{k=1}^{n} \phi_{ik} u_{kj}.
\]

Thus

\[
\mathbb{E}[\phi'_{ij}] = \sum_{k=1}^{N} \mathbb{E}[\phi_{ik}] u_{kj} = 0,
\]

whilst due to iid assumption

\[
\text{Var}[\phi'_{ij}] = \sum_{k=1}^{N} \text{Var}[\phi_{ik}] u_{kj} = \sum_{k=1}^{N} \sigma^2 u_{kj}^2 = \sigma^2 \sum_{k=1}^{n} u_{kj}^2 = \sigma^2.
\]

Next, we prove the iid in a similar way, i.e.

\[
\mathbb{E}[\phi'_{ij} \phi_{mn}] = \mathbb{E}\left[ \sum_{k=1}^{N} \phi_{ik} u_{kj} \sum_{k'=1}^{N} \phi_{mk'} u_{k'n} \right]
\]

\[
= \sum_{k=1}^{N} \sum_{k'=1}^{N} \mathbb{E}[\phi_{mk'} \phi_{ik}] u_{k'n} u_{kj}
\]

\[
= 0.
\]

\[\blacksquare\]

REFERENCES


