Preprint manuscript No.
(will be inserted by the editor)

Anomaly Detection in Large-Scale Data Stream Networks

Duc-Son Pham[†] · **Svetha Venkatesh**[‡] · **Mihai Lazarescu**[†] · **Saha Budhaditya**[‡]

Received: date / Accepted: date

Abstract This paper addresses the anomaly detection problem in large-scale data mining applications using residual subspace analysis. We are specifically concerned with situations where the full data cannot be practically obtained due to physical limitations such as low bandwidth, limited memory, storage, or computing power. Motivated by the recent compressed sensing (CS) theory, we suggest a framework wherein random projection can be used to obtained compressed data, addressing the scalability challenge. Our theoretical contribution shows that the spectral property of the CS data is approximately preserved under a such a projection and thus the performance of spectral-based methods for anomaly detection is almost equivalent to the case in which the raw data is completely available. Our second contribution is the construction of the framework to use this result and detect anomalies in the compressed data directly, thus circumventing the problems of data acquisition in large sensor networks. We have conducted extensive experiments to detect anomalies in network and surveillance applications on large datasets, including the benchmark PETS 2007 and 83GB of real footage from 3 public train stations. Our results show that our proposed method is scalable, and importantly, its performance is comparable to conventional methods for anomaly detection when the complete data is available.

Keywords anomaly detection · random projection · sensor network data · spectral methods · compressed sensing · residual subspace analysis · stream data processing

1 Introduction

The problem of detecting anomalies in data streams captured by large-scale sensor networks has received much interest [5,13,23,36,40] over the past decade. As large-scale networks become prevalent, there is an increasing need to develop approaches that can address the challenges arising from large amounts of data. The problem affects a wide range of applications as the data captured by sensor networks constitutes multimedia content from the web, video from surveillance camera networks, satellite imagery or typical network traffic.

[†] Department of Computing, Curtin University, Perth, Western Australia
E-mail: dspham@ieee.org
[‡] Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Geelong, Victoria, Australia
E-mail: svetha.venkatesh@deakin.edu.au

Approaches to anomaly detection vary significantly in the scope of the detection, the underlying statistical methods, as well as the assumption about the data. As there are varying differences between the definition of anomalies in different settings, it is usually difficult to directly compare the methods in the literature. Examples include Bayesian methods [27, 28], SVM [19], example-based [43] and spectral methods [6], mixed-type data [29]. A more complete survey of anomaly detection is documented in [12].

We restrict our attention to spectral methods for anomaly detection, in particular to residual subspace analysis. This method was originally developed for control system theory [24], [25], [26]. It decomposes data into the principal subspace that characterizes the normal behavior of data, and the residual subspace where anomalies are to be found. Under the null hypothesis that the data is *normal*, the squared prediction error (SPE) which is the l_2-norm of the residual vector, follows a non-central chi-square distribution. Hence, the rejection of the null hypothesis can be based on whether the norm of the residual vector exceeds a certain threshold corresponding to a desired false alarm rate. The threshold is computed based on a statistical measure called Q-statistic [24], [31], [23], which can be computed as a function of residual eigenvalues. Recently, this method has found use in some network anomaly detection problems [30].

Let $\mathbf{x}_i \in \mathbb{R}^N$ be a N-dimensional vector that represents the status of a data network with N sensors at time i, and denote as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$ the network data matrix. We consider two cases:

Case 1: It is difficult to collect *all rows* of the data matrix $\mathbf{X}$ to a central monitor. This could be because of limited bandwidth between the central node and the sensors such that only M readings can be sent over the communications link, where typically $M \ll N$. In other circumstances, many sensors of the network may be physically far away from the central node and such direct communication with the central node may be impossible. Furthermore, the central node may not have storage for all sensors due to memory limitations. In this situation, it is desirable to summarize the status of the sensor network with an M-dimensional reading such that it meets both the bandwidth and memory constraints.

Case 2: It is difficult to collect *all columns* of the data matrix $\mathbf{X}$ centrally. This is equivalent to sub-sampling the temporal stream. This is useful in situations where anomalies have to be found retrospectively. For example, the video data of a network of surveillance cameras may be fully available at remote nodes, and when an incident occurs, the authorities want to access the data centrally. It is however impossible to transmit the entire temporal stream to a central location. In this case, the temporal stream must be sub-sampled.

Inspired by the recent compressed sensing (CS) theory [14, 11], we propose a new framework for the detection of anomalies in such large-scale networks. The proposed framework consists of a strategy to handle large data using compressed sensing/random projection and a traditional spectral-based anomaly detection algorithm. We formalize the application of CS to effectively acquire the data in a compressed way to conform to the physical constraints. This compressed data acquisition permits either sub-sampling of the number of sensors, or the number of frames in a temporal stream and is then used for anomaly detection. Though for detection purpose, the acquisition part is essentially random projection [41], we note that the CS theory certainly enables post anomaly detection tasks such as finding the origins of detected anomalies. As the CS theory is more general, hereinafter we refer to the compressed data using random projection as CS data.

To address the issues, we propose to obtain the compressed data and then perform anomaly detection using residual subspace analysis on it. We show how such a linear transformation can be implemented in some large-scale networks. Our theoretical contributions are:

- First, we extend the theory of random projection/CS by establishing the relation between the spectral properties of the original and compressed data. Specifically, we show that the principal subspace is approximately preserved under the random projection with high probability and we derive two-sided theoretical bounds. This demonstrates that the intrinsic structure of the data is preserved under a random projection, and yields the intuition that anomalies can be detected in the residual subspace of the compressed data.
- Second, we derive the theoretical bounds on the false alarm rate with compressed data relative to complete data. The result shows that the bound is directly related to the dimension of random projection M.

We emphasize that the strategy using compressed sensing/random projection that we analyze in this work is mostly suitable for spectral-based anomaly detection algorithms, which mainly depend on the singular values of the data. However, this scalability strategy might also be useful for other non-spectral anomaly detection algorithms. Such a study to examine the possible benefit is certainly beyond the scope of current work.

We validate our method by considering both network and surveillance anomalies. For the network data, we evaluate the proposed method on real traffic traces collected from the Abilene network [1] over four weeks and synthetic network data. Our experiment verifies that on the real dataset, the proposed method using compressed data achieves equivalent performance as with complete data at a detection rate of more than 94%. For synthetic data, our experiments show that the PCA technique performs even better in compressed domain than original domain for high dimensional data. Importantly, the proposed method requires less memory and storage and can be as much as 100 times faster than the original spectral method using raw data.

For the surveillance data, we validate our method on both a benchmark dataset [2] (64MB) and real-world data collected from multiple surveillance cameras from 3 train stations over a whole week, resulting in over 83GB of video. To the best of our knowledge, the latter is the largest dataset mentioned in the video surveillance literature. It contains anomalous events that were not artificially created and were ground-truthed in conjunction with the transport authorities. The satisfactory performance of our method over different datasets offers promise for deployment in real-world scenarios [1]

The significance of our contributions is the demonstration that spectral-based methods can be applied to CS data, and that anomaly detection can be effectively performed *without an explicit reconstruction of the input signal*. Thus, anomaly detection is equivalent to the uncompressed case, but with the advantage of working with lower number of measurements. Accordingly, the computational cost is also reduced.

In terms of novelty, the framework we present integrates anomaly detection and random projection into a deployable paradigm to overcome the problem with partial data, a reality in most real-world situations. Though it is intuitive from the approximate geometry preservation of random projection that the eigenvalues should be "similar", establishing the precise bounds on eigenvalues and on the false alarm rates here is new and significant. The closest theoretical work on the bounds of eigenvalues due to random projection is given in [41, Section 8.2]. However, Lemma 8.4 in [41] only provides the upper bound, whilst our result provides both upper and lower bounds using the theory of invariant subspaces. Furthermore, their result in [41, Section 8.2] is not probabilistic, thus ignores the essence of random projections. Some similar theoretical results to [41] are presented in [32], but the application

[1] Though we only report these amounts of data in this paper, we note that the proposed method forms a core of a more complex commercial system that has been successfully tested over thousand hours of video, equivalent to hundred of Tetrabyes. For detail see `http://www.icetana.com.au`

is concentrated on using random projection as a privacy preserving mechanism rather than anomaly detection in large-scale data. This work extends our our preliminary investigation [9] by theoretical results with detailed proofs and a more extensive experimental evaluation to validate our claims.

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

2 Related Work

There are mainly three major approaches to address scalability issues in large-scale networks that are generally applicable to many problems.

The first approach is column sampling [15], [16], which is only suitable for static database applications. In this approach, an empirical distribution over the columns of $\mathbf{X}$ is constructed and a small number of columns of $\mathbf{X}$ are selected based on sampling from the empirical distribution. Due to the nature of having the full knowledge about the empirical distribution to do selective sampling, this approach appears unsuitable for online applications.

The second approach is *decentralization* where nodes in the network actively make decisions about communication and processing so as to reduce bandwidth consumption. For example, Huang *et al.* [23] propose a decentralization method for *streaming* data in which the sensors only send information to the fusion point if the observed value falls outside the normal range, which is a typically pre-defined window. If a sensor does not send any data, the fusion point will assume a nominal value. The essence of [23] is an optimized trade-off between the pre-defined window length (which implies the amount of reduction in communication overhead) and the changes in the detection performance of the matching spectral method. However, there is still a likelihood that the communication overhead exceeds the bandwidth and the fact that the central node would need to store a data matrix of the same size as $\mathbf{X}$.

The third approach is *dimensionality reduction*, where the data is transformed to a (much) lower dimension. Within this approach, there are supervised methods that require a complex optimization problem to be solved such as [42]. A recent work [21], which addresses a slightly different problem, uses a local sensitive hashing scheme to reduce the dimensionality of wireless sensors readings. Our proposed framework falls in the realm of unsupervised linear transformation, in particular that of random projection [41] (We note that from CS theory, some deterministic linear transformation might work equivalently but such a discussion is beyond the scope of this paper). This dimensionality reduction technique exploits a special statistical property of high-dimensional data wherein the geometry is approximately preserved under such transformation. Whilst the applications of random projection, more generally CS, have been found in a wide range of domains, such as preserving privacy [32] (see [41] for a more comprehensive list), to the best of our knowledge there is no work in its application to anomaly detection, especially with residual subspace analysis.

3 Background

3.1 Residual Subspace Projection and Anomaly Detection

Let the information about a network be represented by a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$ where each data instance $\mathbf{x}_i \in \mathbb{R}^N$. For notational simplicity, we assume the data matrix has been centralized. If $\mathbf{X}$ is available, the residual method performs the eigenvalue decomposition of the sample covariance matrix as:

$$\mathbf{\Sigma_x} = (1/L)\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{1}$$

from which the K principal eigenvectors $\mathbf{U}$ corresponding to the largest K eigenvalues can be found. The projection of any data instance $\mathbf{x}$ onto the residual subspace is given as:

$$\mathbf{z} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{x}. \tag{2}$$

In residual subspace analysis [24], the error signal ($\mathbf{z}$) is assumed to be multivariate normally distributed and hence the squared prediction error (SPE) $\|\mathbf{z}\|_2^2$ follows a non-central chi-square distribution under the null hypothesis that the data is normal. Hence, rejection of the null hypothesis can be based on whether norm of the error vector exceeds a certain threshold corresponding to a desired false alarm rate. The threshold is called Q-statistic, and it is a function of non-principal eigenvalues in residual subspace. For a significance level β, the Q-statistic is usually computed as:

$$Q_\beta = \theta_1 \left[\frac{c_\beta \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{\frac{1}{h_0}}, \tag{3}$$

where $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $\theta_i = \sum_{j=K+1}^{N} \lambda_j^i$ for $i = 1, 2, 3$, c_β = $(1 - \beta)$ percentile in a standard normal distribution, and $\lambda_j, i = 1, \ldots, M$ are the eigenvalues of $\mathbf{\Sigma_x}$. An anomaly is detected when $\|\mathbf{z}\|_2^2 > Q_\beta$. (see Fig. 1 for an illustration).

In practice, it is important to select a suitable value for K. Like most other spectral methods, the general principle for selecting K is the smallest number of principal components that capture most of the energy. For residual subspace methods, the selection of K is always a trade-off. Selecting small K makes the residual subspace large, and hence can improve detection but may increase false positives. On the other hand, selecting large K makes the residual subspace small, and hence reduces false positives but may increase false negatives. In our work, we select K to capture about 90% of energy in the principal subspace.

3.2 From Random Projection To Compressed Sensing

It has been observed in the literature that though the dimension of the data may be large, the intrinsic dimension which carries most information about the data is typically much smaller. This has motivated a large number of works on dimensionality reduction. They all aim at yielding compressed data that is easier to work with. In most cases, this contains less noise than that with original data. A particular common class of dimensionality reduction is linear transformation, wherein the compressed data $\mathbf{y}$ is obtained through a linear transformation

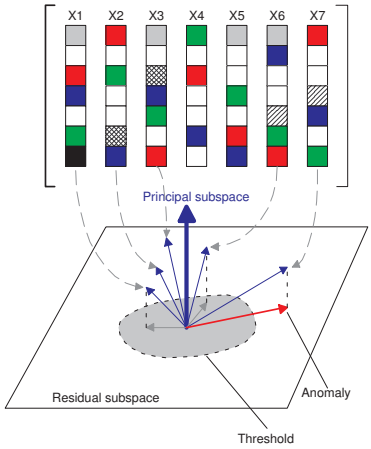


Fig. 1 Anomaly detection with residual subspace analysis. Here the data samples $\mathbf{x}_1$ to $\mathbf{x}_6$ mostly align with the principal subspace as their projection to the residual subspace, which is illustrated by a plane, is small than a threshold. However, the projection of $\mathbf{x}_7$ onto the residual subspace is large, implying an anomaly.

$\mathbf{y} = \mathbf{\Phi x}$. Here $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ effectively reduces the dimension of the data from N to M and its columns are normalized to unit norm. If the intrinsic dimension is K then obviously $M \geq K$. In compressed sensing theory, suppose that $\mathbf{x}$ is a K sparse vector, then a a linear transformation $\mathbf{\Phi}$ is characterized by a so-called restricted isometry constant (RIC) δ_K, which satisfies

$$(1 - \delta_K)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi x}\|_2^2 \leq (1 + \delta_K)\|\mathbf{x}\|_2^2. \tag{4}$$

These inequalities describe the approximate geometry preservation property of $\mathbf{\Phi}$. Ideally, a good linear transformation corresponds to small δ_K. To achieve this, the columns of $\mathbf{\Phi}$ need to be as close to orthogonal as possible. Under CS theory, $\mathbf{\Phi}$ does not have to be a random matrix and in fact there are published works that construct $\mathbf{\Phi}$ deterministically. However, it is found that many classes of random matrices often have small RIC and can be easily generated, such as Bernoulli random matrices, database friendly random matrices, and Gaussian random matrices [11,3,41]. It also follows from CS theory that the dimensions of random projection is $M = \mathcal{O}(K \log N) \ll N$ for large N. This implies that by using random projection, the compressed data could have a smaller dimension than the original data without losing its geometrical property. Because of the non-adaptive nature, such compression is suitable for large-scale problems (For further detail of these random matrices, see [11,3, 41]).

Before proceeding, we note that the above geometry preserving projection holds for data $\mathbf{x}$ that is sparse via some unitary transform, i.e. $\mathbf{x} = \mathbf{\Psi}\alpha$ and that α is K-sparse. In this case $\mathbf{\Phi\Psi}$ is effectively the linear projection. Because $\mathbf{\Phi}$ is random and $\mathbf{\Psi}$ is unitary, the statistical property remains unchanged (This can be proved easily). Secondly, though we only focus on anomaly detection in this work, the CS theory states that the original data $\mathbf{x}$ can be *reconstructed* from the compressed data $\mathbf{y}$ by solving the following optimization problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{5}$$

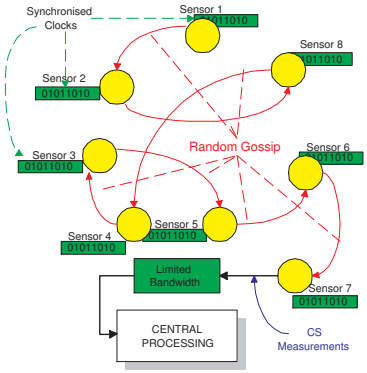


Fig. 2 Sensor subsampling. Here the synchronized sensors effectively perform cross-network compressed measurements via gossips, indicated by the red arrows. After a number of gossips, the final compressed measurements will arrive at the center node.

Here λ is a regularization parameter, which controls the sparsity of the solution, and is typically found by cross-validation techniques, the detail of which can be found more in the CS literature. From an anomaly detection point of view, this implies that post-processing tasks of anomaly detection, such as identification of anomalies might be possible under CS theory. However, we leave this for future work. In what follows, we focus on obtaining an equivalent random projection in large-scale networks by borrowing some concepts from the CS literature.

4 Proposed framework

4.1 System setup

In the first step of the proposed framework, we obtain compressed data using random projection. Mathematically speaking, we denote the complete data matrix as $\mathbf{X} \in \mathbb{R}^{N \times L}$ and the actual data matrix available for processing as $\mathbf{Y} \in \mathbb{R}^{N' \times L'}$ after applying random projection on a large network. The reduction in either N' or L' depends on whether this linear compression is deployed for reducing the feature dimension or time instances to meet the network constraints. We revisit the two cases considered previously:

Case 1: Sensor sub-sampling: We seek a linear transformation on the data $\mathbf{y} = \mathbf{\Phi x}$ where the random matrix $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$ has entries as random variables. There are some known classes of random matrices suitable for large-scale networks. For example, in the database friendly matrices [3] the entries can take values of either 0 with probability 2/3 or ± 1 with probability 1/6. If all sensors have synchronized clocks and the same random generator, a rule can be set up so that the sensors send their pre-modulated reading with ± 1 depending on the value of the random generator. Alternatively, when the sensors cannot directly reach the central node, the random gossip algorithm [38] can be applied to propagate the projection $\mathbf{y}$ to the central node (see Fig. 2 for an illustration). The additional advantage over the decentralization approach is that the central node can now perform the analysis using the residual method on the compressed data $\mathbf{y}$. We show both theoretically and experimentally in subsequent sections that the performance of the detector is nearly as optimal as if the whole data matrix $\mathbf{X}$ were available.

Case 2: Temporal stream frame sub-sampling: In this scheme, the operator can request the server to generate random numbers having values ± 1 and modulate the data with these random numbers, accumulating the values for L' different iterations where $L' \ll L$ (see Fig. 3 for an illustration). We show that by doing sub-sampling, limited bandwidth and storage can be efficiently utilized to detect anomalies as successfully as if the full data $\mathbf{X}$ is available.

In the second step, we perform *anomaly detection using compressed data.* Instead of using $\mathbf{X}$ which is not available, we now apply the residual method on the compressed data $\mathbf{Y}$, i.e. compute its eigenvalues and hence obtain the Q-statistic to determine the presence of an anomaly.

4.2 Theoretical analysis

With the following theoretical analysis, we aim at proving that doing anomaly detection with residual subspace analysis using compressed data obtained from random projection is approximately equivalent to that using complete data. Our theoretical analysis is based on relative performance to the complete data $\mathbf{X}$. Even though this complete data $\mathbf{X}$ is not practically available, a relative comparison can provide a guarantee on near optimal performance of any spectral-based detection method. To do this, we first study the changes in the eigenvalues (spectral properties) reflected in the compressed data as they constitute an important factor for detection as shown in (3). The bounds on the eigenvalues of compressed data then allow us to study the bounds on false alarm rates when the residual subspace method is applied to the compressed data for anomaly detection.

4.2.1 Case 1: M readings from N sensors.

Random projection is used to compress the columns of $\mathbf{X}$ from N to M and the its relation with the compressed data is given by

$$\mathbf{y}_i = \mathbf{\Phi}\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^M, i = 1 \dots L \tag{6}$$

Denote the eigenvalues of the complete data $\mathbf{X}$ as $\lambda_1, \dots, \lambda_N$, the eigenvalues of the compressed data $\mathbf{Y}$ as $\xi_i, i = 1, \dots, M$, K as the number of principal eigenvalues in the complete data $\mathbf{X}$, such that $K < M \ll N$. For simplicity, we assume that the CS matrix is a random Gaussian matrix. Similar results can also be obtained for other random matrices,

Theorem 1 *With a probability of at least $1 - \delta$, the changes in the eigenvalues are bound by*

$$|\lambda_i - \xi_i| \leq 4\sqrt{2}\lambda_1 \left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2 \ln \frac{1}{\delta}}{M}} \right) \tag{7}$$

for $i = 1, \dots, K$, where λ_1 is the largest eigenvalue of $\mathbf{\Sigma}_\mathbf{x}$.

The theorem is a direct consequence of the concentration property of Gaussian ensembles and the proof is detailed in Appendix 6.

Theorem 1 suggests that as the principal subspace spanned by $\mathbf{X}$ is approximately preserved in the compressed domain with high probability, and the *intrinsic* structure of the data in the original input domain is unchanged under random projection (see Fig. 4 for an illustration).

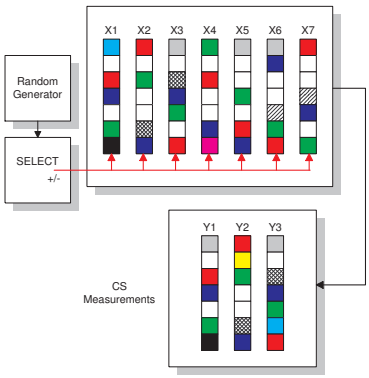


Fig. 3 Temporal subsampling. Here the data $\mathbf{x}_1, \ldots, \mathbf{x}_7$ reside in a database. The data are ± 1-randomly modulated through a random generator, and the summation is taken over all data points to obtain temporally-compressed data $\mathbf{y_1}, \ldots$.

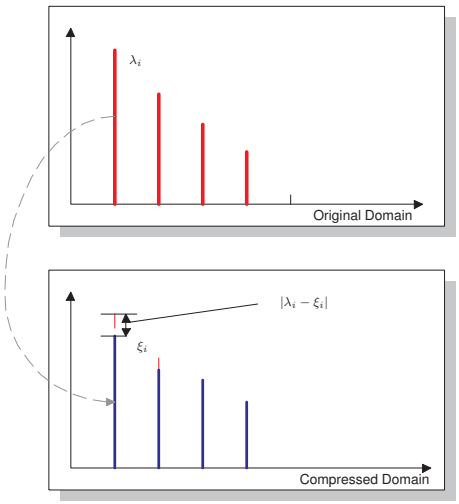


Fig. 4 Illustration of Theorem 1. The top plot depicts the eigenvalue distribution of data in the original domain, whilst the bottom one shows that of the compressed domain. Theorem 1 essentially quantifies the variation of the distribution for the principal (largest) eigenvalues.

We now discuss the implication of this result on anomaly detection in compressed data. The detection of volume anomalies using the residual subspace method is entirely based on the total power of the residuals, i.e. $\|\mathbf{z}\|^2$, rather on the actual residual subspace itself as long as it remains noise-like, i.e. no salient spectral features. It can be easily shown that when the random matrix $\mathbf{\Phi}$ is normalized (each column to unit norm), the total power is unchanged. Thus, a small variation in the principal subspace directly translates to a small change in the total power of the residual subspace. It means that as far as the statistic $t = \|\mathbf{z}\|^2$ is concerned, its distribution will also experience a small change when the compressed data is used. This intuitive argument can be more formally stated by the following result, which forms the basis for our proposed framework.

Theorem 2 *If the residual method is applied to the compressed data, with a probability of at least* $1-\delta$*, the change in the false alarm rate is bounded by:*

$$\Delta FA \leq \mathcal{O}\left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2\ln(1/\delta)}{N}}\right). \tag{8}$$

The proof is detailed in Appendix 6.

We now investigate the effect of different factors on the changes in the false alarm rate. If we fix δ in advance, the second term on the left hand side of (12) becomes significantly small as the problem size, and thus M, becomes large. Therefore, for large-scale networks, the first term is dominant. As mentioned above, the number of measurements M is related to the sparsity via $M = \mathcal{O}(K \log N)$ under CS theory. This implies that the first term will decay at the rate $\mathcal{O}(\sqrt{Klog(N)/N})$ and thus for large networks, this term is also small if $K \ll N$. Fortunately, for volume anomalies, the intrinsic dimension appears consistent with this assumption [30].

4.2.2 Case 2 : ***Sub-sampling the number of data instances.***

In the previous case, we used random projection to reduce the number of readings in data streams. Effectively, this reduces the number of rows in the data matrix $\mathbf{X}$, which is useful when N is large. In a similar manner, we now show that the proposed framework can be applied to the case when the number of instances is large. Effectively, we use random projection to compress each L-dimensional row of the complete data matrix $\mathbf{X}$ to a M-dimensional row of the matrix $\mathbf{Y}$ using a random matrix $\mathbf{\Phi} \in \mathbb{R}^{M\times L}$, where $M < L$. Mathematically, the relation between $\mathbf{Y}$ and $\mathbf{X}$ can be written as:

$$\mathbf{Y}^T = \mathbf{\Phi}\mathbf{X}^T. \tag{9}$$

In this case, $N' = N$ and $L' = M$. We now show that the results for the previous case are applicable in this case. To see this, we start from the basic result in linear algebra that:

$$\lambda_i(\mathbf{X}\mathbf{X}^T) = \lambda_i(\mathbf{X}^T\mathbf{X}), i = 1, \ldots, \min(N, L). \tag{10}$$

This implies that the changes in the principal eigenvalues of $\mathbf{Y}\mathbf{Y}^T$ relative to $\mathbf{X}\mathbf{X}^T$ is the same as the changes in eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ relative to $\mathbf{X}^T\mathbf{X}$ and as $\mathbf{Y}^T$ and $\mathbf{X}^T$ are related in a similar manner as shown in (9), the previous result applies. The only minor difference is that N should be replaced by L as the reduction is perform on the rows of $\mathbf{X}$. The changes in the principal eigenvalues are bounded by:

$$|\lambda_i - \xi_i| \leq 4\sqrt{2}\lambda_1\left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2\ln\frac{1}{\delta}}{M}}\right), \tag{11}$$

whilst the changes in the false alarm rate is bounded by:

$$\Delta FA \leq \mathcal{O}\left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2\ln(1/\delta)}{N}}\right). \tag{12}$$

with probability of at least $1-\delta$.

4.2.3 Complexity analysis

If the complete data $\mathbf{X}$ were available, the covariance matrix formation and eigenvalue decomposition requires a computational power of $\mathcal{O}(N^3)$ and memory storage of $\mathcal{O}(N^2)$ in the case of PCA. In a similar fashion, the complexity for SVD computation is $\mathcal{O}(LN^2 + L^2N)$. In contrast, the complexities for the proposed framework (both computational and storage) are only $\mathcal{O}(M^3)$ and $\mathcal{O}(M^2)$ respectively, where $M = \mathcal{O}(K \log N)$. As previously discussed, when the intrinsic dimension of the complete data is small relative to its size, significant reduction in both storage and complexity is achieved with the proposed method. We also note that if the data is sparse in the original domain, then sparse-SVD or PCA may be directly applicable. However, it is much more likely for the data to be sparse only through an (unknown) orthogonal transformation. Since data is generally dense in the original domain, sparse SVD is generally not applicable.

4.3 Justification of Theoretical Bounds

In what follow we examine whether the theoretical bounds derived previously are tight enough so that they can be a general guidance for practical purposes. In other words, *are the bounds approximately at the order of the real deviation?*

As the theoretical bound on the false alarm is naturally dependent on the bound of the eigenvalues, we restrict the discussion to the bounds in Theorem 1. In this case, the Theorem states that such a deviation of the eigenvalues should not exceed $4\sqrt{2}\lambda_1\Delta$ where Δ is dependent on $M^{-0.5}$ as

$$\Delta = \frac{\sqrt{K} + \sqrt{2\ln(1/\delta)}}{\sqrt{M}}. \tag{13}$$

For a 90% confidence, the term $\sqrt{2\ln(1/\delta)}$ is 2.1460, which is small. Thus, the dependence is approximately $\sqrt{K/M}$. This dependence provides the implications for the application of compressed sensing/random projection

- The bounds only make sense if the intrinsic dimension K is sufficiently small compared with the reduced dimension M
- *Compression-error trade-off.* For a fixed intrinsic dimension K, better compression would be achieved with a smaller M, however it also results in a larger possible deviation. On the contrary, error is made small when letting $M/N \to 1$, but this defeats the goal of compression. This mean in practice a value of M such that $K \ll M \ll N$ would provide a natural trade-off. In the compressed sensing theory, such a value of M is typically chosen as $M = \mathcal{O}(K \ln(N))$, which appears to be suitable for the compressed sensing recovery problem. In large-scale network problem, there are two possible scenarios:
  - The designer is given a maximum affordable M: in that case, the bounds serve as a rough estimate of the the possible deviation due to using compressed data
  - The designer is given a tolerance on the deviation: in this case, the bounds serves as a rough estimate of the compression required to attain accuracy within the allowable tolerance.

The theoretical bounds are derived on the assumption of distinct principal components. In other words, the principal K eigenvalues are assume sufficiently larger than the residual eigenvalues. There are two practical issues that we emphasize:

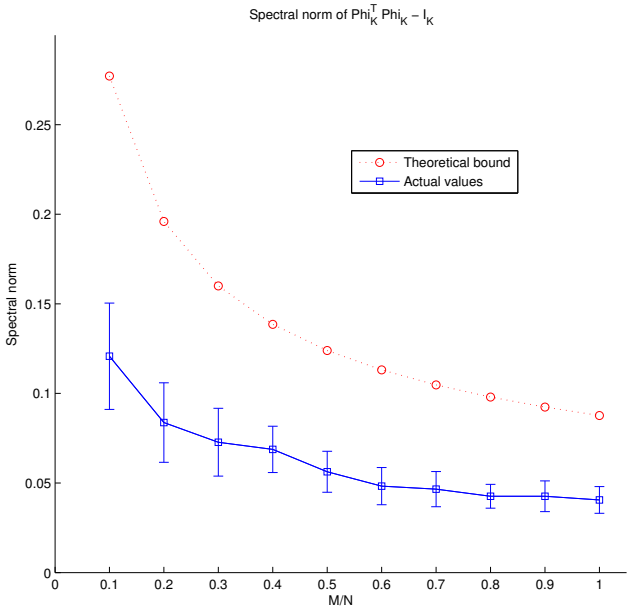


Fig. 5 Theoretical and actual values of $\|\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{I}_K\|_2$

– How do we determine K, especially since we are unable to work on the original data if it is too large. The theoretical results indicates that for sufficiently large M, one may determine K using the compressed data instead. It means that one can start with the maximum affordable value of M and analyze the compressed covariance matrix to determine K.
– In practice, the eigenvalues may follow a decay distribution in many cases. When this happens, there is no clear choice of K. As with the well-known principal component analysis in statistics, one typically chooses a cut-off point at which at least, for example, 90% energy is retained. This is what we use in our work and works rather well.

We now turn the discussion to the tightness of the eigenvalue deviation bound. Upon examining the proof, we found that the bounds are reasonably tight and it is difficult to improve any further. Our bounds depend on the concentration result of Gaussian random matrices which state that for a particular Gaussian random matrix $\mathbf{\Phi}_K$ of size $M \times K$ where each entry follows $\mathcal{N}(0, 1/\sqrt{M})$ the singular values are bounded by $1 \pm \Delta$ with a probability of at least $1 - \delta$, where Δ is defined as (13). This implies that the singular values of $\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{I}_K$ are bounded by $1 \pm 2\Delta$. *How tight is this theoretical bound in the literature?* To do so, we study the case where $K = 5, N = 10^4, \delta = 0.1$ and vary M. The theoretical and actual values of $\|\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{I}_K\|_2$ are shown in Fig. 5. From this plot, we observe that the theoretical bound appears to be reasonable, though it is slightly conservative. It is about twice the actual value on average. Thus, when using this several times, we expect the bounds on the eigenvalues are also a few more times larger than the actual values. To verify this, we consider the same setting and create a covariance matrix whose the 5 principal eigenvalues are 100. Then we directly compute the average eigenvalues of the compressed covariance matrix obtained from random projection whilst M is varied. Fig. 6

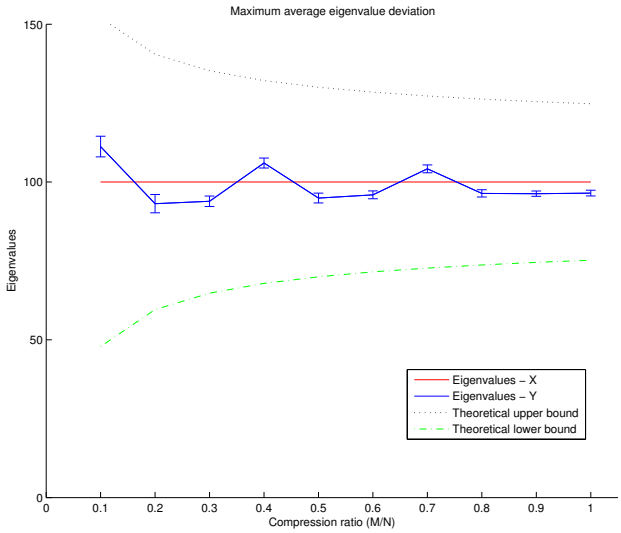


Fig. 6 Maximum average principal eigenvalue deviations

shows the maximum average deviation of the compressed eigenvalues, whilst Fig. 7 shows the deviation of all principal eigenvalues at a particular $M = N/2$. The plots indicate that the theoretical bounds are about 4-5 times larger than the actual one, but this is as expected. Obviously, if the bound on spectral norm of Gaussian random matrix is tighter, our theoretical bound will be also tighter. However, it is noted that we have used the best bounds available for Gaussian random matrices in the literature to date.

Whilst improving the bounds might be of future interest, the theoretical results importantly gives us a justification to the scalable framework that deals with large-scale network data using random projection/compressed sensing via spectral methods. In the experimental section, we shall illustrate this more clearly.

4.4 Discussion

From the above cases, one may also apply sub-sampling in both dimensions if they are both large. In this case, it is natural to assert that the bound of the change in the false alarm is the sum of the bounds derived in the above theorems. Thus, the bound is unavoidably increased for trade-off both in the feature and instances dimensions. Nevertheless, it is only linear to the reductions.

We note that our proposed method is rather general, and has not yet taken into account the specific structure of a particular network. Without specialization, it is more suited to the abnormality detection in general centralized networks. It is of interest to extend the general theory here to the case of particular network structure. For example, in the case of sensor networks where the sensors are organized in clusters or tree [21], it might be more desirable to do in-network processing to further reduce the bandwidth. We thus leave the specialization as future extension of the developed theory.

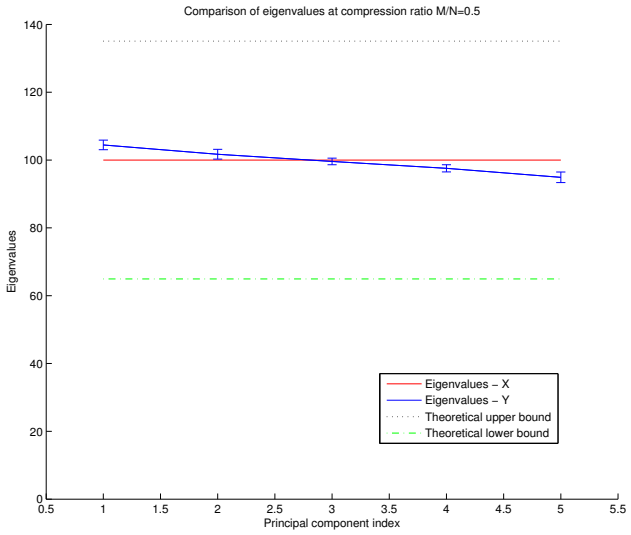


Fig. 7 Principal eigenvalue deviation at $M/N = 0.5$

Finally, even though it is intuitive that random projection approximately preserves the geometry, and thus one should expect the eigenvalues to follow suit, our theoretical results are the first in the literature to provide two-sided bounds on *all* principal eigenvalues, and thus this is a significant contribution to the literature. In comparison, we note the limitation of some previous work:

- The text-book result [41] only provides a non-probabilistic one-sided bound for the eigenvalues. Here, the bound is expressed in terms of the ε deviation due to random projection, but this value is too difficult to compute for a given projection matrix.
- The early KDD paper [7] only gives practical demonstration that principal component analysis works with random projection, but no theoretical justification is given as thoroughly as what we present here.
- The recent work in [18] studies *only the first* eigenvalue, but with a deterministic orthogonal projection matrix. Here, we use random Gaussian matrices that are only approximately orthogonal, and that we provide probabilistic bounds for *all* principal eigenvalues.

Moreover, we also derive the bound on the deviation of the false alarm, which is specifically tailored to the anomaly detection problem.

5 Experimental Results

We evaluate our framework using both real-world and synthetic datasets experiments from two application areas: network traffic analysis and video surveillance analysis. We evaluate the performance of our framework for the sensor sub-sampling case with network data while the temporal frame sub-sampling case performance is studied using video data.

In all cases, we compare the proposed anomaly detection using the proposed strategy (random projection) with the other relevant alternatives including the full data [30] and the decentralization approach of [23]. We emphasize the following remarks

- As this work is concentrated on scalability strategy rather than outlier detection algorithms, we compare different *strategies* for the same outlier detection algorithm using residual subspace analysis to verify the scalability solution analyzed in this work. As mentioned earlier in the introduction, the analyzed random projection strategy might also work with other non-spectral algorithms, but it is beyond the scope of current work. Furthermore, some other non-spectral algorithms might be application specific, which requires additional assumptions and settings and comparing them directly is almost impossible. Thus, we restrict out attention to residual subspace analysis as the outlier detection algorithm for consistency.
- Our main goal is to demonstrate that in a wide range of circumstances in data mining, the proposed framework is indeed useful. In other words, the trade-off between compression and deviation can be satisfactorily achieved. Thus, the compression parameter M does not mean as a tuning hyperparameter, but should be viewed as a trade-off parameter in the context. Also, for consistency, we determine the parameter K by using the 90% principal energy as the guiding principle wherever we do not know K *a priori*.

5.1 Sensor Subsampling Results

The anomaly detection capability of the proposed framework is evaluated on a real-work benchmark data set Abilene [1] and a synthetic dataset specifically designed to simulate a large network.

5.1.1 Experiment on the Abilene Data

Abilene Dataset. The Abilene dataset [1] is a well-known dataset for network research, and captured from a real-work backbone network. Its first use for volume anomaly detection is documented in [30]. Here, we are primarily interested in anomalies resulting from abnormal changes of the network traffic. The changes arise because of events such as abnormal DNS transaction, network equipment failure, flash crowd occupancies, distributed denial-of-service (DDoS) attacks [30]. Importantly, these changes cannot be detected from a simple thresholding due to the varying characteristics of normal traffic during a day. In Figs. 18(a) and 8, we plot the total network traffic for a period with and without volume anomalies (highlighted in red). As can be seen, volume anomalies are often hidden under normal network traffic. The preliminary investigation in [30] reveals that the spectral approach is capable of detecting these anomalies. Our purpose in this experiment is to extend [30] to the case where the data is not complete and verify the proposed method.

Data collection. In the Abilene network, the traffic flow is the amount of traffic flowing in between each pair of ingress and egress nodes in the network. It is also known as an origin-destination (OD) flow, which is the traffic that enters the backbone at the origin *point of presence* (PoP) and exits at destination PoP [30]. The Abilene dataset consists of the readings collected from 41 network links over a period of several months. The OD trace contains the measurement from each link for each 10-second interval. We use a subset of the data which covers a period of 2 weeks (1008 measurements per week). Most of the data reflects normal network conditions with only 6 real anomalies (verified manually) in

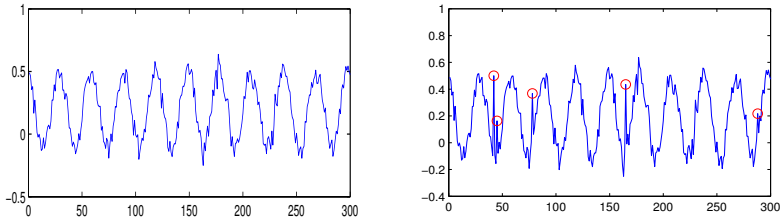


Fig. 8 Typical network link data: Normal (left) and Abnormal (right). Horizontal axis denotes the aggregated window index, vertical axis denotes the ℓ_1-norm of the vector of total traffic.

the original dataset. In addition, we inject 45 synthetic anomalies of different magnitudes following the procedure described in [30]. We use the first week data for training and the second week data for testing.

Improving Random Projection. As discussed earlier, the random projection is represented by a matrix $\mathbf{\Phi}$ from which the compressed data is obtained via $\mathbf{y} = \mathbf{\Phi x}$. A good random projection matrix should have columns as close to orthogonal as possible (in other words small RIC) so that the geometry of data in high-dimensional space is preserved better in the low-dimensional space. In other words, the mutual coherence of the overcomplete system $\mathbf{\Phi} = [\phi_1, \ldots, \phi_N]$, which is defined as:

$$\mu(\mathbf{\Phi}) = \max_{i \neq j} |\langle \phi_j, \phi_k \rangle|, \tag{14}$$

must be as small as possible. For a real-valued matrix $\mathbf{\Phi}$, the lower bound on the mutual coherence is known as the Welch bound [39]:

$$\mu(\mathbf{\Phi}) \geq \sqrt{(N-M)/(M(N-1))}. \tag{15}$$

For many classes of random matrices, the mutual coherence can be small with high probability. In cases where the problem size is not very large, such as in this Abilen dataset with N is only 41, a random Gaussian matrix might not have good approximate orthogonality property, which necessitates improvement in practice. To further improve the random projection matrix, we start with a random Gaussian matrix and then apply the recently proposed algorithm by Elad [17]. This algorithm exploits the fact that the mutual coherence of $\mathbf{\Phi}$, with each column normalized to unit norm, is the maximum magnitude of the off-diagonal elements of the Gram matrix $\mathbf{G} = \mathbf{\Phi}^T\mathbf{\Phi}$, where the Gram matrix has rank M. Hence, by iteratively shrinking the entries of the Gram matrix, forcing its rank to M, and taking its square root, a smaller mutual coherence for $\mathbf{\Phi}$ with a specified rank M is achieved. Though the algorithm could be sensitive to the parameter setting and its convergence is yet to be studied, we found that in practice this method improves the mutual coherence considerably. In practice, the actual signal $\mathbf{x}$ might not be sparse in the basis $\mathbf{I}$ but in some $\mathbf{\Psi}$. In this case, $\mu(\mathbf{\Phi\Psi})$ needs to be small instead. If $\mathbf{\Phi}_o$ is the optimal sensing matrix for the basis $\mathbf{I}$, then the optimal matrix $\mathbf{\Phi}$ for the basis $\mathbf{\Psi}$ is found from $\mathbf{\Phi} = \mathbf{\Phi}_o\mathbf{\Psi}^{-1}$, assuming that $\mathbf{\Psi}$ is invertible. For the Abilene network data, $N = 41$, $M = 16$, $K = 6$ and the Welch bound of the sensing matrix is 0.1976. Using Elad's algorithm [17], we achieve a mutual coherence of 0.36 from the initial coherence of 0.55.

Anomaly detection. Residual subspace analysis is applied using different strategies: complete, compressed, and decentralized data. Figures 9 and 10 show the similarity between the principal eigenvalues and the residual vectors using these strategies. We then

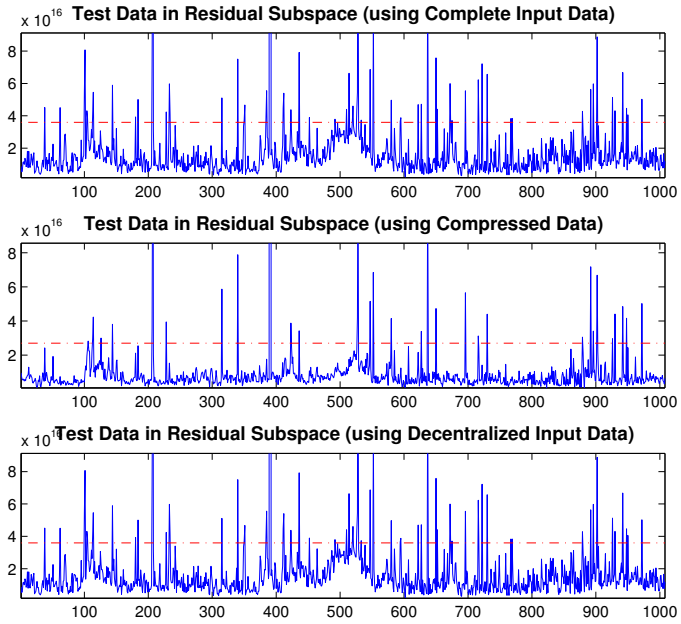


Fig. 9 Residual vector plots for Abilene data.

conduct experiments to obtain the receiver operating characteristics (ROC) curve which is shown in Figure 11. We notice that the performance is very similar with all three strategies.

The performance on the compressed data is very close to that on the uncompressed and decentralized strategies. To further quantify this, we also compare the ROC curves using (i) the area under the ROC curve (AUC) and (ii) equal error rate (EER). An effective classifier should achieve an AUC close to 1 and small ERR. From the ROC curves, we determine that the AUC/EER values are 0.95/0.09, 0.96/0.11, and 0.95/0.10 for the original, compressed, and decentralized data respectively.

5.1.2 Experiment on Synthetic Network Traffic Data

Synthetic Data Generation. We generate synthetic network traffic data following the procedure in [30]. In particular, we consider a network where the number of local monitors N ranges from 500 to 2000 and the number of time instances $L = 2000$. The network traffic signal is modeled as $\mathbf{x}$ as $\mathbf{x} = \mathbf{s} + \mathbf{n}$ where $\mathbf{x} \in \mathbb{R}^N$. It consists of two parts: $\mathbf{s}$ characterizes the long-term structure in the data and $\mathbf{n}$ represents the local temporal variation. For the long-term network traffic signal, $\mathbf{s} = \mathbf{\Psi}_s \alpha_s$. Here, $\mathbf{\Psi}_s$ is the basis for the intrinsic network data. Due to its daily periodic characteristics, we select the discrete cosine transform (DCT) matrix as the basis. The number of principal components is $K = 4$. We simply model the noise as zero-mean Gaussian with variance $\sigma^2 = 0.01$. To simulate abnormal network conditions, we inject 70 anomalies of different magnitudes.

Anomaly detection. When specifying the dimension M for random projection, we need to consider the trade-off between performance and error rates. Selecting a smaller value of

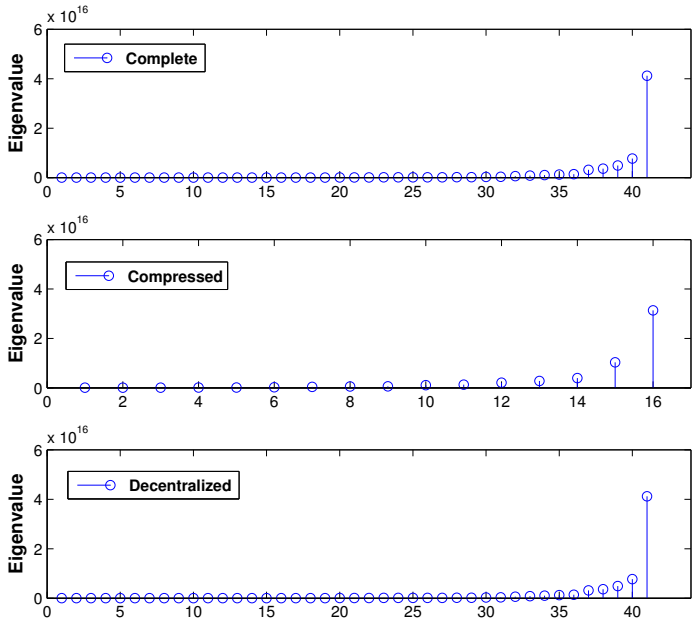


Fig. 10 Eigenvalue distribution for Abilene data.

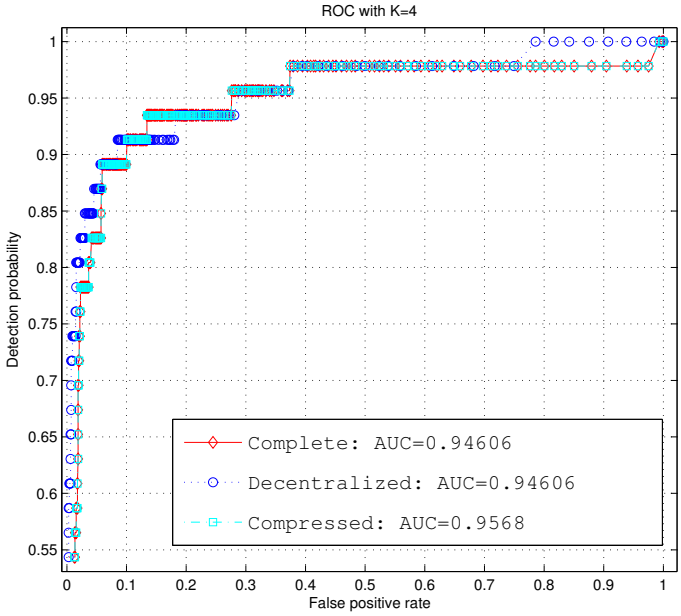


Fig. 11 ROC curves for the Abilene data

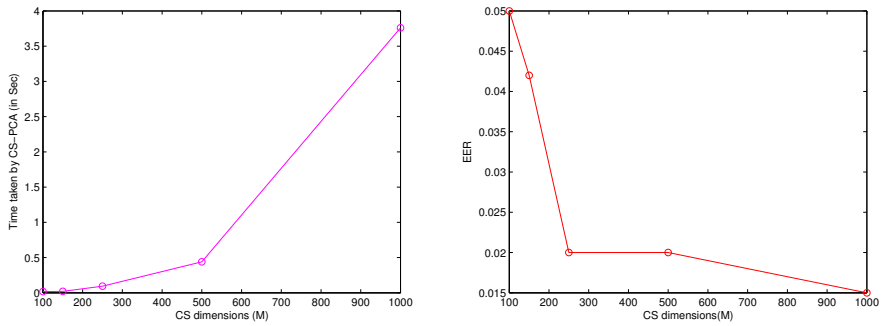


Fig. 12 Trade-off: computation (left) and error rate (right).

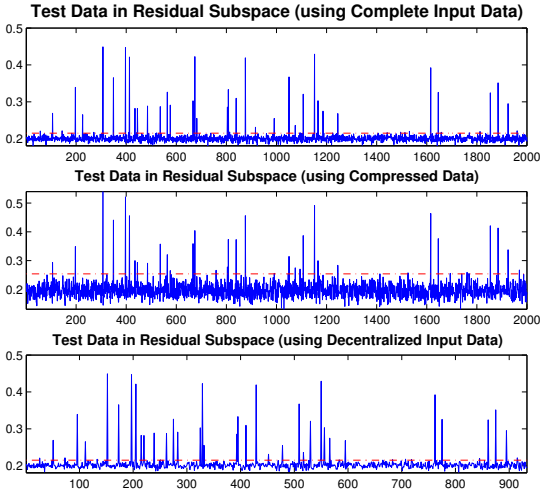


Fig. 13 Residual vector plots for synthetic data.

M reduces the computational complexity at the cost of a potentially lower performance due to the increase in the mutual coherence of the random projection matrix. In the CS literature, the value of $\mathcal{O}(K \log N)$ has been frequently suggested. We set $N = 2000$, vary M between 100 to 1000 and measure the EER and computational time. The results are shown in Figs. 12(a) and 12(b). Selecting M in the range $250 - 300$ gives moderately low error rates at a large reduction in computational time. If M is too low, the error rate becomes larger. If M is too large, the reduction in error rate is not significant whilst the computational time increases somewhat quadratically. Therefore, we determine that suitable values of M are 118, 280, and 450 when number of nodes are 500, 1000 and 2000 respectively. The random projection matrices are improved from random Gaussian with a final mutual coherence of 0.37, 0.35 and 0.20 respectively.

With the specified random projection matrix, we then examine the behavior of the residual vectors using complete and compressed data. Figs. 13 and 14 demonstrate the result.

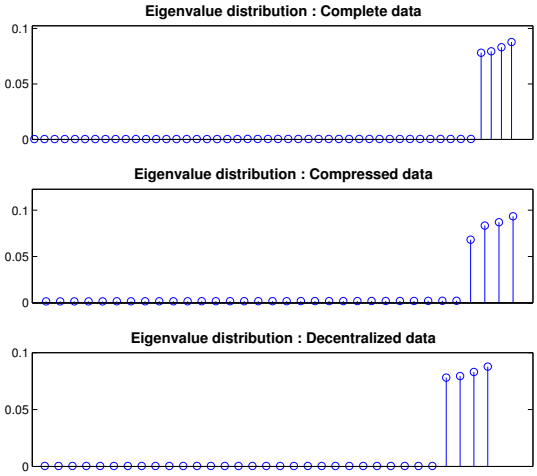


Fig. 14 Eigenvalue distribution for synthetic data.

Once again, we observe that the patterns of the eigenvalue distribution and the residual vectors are similar in both complete and compressed data cases. We then explore the detection performance by comparing residual subspace analysis using compressed data with that using complete data [30], and the decentralized version presented in [23]. The ROC curves for these three cases are shown in Fig. 15. In terms of AUC, the detection with the compressed data is approximately equivalent (even sightly better than) to the other cases.

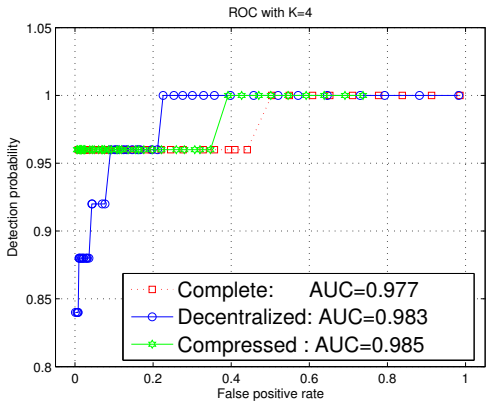


Fig. 15 ROC performance

Next, Figure 16 compares these three strategies in terms of communication, computation and storage overhead. It shows that by using compressed data compared to other two strategies we can reduce the communication bandwidth by 45% to 60% , computational cost by 80% to 90%, and storage requirement 45% to 70% as compared with other two strategies.

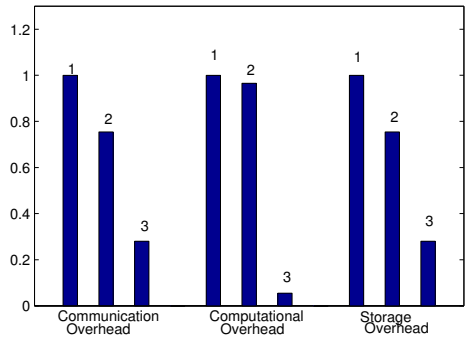


Fig. 16 Communication, computational and storage complexity. The strategies are: Complete Data (1), Decentralization (2), and Compressed Data (3)

Finally, we vary the size of the network from 100 to 2000 and the result is shown in Table 1. The AUC and EER indicators are as competitive as those using complete data, with a better computational performance, it is from 6 to 1000 times faster.

Table 1 Anomaly detection performance on synthetic data.

Metric	N	100	500	1000	2000
Time (seconds)	Complete	0.023	0.430	3.364	20.932
	Compressed	0.004	0.023	0.097	0.203
	Decentralized	0.021	0.425	3.360	20.90
AUC	Complete	0.993	0.996	0.982	0.986
	Compressed	0.997	0.991	0.984	0.979
	Decentralized	0.995	0.994	0.981	0.98
EER	Complete	0.060	0.080	0.090	0.090
	Compressed	0.060	0.080	0.020	0.020
	Decentralized	0.060	0.080	0.090	0.10

5.2 Temporal Frame Subsampling

5.2.1 Problem Background and Datasets

Next we evaluate the performance of our framework when dealing with temporal sampling (Section 4.2.2) and determine the suitability of our work in addressing a real-world surveillance problem faced of a local public transport authority. The local train network monitoring system consists of over 3000 cameras operating 24 hours a day. Constant human operator supervision of video is impossible. The problem of detecting anomalies in the video is challenging because (1) most anomalous patterns occur in the presence of normal patterns (a majority of people behave normally) and (2) there is no predefined description of anomalous behavior - the anomaly changes with the context of the scene (a person walking at normal pace on the rail tracks is considered as anomalous whereas the same behavior would be normal if it occurs on the station platform).

We use two video surveillance datasets: one provided by the local public transport authority and the PETS 2007 benchmark dataset [2]. The PETS 2007 data was used to demonstrate the effectiveness of our work on an established dataset as the video sequences are freely available from the PETS archive [2]. Both video datasets are pre-processed to extract motion features. In both cases, the ground truth data is available. Training is done offline and testing is performed on the incoming data streams

5.2.2 Video Data Pre-Processing and Feature Extraction.

We use optical flow [33] to define the motion in the scene. The advantage of such low-features, collected in a grid superimposed on the images, is that they provide good information about motion whilst alleviating the need for object tracking [4]. The limitation of this type of feature is that it is limited to motion anomalies. In many practical situations, this is sufficient.

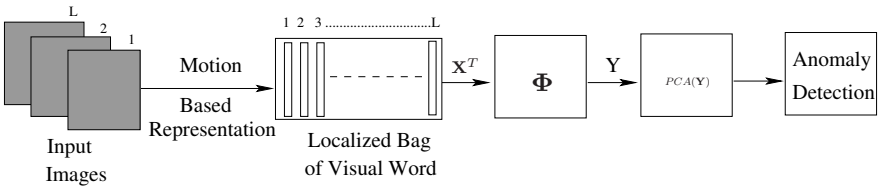
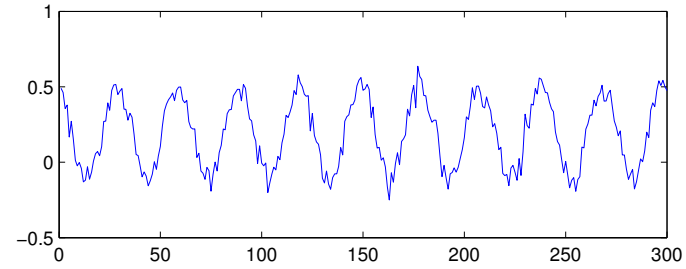


Fig. 17 Schematic description of the proposed method.

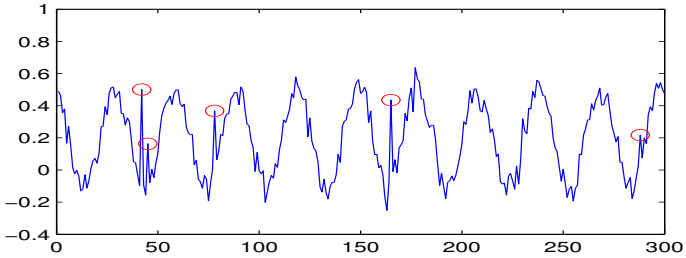
5.2.3 Motion-based representation:

Each image is divided into grid-based regions and the motion statistics of each cell in an image is computed over a pre-defined time bin. The motion distribution of each cell is simply calculated as the number of optic flow vectors in that cell. Figures 18(a) and 18(b) show the the amalgamated motion flows over 300 frames for 2 sequences from the data collected from surveillance cameras in a train station. In Figure 18(a), we plot the volume of the motion flows, which is the squared norm of the vector of motion statistics from all cells. Each point shows the motion statistics amalgamated over one minute. Generally, the characteristics of the *normal* motion volume is "high" if there is a train in the station, otherwise it remains "low" giving rise to periodic rise and fall.

The majority of past relevant work [8, 34, 37] has treated the whole scene as either "normal" or "abnormal", but these examples suggest that a framework which can detect abnormality in presence of the normal behavior is needed. Our intuition is to capture the structure of the overall normal pattern in the principal subspace. We now give some analytical arguments to justify the proposed method. In the subway example, passengers would normally follow the "walk" path to enter and exit. The normal activities induce a distribution of motion vectors over the cells in the subway. Importantly, this distribution also signifies the relationship between the cells. For example, some cells tend to be highly correlated due to the average flow of the traffic through the cells. This dependence gives us important information about the structural pattern of normal events. Thus, if, for example, a person or a group crosses the subway tunnel in an unusual manner, the observed motion distribution will



(a) During a normal period



(b) During a period with anomalies

Fig. 18 Amalgamated optical flows over one minute interval. Horizontal axis denotes aggregated window index, vertical axis denotes normalized ℓ_2-norm of the aggregated motion count vector.

carry totally new structural information. This is the basis for our abnormality detection technique. Similarly, a loitering person is likely to lead to the observation that the cells covering the loitering trajectory become more correlated than normal. If this structural information is known, we can separate the normal activity by projecting the observed motion pattern onto the space induced by the structure, so that abnormal activities can be easily investigated in the residual subspace.

Localized bag-of-visual-words: We are motivated by [35] to use *bag-of-visual words* for representing the optic flow count in the cells. Niebles *et al.* [35] derive visual words from the the human activity in the spatio-temporal domain. Using a grid-based approach we extract optic flow counts for each cell. We consider each cell similar to a *term* and motion statistics (i.e. number of motion flows) of each cell as equivalent to *word* frequency. Hence, the number of *terms* is equal to the number cells in the image. We construct the feature-frame matrix in an analogous manner to the term-document matrix. Denote the number of cells as N and the motion statistics of cell i at frame l as $x_i(l)$. The vector of motion statistics is defined as $\mathbf{x}_l = [x_1(l), \ldots, x_N(l)]^T$. For a sequence of L frames, the feature-frame matrix is defined as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$. In document analysis, the semantic variables (topics) govern the probabilistic occurrence of the terms. Similarly in our case, the structural variables of the normal behavior govern the distribution of motion vectors over cells.

For detection, we project the compressed data onto the compressed residual subspace and use the Q-statistic as described in Section 3.1. Figure 17 shows the schematic description of the proposed method.

5.2.4 Results on the PETS2007 Benchmark Video Data.

The PETS2007 [2] dataset consists of video footage obtained from a multiple camera surveillance system. For the PETS datasets processing, the time granularity for aggregating the motion count was set to 100 frames. For training, we use the S0 sequences which consisted of 4500 frames captured at a resolution of 720×576. The sequences containing no *unusual* events and no externally injected "actors", and the crowd density typically (depending on the camera) is medium.

The first PETS2007 test sequence used was S3 captured by camera 1. The sequence consists of 2970 frames and the anomaly was a theft event. It involved two actors walking normally towards the middle of the scene where after a brief stop, they proceed to pick a bag and leave the area. As this event takes place, there is a significant flow of people in the top part of the scene. Hence, for our approach to produce the correct results, it would need to detect the anomaly and correctly highlight the time interval over which the anomalous event take place. As the sequence is short, PCA was applied directly on the features extracted from the training sequence and the eigenvalues are plotted in Figure 19. We chose the largest three eigenvalues (i.e. $K = 3$) for the principal subspace, while the rest of the eigenvectors span the residual subspace. The threshold Q_β was computed according to the $1 - \beta$ confidence level and we chose $\beta = 0.005$. Figures 20(i) and (ii) show the projection of each column of $\mathbf{X}_{\mathrm{Train}}$ and $\mathbf{X}_{\mathrm{Test}}$ into the residual subspace and the horizontal line denotes the threshold Q_β. The theft event is clearly highlighted in the residual subspace as the threshold Q_β is exceeded. It should be noted that the plot for the residual domain shows two peaks, which correspond to events which are 50 frames apart and are thus considered to be part of the same anomalous event. A number of corresponding frames within the area where the anomaly takes place are shown in Figure 21.

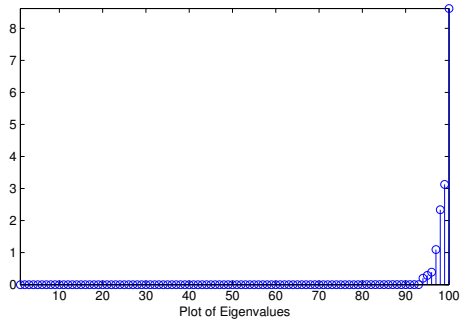


Fig. 19 The magnitude of the eigenvalues computed from matrix $\mathbf{X}$ for $\mathbf{X}_{\mathrm{Train}}$ (sequence S0, camera 1).

Similar results were obtained for sequences S6 from camera 1 (anomaly involves more than 3 actors) and sequence S3 from camera 2. To test the robustness of the approach in different environmental conditions, we use sequence S3 captured by camera 3 as a test set (with the sequence S0 from camera 3 used for training). Figure 22 show the residual vectors of the training and testing sets. Despite the different lighting and camera angles, our approach detects the anomalous event successfully. The results from the four sequences are summarized in the top four entries of Table 2.

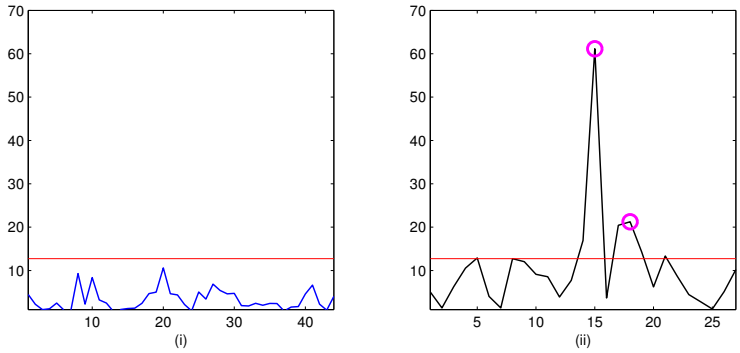


Fig. 20 Plots of the residual vectors over train data (subfigure (i): sequence S0, camera 1) and test data (subfigure (ii): sequence S3, camera 1) from the PETS 2007 dataset. Horizontal axis denotes aggregated window index, vertical axis denote squares of ℓ_2-norm.

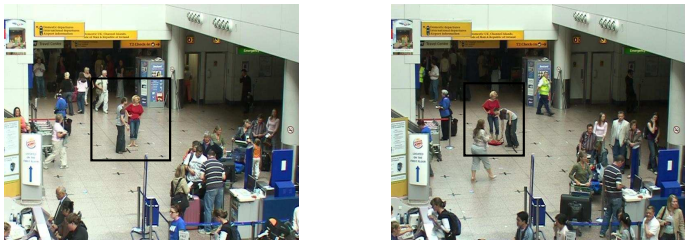


Fig. 21 The detected "anomalous" activity in sequence 3 from camera 1 (PETS2007 dataset).

5.2.5 Results on Public Transport Authority Dataset:

The third set of experiments use video footage from four different train stations. The combined set consists of surveillance video data collected over a week. Importantly, it contains several types of anomalous events that were not artificially created and were ground-truthed in conjunction with the transport authorities. In the previous experiment on the PETS dataset, we have shown the effectiveness of volume anomaly detection framework. Here, we demonstrate the scalability of our proposed approach for this type of data.

For the first evaluation, we used the video data captured from the corridors of the train station in the peak hours of the day (7AM to 11 AM) over a week. The 25fps video data at resolution 570×720 is collected by two different cameras at the entry and exit points of the train station. For the training set $\mathbf{X}_{\mathrm{Train}}$ we used video from five consecutive days where each day has 4 hours continuous video and day 6^{th} ($\mathbf{X}_{\mathrm{Test1}}$) and 7^{th} ($\mathbf{X}_{\mathrm{Test2}}$) are used for testing only. For training, the original number of aggregated time bins is $L = 7200$, the number of grid cells is $N = 100$, and the window length is 10s.

As mentioned Section 4.2.2, we have sub-sampled the temporal stream data, so that the number of snapshots is reduced to M when the length of the snapshots (L) is large and $M \ll L$. The challenge was to select the value of M for an optimal performance. Figure 23 shows the plots for the false positive rate (FPR) and the rate of anomaly detection (both were

normalized to 1), when M varies from 100 to 300 for the above mentioned datasets. When M is in the range of $190 \sim 230$, the FPR is at a minimum and detection rate is maximized. Hence, we have used $M = 220$ for this experiment.

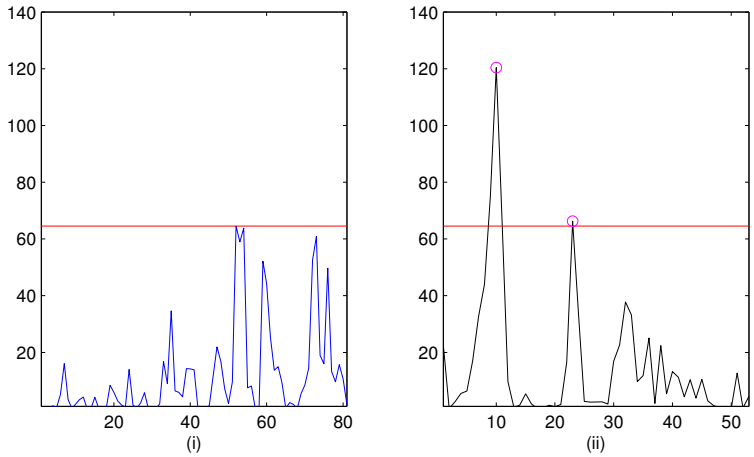


Fig. 22 Plots of the residual vectors over train data (subfigure (i): sequence S0, camera 3) and test data (subfigure (ii): sequence S3, camera 3) from the PETS 2007 dataset. Horizontal axis denotes aggregated window index, vertical axis denote squares of ℓ_2-norm.

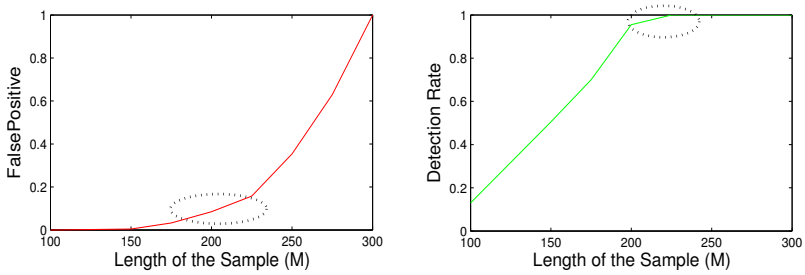


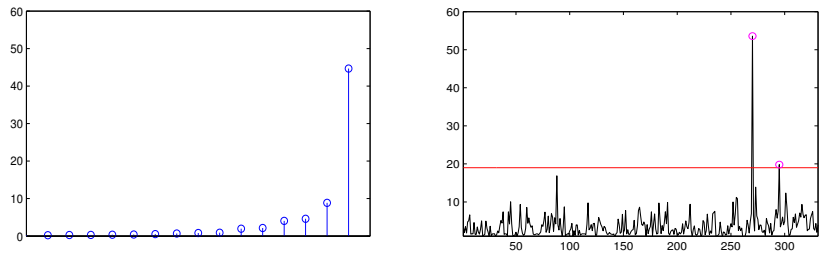
Fig. 23 Plots for FPR and the rate of anomaly detection, when M varies from 100 to 300 for the above mentioned full datasets

We first examine the behavior of the eigenvalue distribution and the residual vectors using different strategies to verify the theoretical contributions. Figs 24, 25, and 26 show the results for complete, compressed, and decentralized data. We select $K = 4$ to cover 90% energy in the principal subspace respectively. The threshold Q_β was computed in a similar way to the previous experiment with the desired false alarm being $\beta = 0.005$. The residual plots are similar in all three cases. We detect two real anomalies out of three from the test data with the detected anomalies corresponding to (1) an adult rubbing a small child against the wall and (2) a group of people loitering (Fig. 27). The missing anomaly was due to the fact that it take place far away from the camera and the motion features are unreliable. We

repeat the same experiment with the second test set ($\mathbf{X}_{\mathrm{Test2}}$) and detect the anomalous event "group loitering" (shown in Fig. 27) which occurred during "off-peak" hours.

For the second evaluation, we use five video sequences captured from cameras covering the stairs (1 sequence), an automated vending machine (1 sequence) and the rail tracks from two different stations (3 sequences). Both the stairs and vending machine sequences are long (8 and 16 hours respectively). In the case of the rail tracks data, two of the train and test sequences are short, while the third sequence is again very long (18 hours). In all cases the video was captured at 25 frames per second with a resolution of 570×720. For the training set $\mathbf{X}_{\mathrm{Train2}}$, we use a total of 27 hours of continuous video (without any anomaly) and 55 hours of video for testing ($\mathbf{X}_{\mathrm{Test3}}$ - some of the video which involve zoom action was removed as we restrict the evaluation to static views).

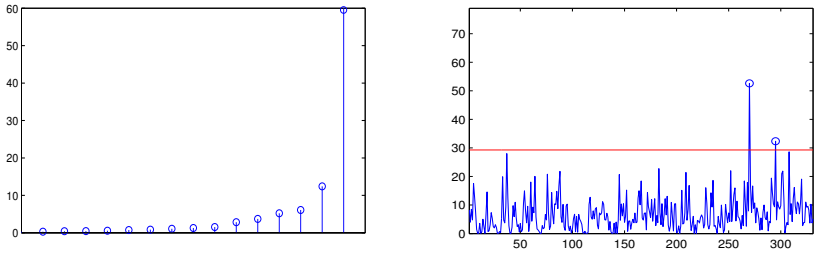
For $\mathbf{X}_{\mathrm{Train2}}$, the parameters were $L = 36,294$ and $N = 100$ while the M value is set to 220 for the long video sequences. For the two shorter video sequences, the value of M is set to 50. For all five sequence, the threshold Q_β is computed in a similar way to the previous experiment with $\beta = 0.005$. A total of 14 anomalies are present in the video streams and our approach is able to identify 13 anomalies correctly while producing one false negative and 8 false positives. One false positive is due to difficulty of differentiating between a person breaking into the vending machine (who opened the machine with a cordless drill) and the maintenance person (who also opened the vending machine with a drill). The false negative is due to the movement in the camera far field.



(a) Eigenvalue distribution. Horizontal axis: component index, vertical axis: eigenvalue.

(b) Residual vectors. Horizontal axis: aggregated window index, vertical axis: squared ℓ_2-norm.

Fig. 24 Data statistics when using complete PTA dataset.



(a) Eigenvalue distribution. Horizontal axis: component index, vertical axis: eigenvalue.

(b) Residual vectors. Horizontal axis: aggregated window index, vertical axis: squared ℓ_2-norm.

Fig. 25 Data statistics when using compressed PTA dataset.

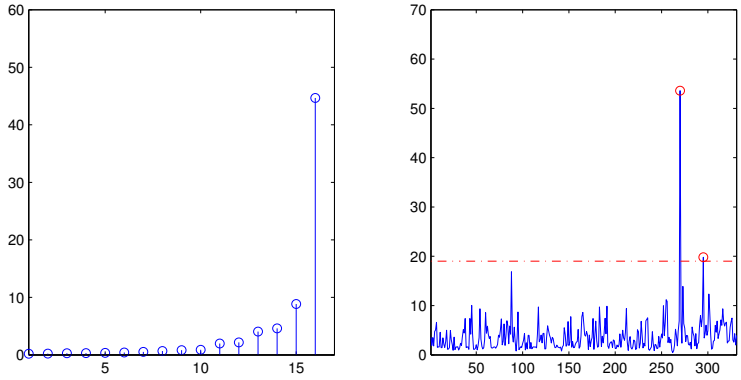


Fig. 26 Data statistics when using compressed PTA dataset. Horizontal axis denotes component index (left) and aggregated window index (right). Vertical axis denotes eigenvalue (left) and square ℓ_2-norm (right).

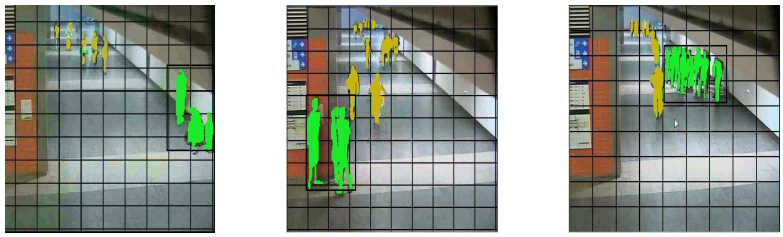


Fig. 27 Anomaly detection in Public Surveillance Data: leaning on the wall (left), hanging out in groups (center) and group loitering (right)

Table 2 Summary of anomaly detection results: temporal subsampling case.

DataSets	No. of Frames used for Training	No. of Frames used for Testing	Real Anomalies	Detected Anomalies	False Positives	False Negatives
PETS(CAM 1 SEQ 3)	4,500	2,971	1	1	0	0
PETS(CAM 1 SEQ 6)	4,500	2,735	1	1	0	0
PETS(CAM 2 SEQ 3)	4,500	2,971	1	1	1	0
PETS(CAM 3 SEQ 3)	4,500	2,972	1	1	0	0
PTA Dataset (Tunnel)	90,000	90,000	3	2	0	1
PTA Dataset (Stairs)	100,000	36,770	2	2	0	0
PTA Dataset - Rail Track (Station1 Cam 1)	6,234	10,504	2	2	0	0
PTA Dataset - Rail Track (Station1 Cam 2 (Far view)	6,363	10,479	2	2	3	0
PTA Dataset - Rail Track (Station2 Cam 1 (Far view)	450,239	660,002	3	3	5	0
PTA Dataset - Soft Drink Vending Machine	529,410	1,311,890	2	2	1	0

6 Conclusions

We have presented a framework for detecting anomalies in data streams captured by large-scale sensor networks. The work addresses a key problem of dealing with incomplete data because of the physical constraints imposed in large-scale networks. The framework further establishes the usefulness of random projection/CS as an effective solution for anomaly detection for both the case when the number of sensors or the number of data instances exceeds the communication bandwidth in a sensor network. The work exploits the fact that the intrinsic dimension of the data in typical sensor network applications is generally small relative to the raw dimension and the fact that compressed sensing is capable of capturing most information with few measurements. We have shown that spectral methods used for anomaly detection can be directly applied to the compressed data with guarantees on performance and we have demonstrated the effectiveness of the framework using both real and synthetic datasets.

Proof of Theorem 1

First, we can assume without loss of generality that the covariance matrix in the original domain $\mathbf{\Sigma_x}$ is diagonal. Indeed, suppose that the its eigenvalue decomposition is

$$\mathbf{\Sigma_x} = \mathbf{\Psi \Lambda \Psi}^T, \tag{16}$$

then the covariance matrix in the compressed domain is

$$\begin{aligned} \mathbf{\Sigma_y} &= \mathbf{\Phi \Sigma_x \Phi}^T \\ &= \mathbf{\Phi \Psi \Lambda \Psi}^T \mathbf{\Phi}^T \\ &= (\mathbf{\Phi \Psi})\, \mathbf{\Lambda}\, (\mathbf{\Phi \Psi})^T . \end{aligned} \tag{17}$$

But as $\mathbf{\Psi}$ is an unitary matrix due to the definition of the eigenvalue decomposition, it follows that $\mathbf{\Phi \Psi}$ is also a random Gaussian matrix with the same statistical properties as $\mathbf{\Phi}$ due to Lemma 1. Thus, in the study of the eigenvalues in the compressed domain, we can safely assume $\mathbf{\Psi} = \mathbf{I}$ to simplify the maths. This means we can express

$$\mathbf{\Sigma_x} = \begin{bmatrix} \mathbf{\Lambda}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_R \end{bmatrix} \tag{18}$$

where $\mathbf{\Lambda}_K = \text{diag}(\lambda_1, \ldots, \lambda_K)$ is the diagonal of K *principal* eigenvalues in the original domain, where $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_K$, and $\mathbf{\Lambda}_R$ is the diagonal sub-matrix of the residual eigenvalues. Assume that the residual eigenvalues are sufficiently smaller compared with the principal eigenvalues, and denote the first K columns of the projection matrix $\mathbf{\Phi}$ as $\mathbf{\Phi}_K$, then the compressed covariance matrix can be written as

$$\mathbf{\Sigma_y} = \mathbf{\Phi}_K \mathbf{\Lambda}_K \mathbf{\Phi}_K^T. \tag{19}$$

We now shall show that $\mathbf{\Sigma_y}$ also has a matching principal subspace in a sense that the K principal eigenvalues of $\mathbf{\Sigma_y}$, which we denote by $\xi_1, \ldots, \xi_K$ are close to the K principal eigenvalues of $\mathbf{\Sigma_x}$, or equivalently $\mathbf{\Lambda}_K$, while the rest is small. To do so we define the intermediate covariance matrix

$$\begin{aligned} \mathbf{\Sigma_z} &= \mathbf{\Phi}_K^T \mathbf{\Sigma_y} \mathbf{\Phi}_K \\ &= \mathbf{\Phi}_K^T \mathbf{\Phi}_K \;\; \mathbf{\Lambda}_K \;\; \mathbf{\Phi}_K^T \mathbf{\Phi}_K. \end{aligned} \tag{20}$$

Denote as $\kappa_1, \kappa_2, \ldots, \kappa_K$ the eigenvalues of $\mathbf{\Sigma_z}$. Our strategy is first to show that $|\xi_i - \kappa_i|, i = 1, \ldots, K$ are small, and $|\lambda_i - \kappa_i|, i = 1, \ldots, K$ are also small. Then we deduce the bound on $|\xi_i - \lambda_i|, i = 1, \ldots, K$.
The mathematical foundation of our proof consists of

– The concentration bound of the spectral norm of Gausian random matrices (see see [20] [10] for example). It follows from the theory that for the Gaussian random matrix $\mathbf{\Phi}_K$ of size $M \times K$ where each entry follows $\mathcal{N}(0, 1/\sqrt{M})$, the extreme singular values satisfy for some $t > 0$:

$$\Pr(\sigma_{\max}(\mathbf{\Phi}_K)) < 1 + \sqrt{K/M} + t) \geq 1 - e^{-M\frac{t^2}{2}}, \tag{21}$$

$$\Pr(\sigma_{\min}(\mathbf{\Phi}_K)) > 1 - \sqrt{K/M} - t) \geq 1 - e^{-M\frac{t^2}{2}}. \tag{22}$$

Let $\delta = e^{-N\frac{t^2}{2}}$ or equivalently $t = \sqrt{2\ln(1/\delta)/N}$, then the following hold with a probability of at least $1 - \delta$,

$$\sigma_{\max}(\mathbf{\Phi}_K) \leq 1 + \sqrt{K/M} + \sqrt{2\ln(1/\delta)/M}, \tag{23}$$

$$\sigma_{\min}(\mathbf{\Phi}_K) \geq 1 - \sqrt{K/M} - \sqrt{2\ln(1/\delta)/M}. \tag{24}$$

Denote as $\Delta = \sqrt{K/M} + \sqrt{2\ln(1/\delta)/M}$, then the concentration bound implies that the Gaussian random matrix $\mathbf{\Phi}_K$ is approximately unitary in a sense that the singular values are close to 1 when $K \ll M$, with a probability of at least $1 - \delta$

$$1 - \Delta \leq \sigma_{\min}(\mathbf{\Phi}_K) \leq \sigma_{\max}(\mathbf{\Phi}_K) \leq 1 + \Delta. \tag{25}$$

It is also useful to note that when $K \ll M$, the variation Δ is sufficiently small and thus we can deduce

$$1 - 2\Delta \leq \sigma_{\min}(\mathbf{\Phi}_K^T\mathbf{\Phi}_K) \leq \sigma_{\max}(\mathbf{\Phi}_K^T\mathbf{\Phi}_K) \leq 1 + 2\Delta. \tag{26}$$

– The approximate invariant subspace Theorem 8.1.11 in [22]. This theorem governs the bound on the sigular values of a covariance matrix when projected from high to low using an approximate unitary transformation. Suppose $\mathbf{\Sigma}_1 \in \mathbb{R}^{n\times n}$ is a symmetric matrix, and $\mathbf{T} \in \mathbb{R}^{n\times k}$ is an approximate unitary transformation matrix. Then the k largest singular values of $\mathbf{\Sigma}_2 = \mathbf{T}^T\mathbf{\Sigma}_1\mathbf{T}$ are close to those of $\mathbf{\Sigma}_1$ by the following

$$|\sigma_i(\mathbf{\Sigma}_2) - \sigma_i(\mathbf{\Sigma}_1)| \leq \sqrt{2}\left(\frac{\|\mathbf{\Sigma}_1\mathbf{T} - \mathbf{T}\mathbf{\Sigma}_2\|_2}{\sigma_k(\mathbf{T})} + \|\mathbf{T}^T\mathbf{T} - \mathbf{I}_k\|_2\|\mathbf{\Sigma}_1\|_2\right), \tag{27}$$

where $\| \bullet \|_2$ denotes the matrix norm.

With the above results, we are now ready to obtain the bounds as follows.
For $\mathbf{\Sigma_y}$ *and* $\mathbf{\Sigma_z} = \mathbf{\Phi}_K^T\mathbf{\Sigma_y}\mathbf{\Phi}_K$:

$$|\xi_i - \kappa_i| \leq \sqrt{2}\left(\frac{\|\mathbf{\Sigma_y}\mathbf{\Phi}_K - \mathbf{\Phi}_K\mathbf{\Sigma_z}\|_2}{\sigma_K(\mathbf{\Phi}_K)} + \|\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{I}_K\|_2\|\mathbf{\Sigma_y}\|_2\right). \tag{28}$$

Using (25) and (25), we bound each term with a probability of at least $1-\delta$ as follows

$$\begin{aligned}
\|\mathbf{\Sigma_y}\mathbf{\Phi}_K - \mathbf{\Phi}_K\mathbf{\Sigma_z}\|_2 &= \|\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{\Phi}_K\mathbf{\Phi}_K^T\mathbf{\Sigma_y}\mathbf{\Phi}_K\|_2 && (29)\\
&= \|\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{\Phi}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K\|_2 && (30)\\
&= \|\mathbf{\Phi}_K(\mathbf{I}_K - \mathbf{\Phi}_K^T\mathbf{\Phi}_K)\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K\|_2 && (31)\\
&\le \|\mathbf{\Phi}_K\|_2\|(\mathbf{I}_K - \mathbf{\Phi}_K^T\mathbf{\Phi}_K)\|_2\|\mathbf{\Lambda}_K\|_2\|\mathbf{\Phi}_K^T\mathbf{\Phi}_K\|_2 && (32)\\
&\le (1+\Delta)\times 2\Delta\times\lambda_1\times(1+2\Delta)\approx 2\Delta\lambda_1, && (33)\\
\|\mathbf{\Phi}_K^T\mathbf{\Phi}_K - \mathbf{I}_K\|_2 &\le 2\ \Delta, && (34)\\
\|\mathbf{\Sigma_y}\|_2 &= \|\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\|_2 && (35)\\
&\le \|\mathbf{\Phi}_K\|_2\|\mathbf{\Lambda}_K\|_2\|\mathbf{\Phi}_K^T\|_2 && (36)\\
&\le (1+\Delta)\times\lambda_1\times(1+\Delta)\approx(1+2\Delta)\lambda_1, && (37)\\
\|\sigma_K(\mathbf{\Phi}_K)\|_2 &\ge 1-\Delta. && (38)
\end{aligned}$$

Using these bounds, it follows that with a probability of at least $1-\delta$

$$|\xi_i - \kappa_i| \le \sqrt{2}\left(\frac{2\Delta\lambda_1}{1-\Delta} + 2\Delta\lambda_1\right) \approx 4\sqrt{2}\Delta\lambda_1, \quad i=1,2,\ldots,K. \tag{39}$$

For $\mathbf{\Sigma_x} = \mathbf{\Lambda}_K$ *and* $\mathbf{\Sigma_z} = \mathbf{\Phi}_K^T\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K$: Let $\mathbf{T} = \mathbf{\Phi}_K^T\mathbf{\Phi}_K$ then

$$|\lambda_i - \kappa_i| \le \sqrt{2}\left(\frac{\|\mathbf{\Lambda}_K\mathbf{T} - \mathbf{T}\mathbf{\Sigma_z}\|_2}{\sigma_K(\mathbf{T})} + \|\mathbf{T}^T\mathbf{T} - \mathbf{I}_K\|_2\|\mathbf{\Lambda}_K\|_2\right). \tag{40}$$

Again, we bound each term with a probability of at least $1-\delta$ as follows (note that $\mathbf{T}^T = \mathbf{T}$)

$$\begin{aligned}
\|\mathbf{\Lambda}_K\mathbf{T} - \mathbf{T}\mathbf{\Sigma_z}\|_2 &= \|\mathbf{\Lambda}_K\mathbf{T} - \mathbf{T}\mathbf{T}\mathbf{\Lambda}_K\mathbf{T}\|_2 && (41)\\
&= \|(\mathbf{I}_K - \mathbf{T}^T\mathbf{T})\mathbf{\Lambda}_K\mathbf{T}\|_2 && (42)\\
&\le \|(\mathbf{I}_K - \mathbf{T}^T\mathbf{T})\|_2\|\mathbf{\Lambda}_K\|_2\|\mathbf{T}\|_2 && (43)\\
&\le 4\Delta\times\lambda_1\times(1+2\Delta)\approx 4\Delta\lambda_1, && (44)\\
\|\sigma_K(\mathbf{T})\|_2 &\ge 1-2\Delta, && (45)\\
\|\mathbf{T}^T\mathbf{T} - \mathbf{I}_K\|_2 &\le 4\Delta. && (46)
\end{aligned}$$

Using these bounds, it follows that with a probability of at least $1-\delta$

$$|\lambda_i - \kappa_i| \le \sqrt{2}\left(\frac{4\Delta\lambda_1}{1-2\Delta} + 4\Delta\lambda_1\right) \approx 8\sqrt{2}\Delta\lambda_1, \quad i=1,2,\ldots,K. \tag{47}$$

Thus, from (39) and (47) we can deduce the bound on $|\lambda_i - \xi_i|, i=1,\ldots,K$. We note that a triangle inequality immediately gives

$$|\lambda_i - \xi_i| \le |\lambda_i - \kappa_i| + |\xi_i - \kappa_i|. \tag{48}$$

However, such bound can be still improved. This is because we note that if we take $\mathbf{\Lambda}_K$ as a reference, then the singular values of $\mathbf{\Sigma_y}$ are also the singular values of $\mathbf{\Phi}_K^T\begin{bmatrix}\mathbf{\Lambda}_K & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{\Phi}_K$. Thus, every time we perform an approximate orthogonal projection by $\mathbf{\Phi}_K$, the singular values are moved further away from those of the original matrix. Also, due to the construction of $\mathbf{\Sigma_z} = \mathbf{\Phi}_K^T\mathbf{\Sigma_y}\mathbf{\Phi}_K = \mathbf{\Phi}_K^T\mathbf{\Phi}_K\mathbf{\Lambda}_K\mathbf{\Phi}_K^T\mathbf{\Phi}_K$, we conclude that the singular values of

$\mathbf{\Sigma_z}$ even move further away from those of $\mathbf{\Lambda}_K$. This can be obviously seen with the largest singular values, where we have shown that

$$\xi_1 \leq (1+2\Delta)\lambda_1, \tag{49}$$
$$\kappa_1 \leq (1+2\Delta)\xi_1 \leq (1+4\Delta)\lambda_1. \tag{50}$$

Thus, every time an approximate orthogonal transformation is applied, the bound on the singular values becomes larger. This implies that the tightest bound on the singular values of $|\lambda_i - \xi_i|$ can be obtained by the difference between the bound on $|\lambda_i - \kappa_i|$ and $|\xi_i - \kappa_i|$. It follows that with a probability of at least $1-\delta$

$$|\lambda_i - \xi_i| \leq 4\sqrt{2}\left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2\ln(1/\delta)}{M}}\right) \quad i = 1, 2, \ldots, K. \tag{51}$$

Proof of Theorem 2 (Bound on false alarm rate)

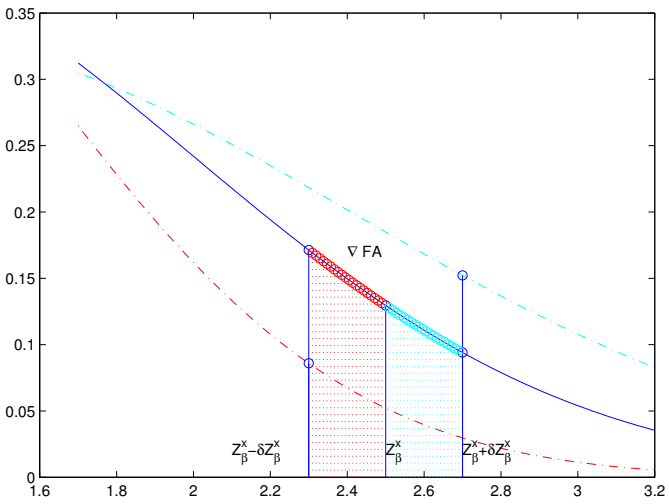


Fig. 28 Tail behavior for complete data and compressed data. The solid curve represents the tail of the distribution of the normalized statistic when complete data is used. The two dashed curved represent two possible tails of the distribution of the normalized statistic but when compressed data is used instead. The shaded areas represent the change in the false alarm in the two cases.

The residual statistics has normal distribution and the false alarm rate depends strongly on the tail. The previous proof has shown that the principal eigenvalues experience a small change under a random projection. Next, we show that there is also a small deviation in the false alarm rate. Our strategy is based on perturbation analysis of the tail of the distribution of the decision statistic. Figure 28 illustrates the tail behavior in the original and compressed domains. Here, Z_β^X denotes the normalized statistic in the original domain. The false alarm

for original data is the area under the distribution curve from z_β^X to ∞. Due to compression, suppose that the normalized statistic associated with the compressed data moves by $z_\beta^X \pm \delta z_\beta^X$. Thus, the change in the false alarm can be calculated as the change of the tail area, which gives

$$\Delta \Pr(FA) = \mathcal{N}(z_\beta^x)\delta z_\beta^x. \tag{52}$$

In what follows, we use the results of the previous proof to evaluate such changes in the false alarm. We note in the above expression, $\mathcal{N}(z_\beta^x)$ is the value of the normal distribution at z_β^x and is assumed known for a given desired false alarm β. For example, with a desired false alarm of 1%, this is approximately 0.0267. Then it remains to compute the change δz_β^x due to random projection to obtain compressed data.

We start with the result in Section 3 of [24], which states that the residual statistics Q has a normal distribution as follows

$$\left(\frac{Q}{\theta_1}\right)^{h_0} = Z \sim \mathcal{N}(\mu, \sigma), \tag{53}$$

where

$$\mu = 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2}, \;\; \sigma = \frac{2\theta_2 h_0^2}{\theta_1^2} \tag{54}$$

and $\theta_1 = \sum_{i=K+1}^N \lambda_i, \;, \theta_2 = \sum_{i=K+1}^N \lambda_i^2, \; \theta_3 = \sum_{i=K+1}^N \lambda_i^3, \; h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$ and K being the number of principal components.

Denote as $z_\beta^x, \mu_x, \sigma_x$ and $z_\beta^y, \mu_y, \sigma_y$ the detection threshold, mean, and standard deviation of the respected distributions in the original and compressed domain, and $C_\beta = 1 - \beta$ as the percentile of the normal distribution with an desired false alarm β. Here,

$$\begin{aligned} \mu_x = 1 + \frac{\theta_2^x h_0^x(h_0^x-1)}{(\theta_1^x)^2}, \;\; \mu_y = 1 + \frac{\theta_2^y h_0^y(h_0^y-1)}{(\theta_1^y)^2}, \\ \sigma_x = \frac{2\theta_2^x (h_0^x)^2}{(\theta_1^x)^2}, \;\; \sigma_y = \frac{2\theta_2^y (h_0^y)^2}{(\theta_1^y)^2} \end{aligned} \tag{55}$$

Then it follows that

$$\frac{z_\beta^x - \mu_x}{\sigma_x} = \frac{z_\beta^y - \mu_y}{\sigma_y} = C_\beta. \tag{56}$$

By writing $Z_\beta^x - Z_\beta^y = (\mu_x - \mu_y) + C_\beta(\sigma_x - \sigma_y)$ and applying the triangle inequality, we obtain

$$\delta z_\beta^x = |Z_\beta^x - Z_\beta^y| \le |(\mu_x - \mu_y)| + C_\beta|(\sigma_x - \sigma_y)|. \tag{57}$$

We next bound each term in (57). Once again, we use perturbation analysis by considering the functions

$$f_\mu(\theta_1, \theta_2, \theta_3) = 1 + \frac{\theta_2 h_0(h_0-1)}{\theta_1^2} = 1 - \frac{\theta_2}{\theta_1^2}\left(1 - \frac{2\theta_1\theta_3}{3\theta_2^2}\right)\frac{2\theta_1\theta_3}{3\theta_2^2} \tag{58}$$

$$= 1 - \frac{2\theta_3}{3\theta_1\theta_2} + \frac{4\theta_3^2}{9\theta_2^3}, \tag{59}$$

$$f_\sigma(\theta_1, \theta_2, \theta_3) = \frac{2\theta_2 h_0^2}{\theta_1^2} = \frac{2\theta_2}{\theta_1^2}\left(1 - \frac{2\theta_1\theta_3}{3\theta_2^2}\right)^2 \tag{60}$$

$$= \frac{2\theta_2}{\theta_1^2} + \frac{8\theta_3^2}{9\theta_2^3} - \frac{8\theta_3}{3\theta_1\theta_2}. \tag{61}$$

These functions allow the computation of the changes in the mean and standard deviation. For example,

$$|(\mu_x - \mu_y)| \approx \sum_{i=1}^{3} \left| \frac{\partial f_\mu(\theta_1, \theta_2, \theta_3)}{\partial \theta_i} \right| |\theta_i^x - \theta_i^y|. \tag{62}$$

We note that the partial derivatives are straightforward, so it remains to derive $|\theta_i^x - \theta_i^y|$. To do so, we use the result in [24] that

$$\theta_i^x = \mathsf{tr}(\mathbf{\Sigma_x})^i - \sum_{j=1}^{K} (\lambda_j^x)^i \;\; i = 1, 2, 3. \tag{63}$$

$$\theta_i^y = \mathsf{tr}(\mathbf{\Sigma_y})^i - \sum_{j=1}^{K} (\xi_j^y)^i \;\; i = 1, 2, 3. \tag{64}$$

From (63) and (64), we obtain

$$\delta_1 = \theta_1^x - \theta_1^y = (\mathsf{tr}(\mathbf{\Sigma_x}) - \mathsf{tr}(\mathbf{\Sigma_y})) + \sum_{j=1}^{K} ((\lambda_j^x) - (\xi_j^y)). \tag{65}$$

We now attempt to bound each RHS term of (65). First, for the trace terms we recall that

$$\mathsf{tr}(\mathbf{\Sigma_x}) - \mathsf{tr}(\mathbf{\Sigma_y}) = \mathsf{tr}(\mathbf{\Sigma_x}) - \mathsf{tr}(\mathbf{\Phi \Sigma_x \Phi}^T) \tag{66}$$

From the previous remark, we can assume $\mathbf{\Sigma_x}$ is a diagonal matrix. $\mathbf{\Sigma_x} = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$. Let, ϕ_i be the i^{th} column of the matrix $\mathbf{\Phi}^T$, then

$$\mathsf{tr}(\mathbf{\Phi \Sigma_x \Phi^T}) = \mathsf{tr}(\sum_{i=1}^{M} (\sum_{j=1}^{N} \phi_{ij}^2 \lambda_j)) = \mathsf{tr}(\sum_{j=1}^{N} \left[\lambda_j (\sum_{i=1}^{M} \phi_{ij}^2) \right]) \tag{67}$$

Since, each column of the matrix $\mathbf{\Phi}$ is normalized to unity then,

$$\mathsf{tr}(\mathbf{\Phi \Sigma_x \Phi^T}) = \mathsf{tr}(\mathbf{\Sigma_x}). \tag{68}$$

Thus, the trace term in (65) is zero. Meanwhile, it follows from (51) with a probability of at least $1 - \delta$ that

$$|\theta_1^x - \theta_1^y| \leq K 4\sqrt{2} \Delta \lambda_1, \tag{69}$$

where $\Delta = \sqrt{K/M} + \sqrt{2\ln(1/\delta)/M}$ is a small number. Similarly, for the second- and third-order terms

$$|\theta_2^x - \theta_2^y| \leq K 4\sqrt{2} \times 2\Delta \lambda_1^2 = 8\sqrt{2} K \lambda_1^2 \Delta, \tag{70}$$

$$|\theta_3^x - \theta_3^y| \leq K 4\sqrt{2} \times 3\Delta \lambda_1^3 = 12\sqrt{2} K \lambda_1^3 \Delta. \tag{71}$$

Thus, we can bound the variation in the mean as follows

$$|\mu_x - \mu_y| \leq 4\sqrt{2} K \lambda_1 \left(\left| \frac{\partial f_\mu}{\partial \theta_1} \right| + 2 \left| \frac{\partial f_\mu}{\partial \theta_2} \right| \lambda_1 + 3 \left| \frac{\partial f_\mu}{\partial \theta_3} \right| \lambda_1^2 \right) \Delta, \tag{72}$$

where the partial derivatives are

$$\frac{\partial f_\mu}{\partial \theta_1} = \frac{2\theta_3}{3\theta_1^2 \theta_2} \tag{73}$$

$$\frac{\partial f_\mu}{\partial \theta_2} = \frac{2\theta_3}{3\theta_1\theta_2^2} - \frac{4\theta_3^2}{3\theta_2^4} \tag{74}$$

$$\frac{\partial f_\mu}{\partial \theta_3} = \frac{2}{3\theta_1\theta_2} - \frac{8\theta_3}{9\theta_2^3} \tag{75}$$

Similarly,

$$|\sigma_x - \sigma_y| \leq 4\sqrt{2}K\lambda_1 \left(\left| \frac{\partial f_\sigma}{\partial \theta_1} \right| + 2 \left| \frac{\partial f_\sigma}{\partial \theta_2} \right| \lambda_1 + 3 \left| \frac{\partial f_\sigma}{\partial \theta_3} \right| \lambda_1^2 \right) \Delta. \tag{76}$$

where the partial derivatives are

$$\frac{\partial f_\sigma}{\partial \theta_1} = -\frac{4\theta_2}{\theta_1^3} + \frac{8\theta_3}{3\theta_1^2\theta_2} \tag{77}$$

$$\frac{\partial f_\sigma}{\partial \theta_2} = \frac{2}{\theta_1} - \frac{8\theta_3^2}{3\theta_2^4} + \frac{8\theta_3}{3\theta_1\theta_2^2} \tag{78}$$

$$\frac{\partial f_\sigma}{\partial \theta_3} = \frac{16\theta_3}{9\theta_2^3} - \frac{8}{3\theta_1\theta_2}. \tag{79}$$

As θ_i, C_β, K are constants when we study the bound on the change in the false alarm as a function of M, all the derivatives are constant. The above results indicate that the change in the false alarm depends linearly on Δ, which implies with a probability of at least $1-\delta$:

$$\Delta \Pr(FA) = \mathcal{O}\left(\sqrt{\frac{K}{M}} + \sqrt{\frac{2\ln(1/\delta)}{M}} \right). \tag{80}$$

Preservation of Gaussianity under Unitary Transformation

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

Lemma 1 *Suppose that $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$ is an iid random matrix whose entries follow a zero-mean Gaussian distribution with variance σ^2. Let $\mathbf{U} \in \mathbb{R}^{N\times N}$ be a unitary matrix. Then $\mathbf{\Phi}' = \mathbf{\Phi}\mathbf{U}$ is also an iid Gaussian random matrix with the same variance σ^2.*

First we prove that $\mathsf{E}[\phi'_{ij}] = \mathsf{E}[\phi_{ij}] = 0$ and $\mathsf{Var}[\phi'_{ij}] = \sigma^2$. We start from $\phi'_{ij} = \sum_{k=1}^n \phi_{ik}u_{kj}$. Thus $\mathsf{E}[\phi'_{ij}] = \sum_{k=1}^N \mathsf{E}[\phi_{ik}]u_{kj} = 0$, whilst due to iid assumption

$$\mathsf{Var}[\phi'_{ij}] = \sum_{k=1}^N \mathsf{Var}[\phi_{ik}]u_{kj} = \sum_{k=1}^N \sigma^2 u_{kj}^2 = \sigma^2 \sum_{k=1}^n u_{kj}^2 = \sigma^2. \tag{81}$$

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

$$\mathsf{E}[\phi'_{ij}\phi_{mn}] = \mathsf{E}\left[\sum_{k=1}^N \phi_{ik}u_{kj} \sum_{k'=1}^N \phi_{mk'}u_{k'n} \right] \tag{82}$$

$$= \sum_{k=1}^N \sum_{k'=1}^N \mathsf{E}[\phi_{mk'}\phi_{ik}]u_{k'n}u_{kj} \tag{83}$$

$$= 0. \tag{84}$$

References

1. http://www.abilene.iu.edu/
2. http://www.cvg.rdg.ac.uk/pets2007/data.html/
3. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 274–281. ACM, New York, NY, USA (2001). DOI http://doi.acm.org/10.1145/375551.375608
4. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. pp. 555–560 (2008)
5. Aggarwal, C.: On abnormality detection in spuriously populated data streams. In: Proceedings of the IEEE International Conference on Data Mining (ICDM) (2005)
6. Barnett, V., Lewis, T.: Outliers in statistical data. Chichester, New York (1984)
7. Bingham, E., Mannila, H.: Random projection in dimensionality reduction: applications to image and text data. In: Proc. KDD, pp. 245–250. ACM (2001)
8. Brand, M., Oliver, N., Pentland, A.: Coupled Hidden Markov Models for Complex Action Recognition. In: IEEE CVPR, pp. 994–999 (1997)
9. Budhaditya, S., Pham, D., Lazarescu, M., Venkatesh, S.: Effective anomaly detection in sensor networks data streams. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 722–727 (2009)
10. Candes, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Info. Theory **52**(2), 489–509 (2006)
11. Candes, E., Tao, T.: Near optimal signal recovery from random projections: Universal encoding strategies. IEEE Trans. Info. Theory (2006)
12. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (2009)
13. Chatzigiannakis, V., Papavassiliou, S., Grammatikou, M., Maglaris, B.: Hierarchical anomaly detection in distributed large-scale sensor networks. In: Proceedings of the 11th IEEE Symposium on Computers and Communications (ISCC), pp. 761–767. Washington, DC, USA (2006)
14. Donoho, D.: Compressed sensing. In: IEEE Trans. Info. Theory, vol. 52, pp. 1289–1306 (2006)
15. Drineas, P., Frieze, A., Kannan, R., Vempala, S., Vinay, V.: Clustering Large Graphs via the Singular Value Decomposition. Machine Learning **56**(1), 9–33 (2004)
16. Drineas, P., Kannan, R., Mahoney, M.: Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. SIAM Journal of Computing **36**(1), 158 (2006)
17. Elad, M.: Optimized projections for compressed sensing. IEEE Trans. Sig. Proc. **55**, 5695–5702 (Dec. 2007)
18. Fowler, J.: Compressive-projection principal component analysis and the first eigenvector. In: Data Compression Conference, 2009. DCC'09., pp. 223–232. IEEE (2009)
19. Fujimaki, R.: Anomaly detection support vector machine and its application to fault diagnosis. In: Proceedings of the IEEE International Conference on Data Mining (ICDM), pp. 797–802. Washington, DC, USA (2008)
20. Geman, S.: A limit theorem for the norm of random matrices. Ann. Probab. **8**, 252–261 (1980)
21. Giatrakos, N., Kotidis, Y., Deligiannakis, A., Vassalos, V., Theodoridis, Y.: Taco: tunable approximate computation of outliers in wireless sensor networks. In: Proceedings of the 2010 international conference on Management of data, pp. 279–290. ACM (2010)
22. Golub, Loan, V.: Matrix computations (3rd ed.). Johns Hopkins University Press, Baltimore, MD, USA (1996)
23. Huang, L., Nguyen, X., Garofalakis, M., Jordan, M., Joseph, A., Taft, N.: In-Network PCA and Anomaly Detection. In: Proc. NIPS, pp. 617–624 (2007)
24. Jackson, E., Mudholkar, G.: Control procedures for residuals associated with principal component analysis. Technometrics **21**(3), 341–349 (1979)
25. Jackson, J.: Quality control methods for several related variables. Technometrics pp. 359–377 (1959)
26. Jackson, J.: Principal components and factor analysis. I- Principal components. Journal of Quality Technology **12**, 201–213 (1980)
27. Janakiram, D., Reddy, V., Kumar, A.: Outlier detection in wireless sensor networks using Bayesian belief networks. In: Proceedings of the First International Conference on Communication System Software and Middleware. (2006)
28. Jiang, X., Cooper, G.: A real-time temporal bayesian architecture for event surveillance and its application to patient-specific multiple disease outbreak detection. Data Mining and Knowledge Discovery **20**(3), 328–360 (2010)
29. Koufakou, A., Georgiopoulos, M.: A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes. Data Mining and Knowledge Discovery **20**(2), 259–289 (2010)

30. Lakhina, A., Crovella, M., Diot, C.: Diagonising network-wide traffic anomalies. In: Proc. ACM SIGCOMM (2004)
31. Li, W., Yue, H., Valle-Cervantes, S., Qin, S.: Recursive PCA for adaptive process monitoring. Journal of Process Control **10**(5), 471–486 (2000)
32. Liu, K., Kargupta, H., Ryan, J.: Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering **18**(1), 92–106 (2006)
33. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. IJCAI, vol. 81, pp. 674–679 (1981)
34. Medioni, G., Cohen, I., Brémond, F., Hongeng, S., Nevatia, R.: Event Detection and Analysis from Video Streams. IEEE Trans. Pattern Anal. Mach. Intell. pp. 873–889 (2001)
35. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79**(3), 299–318 (2008)
36. Noto, K., Brodley, C., Slonim, D.: Frac: a feature-modeling approach for semi-supervised and unsupervised anomaly detection. Data Mining and Knowledge Discovery pp. 1–25 (2011)
37. Phung, D., Duong, T., Venkatesh, S., Bui, H.: Topic transition detection using hierarchical hidden Markov and semi-Markov models. In: Proc. ACM-MM, pp. 11–20 (2005)
38. Rabbat, M., Haupt, J., Singh, A., Nowak, R.: Decentralized compression and predistribution via randomized gossiping. In: Proc. IPSN, pp. 51–59. New York, NY, USA (2006)
39. Strohmer, T., Heath, R.: Grassmannian frames with applications to coding and communication. Applied and Computational Harmonic Analysis **14** (May 2003)
40. Thottan, M., Ji, C.: Anomaly detection in IP networks. IEEE Transactions on Signal Processing **51**(8), 2191–2204 (2003)
41. Vempala, S.: The Random Projection Method. SIAM (2004)
42. Yan, J., Zhang, B., Liu, N., Yan, S., Cheng, Q., Fan, W., Yang, Q., Xi, W., Chen, Z.: Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE transactions on Knowledge and Data Engineering pp. 320–333 (2006)
43. Zhu, C., Kitagawa, H., Faloutsos, C.: Example-based robust outlier detection in high dimensional datasets. In: Proc. ICDM (2005)