

©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Effective Anomaly Detection in Sensor Networks Data Streams

Saha Budhaditya, Duc-Son Pham, Mihai Lazarescu, and Svetha Venkatesh
Department of Computing, Curtin University of Technology, Perth, Western Australia
Emails: b.saha@postgrad.curtin.edu.au, dspham@ieee.org

Abstract—This paper addresses a major challenge in data mining applications where the full information about the underlying processes, such as sensor networks or large on-line database, cannot be practically obtained due to physical limitations such as low bandwidth or memory, storage, or computing power. Motivated by the recent theory on direct information sampling called compressed sensing (CS), we propose a framework for detecting anomalies from these large-scale data mining applications where the full information is not practically possible to obtain. Exploiting the fact that the intrinsic dimension of the data in these applications are typically small relative to the raw dimension and the fact that compressed sensing is capable of capturing most information with few measurements, our work show that spectral methods that used for volume anomaly detection can be directly applied to the CS data with guarantee on performance. Our theoretical contributions are supported by extensive experimental results on large datasets which show satisfactory performance.

Keywords—stream data processing, anomaly detection, spectral methods, residual analysis, compressed sensing

I. INTRODUCTION

Anomaly detection in data streams in large-scale sensor networks is of great research interest. A number of methods which work on either databases or datastreams have been proposed to detect different types of anomalies and the methods incorporate different approaches for example Bayesian networks [13], SVM [9], K -nearest neighborhood [4], clustering [20], and spectral methods [3], [14] (see also [6] and references therein).

However, these techniques fundamentally assume that the complete data is available with sufficient storage and computing power. As the network size increases, it becomes increasingly difficult to acquire all the data streams for processing [11]. Hence, in large-scale networks, complete data information may not be always available at the fusion point either because of low bandwidth or large geometrical distance between sensors [2]. The particular constraints on large-scale networks imply that only *partial* information about the whole network can possibly be *sensed*. This issue has been recently tackled by some recent approaches including *decentralization* [10] or *column sampling* [8]. However, there are still inherent limitations with these approaches. For example, in the decentralization approach, there is still a likelihood that the communication overhead exceeds the bandwidth and the fact that the central node needs to store

a data matrix of the same size as \mathbf{X} . On the other hand, the column sampling approach is not suitable for on-line applications.

Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L]$ denote the network data matrix in which $\mathbf{x}_i \in \mathbb{R}^N$ is a N -dimensional vector representing the status of N nodes at time instance i . A central monitor collects information from these distributed nodes to make a decision, and two cases can be considered:

Case 1: It is difficult to collect *all rows* of the data matrix \mathbf{X} centrally and it is only feasible to sample M out of the total of N sensors. This data from M sensors is streamed continuously to the central point.

Case 2: It is difficult to collect *all columns* of the data matrix \mathbf{X} centrally. This is equivalent to sampling only a subset of the temporal domain, and thus 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 available fully (retrospectively), 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. .

We propose a new framework for the detection of anomalies in large-scale sensor networks to address the above incomplete data challenge arising from physical constraints. The framework is based on the recently developed compressed sensing theory (CS) [7] utilizing its implicit *direct* information sampling capability (detailed in section II-B). We formalise how CS can be used to effectively acquire the data 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, which is then used for anomaly detection. We first formulate two important theoretical results: (1) a theoretical bound which establishes that the principal sub-space is preserved in both the raw and CS domains with high probability and (2) a theoretical bound for the false alarm rate in anomaly detection using this spectral technique. To demonstrate the flexibility of the CS paradigm, experimental results are performed on two datasets: a) a synthetic, large-scale network dataset to demonstrate anomaly detection when the data from sensors is sensed in a compressed way, that is only M of N sensors are sampled, and b) a large video dataset, in which the temporal dimension is sensed in a compressed way and

used for anomaly detection.

The significance of our contributions is the demonstration that spectral-based methods can be applied to CS data, and anomaly detection 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. More exactly, the computational complexity of the proposed method is *sublinear* ($\mathcal{O}(\log N)$ or $\mathcal{O}(\log L)$). The framework we present integrates anomaly detection and CS into a deployable paradigm to overcome the problems of anomaly detection with partial data, a reality in most real-world situations.

The paper is organized as follows. Section 2 describes the problem in detail and provides background information on CS and anomaly detection. Section 3 explains our proposed method and its analysis. Section 4 describes the data, experimental setup and results and the conclusions are covered in Section 5.

II. BACKGROUND

A. Residual Subspace Projection and Anomaly Detection

Let the complete network data matrix be denoted by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_L]$ where each data instance $\mathbf{x}_i \in \mathbb{R}^N$. The residual analysis method [12] seeks a decomposition of the observed data into principal subspace which is believed to govern the normal characteristics, and the residual subspace from which abnormal characteristics can be found. If \mathbf{X} is available, the residual method performs the eigenvalue decomposition of the sample covariance matrix $\Sigma_{\mathbf{x}}$ from which the K principal eigenvectors \mathbf{U} are obtained. The projection of any data instance \mathbf{x} onto the residual subspace is $\mathbf{z} = (\mathbf{I} - \mathbf{U}\mathbf{U}^T)\mathbf{x}$. Under the null hypothesis that the data is ‘normal’, the squared prediction error (SPE) *i.e.* $\|\mathbf{z}\|_2^2$ follows a non-central chi-square distribution. 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 -statistics, which is the function of non-principal eigenvalues in residual subspace, and can be approximated by

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

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 $i = 1, 2, 3$, $c_\beta = (1 - \beta)$ percentile in a standard normal distribution and Q_β , and $\lambda_j, i = 1, \dots, M$ are the eigenvalues of $\Sigma_{\mathbf{x}}$. An anomaly is detected when $\|\mathbf{z}\|_2^2 > Q_\beta$.

B. Compressed Sensing

Assume that a data vector $\mathbf{x} \in \mathbb{R}^N$ admits a linear representation by a set of orthonormal basis functions Ψ with coefficients α , *i.e.* $\mathbf{x} = \Psi\alpha$. Two cases of interest are i) *Sparse signal*: the signal \mathbf{x} is said to be K -sparse if only

K entries of α are nonzero; and ii) *Compressible signal*: the magnitudes of the coefficients α , when ordered, follow an exponential decay [5]. When \mathbf{x} is sparse or compressible, CS theory [5], [7] has proved that it is possible to ‘sense’ \mathbf{x} via a simple, non-adaptive and linear projection $\mathbf{y} = \Phi\mathbf{x}$. The sensing matrix $\Phi \in \mathbb{R}^{M \times N}$ has a significantly smaller number of rows than columns, *i.e.* $M \ll N$, meaning that the dimension of \mathbf{y} is considerably smaller than \mathbf{x} . Importantly, under suitable conditions on the approximate orthogonality between columns of the sensing matrix Φ , it is possible to perfectly recover \mathbf{x} from \mathbf{y} via a convex optimization problem

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

This implies that all the salient information about \mathbf{x} is captured in \mathbf{y} , making CS an universal tool for information preserving projection technique. When classification is needed instead of recovery, the use of CS is clearly an advantage as the number of processing samples is reduced to M (in practice, $M = \mathcal{O}(K \log N) \ll N$) [7]. The advantage of working in CS domain is that it overwhelmingly reduces the communication overhead and increases the scalability of the framework. For example, as the network data is *sparse*, only a small number of non-adaptive measurements M is needed to retain information about the main traffic. We refer interested readers to the CS repository for numerous background materials (<http://www.dsp.rice.edu/cs>).

Whilst the main focus of the CS community is on the recovery problem, *i.e.* to infer \mathbf{x} from \mathbf{y} , our focus here is on anomaly detection. Thus, as the information about \mathbf{x} is preserved in \mathbf{y} , we show subsequently that it is possible to directly detect anomalies from the compressed data \mathbf{y} .

III. PROPOSED FRAMEWORK

A. System setup

In the first step of the proposed framework, we obtain *compressed data acquisition using CS* $\mathbf{Y} \in \mathbb{R}^{N' \times L'}$. The reduction in either N' or L' depends on whether CS is deployed for reducing the feature dimension or time instances to meet the network constraints. We revisit the two cases considered previously:

Case 1: For the sensor sub-sampling case: We seek a linear transformation on the data $\mathbf{y} = \Phi\mathbf{x}$, where $\Phi \in \mathbb{R}^{M \times N}$ is known as the CS measurement matrix, whose entries are random variables. There are many matrices that can be efficiently implemented in practice such as the database friendly CS matrices [1] whose 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.

Case 2: For the temporal stream frames sub-sampling case: By using the CS theory and the CS matrix, the operator

can request the server to generate random number and select instances corresponding to the random values ± 1 , sum these two sets of instances, subtract them, and iteratively send such L' results to the operator where $L' \ll L$.

In the second step, we perform *anomaly detection using compressed measurements*. 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 detect anomalies.

Even though the framework appears simple and that CS or random projection has been well known and residual method is a standard method, the most important thing is to justify this simple scheme in a concrete manner, which is our main contribution. As shown in the following, the linearity of the CS acquisition, sparse spectral characteristic of the data, and the concentration properties of random projection are the main ingredients for the success of this simple framework.

B. Theoretical analysis

Our theoretical analysis is based on relative performance to the complete data \mathbf{X} . To do this, we first study the changes in the eigenvalues (spectral properties) reflected in the CS data as they are the important factor for detection as shown in (1). The bounds on the eigenvalues of CS data then allow us to study further the bound on false alarm rates when the residual subspace method is applied to the CS data in order to detect anomalies. It is noted that the proofs can be found at [18]. Here we only summarize the key results and discuss the implications.

1) *Case 1: M readings from N sensors*: The relation between the CS and complete data samples is $\mathbf{y}_i = \Phi \mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^M, i = 1 \dots L$. In this case $N' = M$ and $L' = L$. Denote the eigenvalues of the complete data \mathbf{X} as $\lambda_1, \dots, \lambda_N$ and those of the CS data \mathbf{Y} as $\xi_i, i = 1, \dots, M$. Denote as K the number principal eigenvalues in the complete data \mathbf{X} such that $K < M \ll N$. We assume that the CS matrix is a random Gaussian matrix. The following result shows that when the spectrum of the complete data is sparse, i.e. K is small relatively to N then the principal eigenvalues, i.e. the principal spectral characteristics, are preserved in the CS data.

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

$$|\lambda_i - \xi_i| \leq \sqrt{2}\lambda_1 \left(3\sqrt{\frac{K}{M}} + \sqrt{\frac{M}{N}} + 3\sqrt{\frac{2 \ln \frac{1}{\delta}}{M}} \right) \quad (3)$$

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

Remarks: A similar result on the bound of eigenvalues due to random projection is given in [19, Section 8.2]. However, it contains some parameters which are unclear. Furthermore, their result is not probabilistic which is the nature of random projections. Lastly, Lemma 8.4 in [19] only provides the

upper bound, whilst our result provides both upper and lower bounds using the theory of invariant subspaces.

The above theorem suggests that as the principal subspace spanned by \mathbf{X} is approximately preserved in CS domain with high probability, the *intrinsic* structure of the data in original input domain is unchanged under CS projection. The theorem is a direct consequence of the concentration property of Gaussian ensembles.

Now, we direct the discussion on the implication of this result on the anomaly detection on compressed data. From the previous discussion, we can clearly see that 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 retains noise-like behavior, i.e. has no salient spectral features. It can be easily shown that when the CS matrix Φ is normalized, which is the standard assumption in CS, 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. This means that as far as the statistic $t = \|\mathbf{z}\|^2$ is concerned, its distribution will also experience a small change when the CS data is used instead. This intuitive argument can be more formally stated by the following result, which forms the basis for our proposed framework for scalable anomaly detection in large sensor networks.

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

$$\Delta FA \leq \mathcal{O} \left(\sqrt{M/N} + \sqrt{2 \ln(1/\delta)/M} \right). \quad (4)$$

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 (4) becomes significantly small as the problem size, and thus M , becomes large. Therefore, for large-scale networks, the first term is dominant. CS theory states that in order to fully capture the information, the number of measurements M is related to the sparsity via $M = \mathcal{O}(K \log N)$. This implies that the first term will decay at the rate $\mathcal{O}(\sqrt{K \log(N)/N})$ and thus for large networks, this term is also very small if $K \ll N$. For volume anomalies, the intrinsic dimension has been observed to be consistent with this CS assumption [14].

2) *Case 2: Sub-sampling the number of data instances in temporal stream*: In Case 1, we have used the CS machinery to reduce the number of readings (i.e. rows) in data streaming applications such as sensor networks. 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 the CS machinery to compress each L -dimensional row of the complete data matrix \mathbf{X} into each M -dimensional row of the matrix \mathbf{Y} using a CS matrix $\Phi \in \mathbb{R}^{M \times L}$ where $M < L$. Mathematically, the relation

between this row-reduced version \mathbf{Y} and \mathbf{X} can be written as $\mathbf{Y}^T = \Phi \mathbf{X}^T$. In this case, $N' = N$ and $L' = M$. To see the analogy to the previous result, we start from the fact that $\lambda_i(\mathbf{X}\mathbf{X}^T) = \lambda_i(\mathbf{X}^T\mathbf{X})$, $i = 1, \dots, \min(N, L)$. 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, the previous result is readily applicable. The only minor difference is that N should be replaced by L as the reduction is performed on the row of \mathbf{X} . The changes in the principal eigenvalues are bounded by

$$|\lambda_i - \xi_i| \leq \sqrt{2}\lambda_1 \left(3\sqrt{\frac{K}{M}} + \sqrt{\frac{M}{L}} + 3\sqrt{\frac{2 \ln \frac{1}{\delta}}{M}} \right), \quad (5)$$

whilst the changes in the false alarm rate is bounded by

$$\Delta FA \leq \mathcal{O} \left(\sqrt{M/L} + \sqrt{(2 \ln 1/\delta)/M} \right). \quad (6)$$

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

3) *Complexity analysis*: If the complete data \mathbf{X} were available, the covariance matrix formation and eigenvalue decomposition in case of PCA requires computational power of $\mathcal{O}(N^3)$ and memory storage of $\mathcal{O}(N^2)$. In 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.

IV. EXPERIMENTAL RESULTS

A. Network Traffic Datasets

In this experiment, we consider anomaly detection in a large network traffic simulation [14] where the number of local monitors N ranges from 500 to 2000 and the number of time instances is $L = 2000$. The data is network traffic flow, which is the amount of traffic in between each pair of ingress and egress nodes in the network. The flow has two main characteristics, that is (i) a *normal* behaviors due to the usual traffic pattern (for example, daily demand fluctuation) and *abnormal* or *anomalous* behavior due to unexpected events like abnormal DNS transaction, network equipment failure, flash crowd occupancies, distributed denial of service (DDoS) attack etc. Specifically, this set of anomaly is called *volume* anomaly in the previous work [14] due to meaningful changes in traffic volume.

We set up the network simulation similar to that described in [14]. For the intrinsic network data, we selected DCT as a basis Ψ and the number of principal components is $K = 4$. The additive noise is Gaussian with $\sigma = 0.01$. To simulate abnormal network conditions we injected 70 anomalies of different magnitudes following the procedure specified

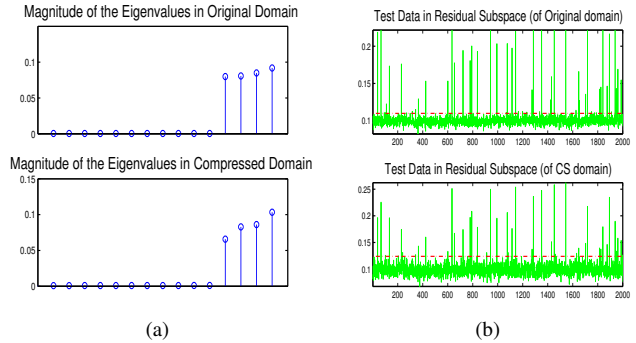


Figure 1. Eigenvalue plot for original and compressed data.

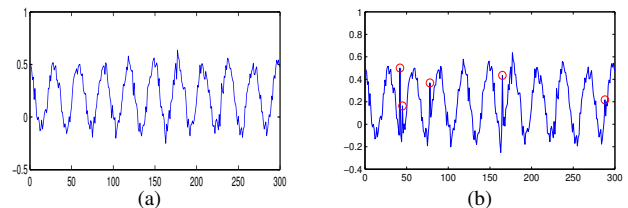


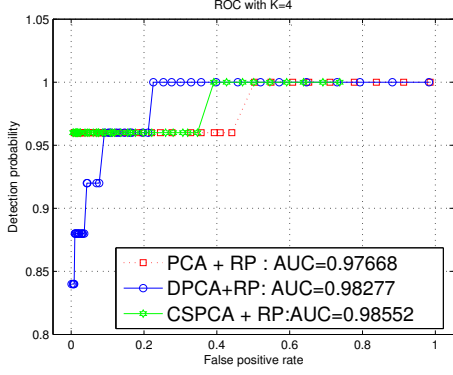
Figure 2. Typical normal (left) and abnormal (right) network link data.

in [14]. The number of CS measurements M is selected according to the CS guidelines, i.e. $M \sim \mathcal{O}(K \log N)$. The sensing matrices (Φ) were random Gaussian with a mutual coherence of 0.37, 0.33 and 0.20 for $N = 500, 100, 2000$ respectively.

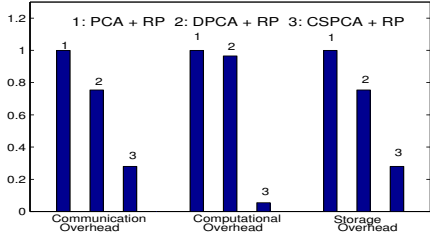
Fig. 1 shows eigenvalue distributions and the observations in residual subspace. This clearly illustrates the result of Theorem 1 for this network problem as the eigenvalues in the original and CS domains exhibit the same pattern.

Fig. 3a shows the receiver operating characteristics (ROC) plot of three different anomaly detection methods which are Lakhina's [14] PCA based residual projection (PCA + RP), Huang's [10] distributed PCA followed by RP (DPCA+RP) and our proposed PCA in CS domain followed by RP (CSPCA+RP). To further quantify this, we compared the plots ROC curves using (i) the area under the ROC curve (AUC) and (ii) equal error rate (EER) where the false positive being equal to false negative. An ideal classifier should achieve an AUC close to 1 and ERR small. The AUC/EER values were 0.976/0.09 for PCA + RP, 0.982/0.02 for DPCA + RP and 0.985/0.02 for CSPCA + RP. The results show that the performance of CSPCA+RP is very close to (even slightly better than) other methods. The reason for the more effective approximation comes from the reduction of the noise level in CS domain for high dimensional data and this leads to a better detection capability.

Fig. 3b compares communication, computation and storage overhead of three methods. From the Figure it can be observed that detecting anomalies in CS domain saves 45% to 60% communication bandwidth, 80% to 90% computa-



(a)



(b)

Figure 3. ROC and cost plots.

Table I
ANOMALY DETECTION PERFORMANCE ON SYNTHETIC DATA.

Metric	Time (seconds)		AUC		EER	
	PCA	CSPCA	PCA	CSPCA	PCA	CSPCA
N						
500	0.430	0.023	0.996	0.991	0.08	0.08
1000	3.364	0.097	0.976	0.985	0.09	0.02
2000	20.932	0.203	0.970	0.980	0.09	0.02

tional cost and 45% to 70% storage requirement with respect to either PCA+RP or DPCA+RP method. Furthermore, Table I provides a comparison of our proposed method and Lakhina’s residual projection (PCA+RP) method when the number of sensors in the network (N) varies from 500 to 2000. The results support our claim that the proposed approach performs equally well to PCA on the original domain.

B. Real-World Video Data

The second set of experiments were conducted on a very large video data stream set provided by the public transport authority. The video data totaling 83GB of compressed video was captured from the city’s central train station. The ground truth was independently verified by the transport authority and incidents ranged from loitering in the station tunnel to unusual behaviour involving infants.

For detecting anomalies in high-speed data streams, we use optical flow [15] as low-level features computed and aggregated over grid-based regions in the images In a similar

manner to the *bag-of-visual words* model for detecting human activity in the spatio-temporal domain [17], we construct the feature-frame matrix where we 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 aggregated over a window of length $w = 10$ seconds. Given a number of non overlapping moving windows L , the feature-frame matrix is defined as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_L]$, where $\mathbf{x}_k(i) = \sum_{l=(k-1)w}^{kw} x_i(l)$, where $k = 1 \dots L, i = 1 \dots N$. The normal activities includes heavy people traffic coming in through the entry point and going out by exit point during peak hours. It also includes few persons or almost no persons in off-peak hours. Any significant change in the motion volume statistics in the spatio-temporal domain would be treated as *unusual*.

We used video data captured at 25fps 570×720 resolution from two cameras in the corridors of the train station from 7AM to 11AM over a whole week. The training set $\mathbf{X}_{\text{Train}}$ is over five consecutive days where each day has 4 hours continuous video. For the testing set $\mathbf{X}_{\text{Train}}$, we used data from days 6th ($\mathbf{X}_{\text{Test1}}$) and 7th ($\mathbf{X}_{\text{Test2}}$). This results in $L = 7200$ and $N = 100$. We investigate the optimal choice for temporal subsampling of stream data by varying M and observe the FPR. The result is shown in Fig. 4 as M varies from 100 to 300. The optimal trade-off is found at $M = 220$ and we use this for subsequent experiments.

Next, we plot the eigenvalue distribution and observation in *residual* subspace for both the original input data (PCA + RP) and the CS data (CSPCA + RP) in Fig. 5. It shows that the energy seems to be concentrated in the 4 principal eigenvalues ($K = 4$). We then project the columns of $\mathbf{X}_{\text{Test1}}$ in the residual subspace as shown in Figs. 6(a) and 6(b). The threshold Q_β was computed in a similar way to the previous experiment with $\beta = 0.005$.

We detected two real anomalies out of three from the test data with the detected anomalies corresponding to (1) leaning and moving a small child against the wall (Fig. 7(a)) and (2) loitering (Fig. 7(b)). These anomalies are detected due to changes in the motion distributions of the cells which though local in nature, are clearly detectable in the residual subspace. The anomaly missed was due to the fact that it took place far away from the camera and as a result, it was difficult to detect because the motion features were not significant.

We repeated the same experiment with the second test set ($\mathbf{X}_{\text{Test2}}$) and detected the anomalous event “group loitering” (shown in Figure 7(c)) which occurred during “off-peak” hours.

V. CONCLUSION

In this paper we have presented a framework for detecting anomalies in data streams captured by large-scale sensor networks. The work addresses the key problem of dealing with incomplete data because of the physical constraints

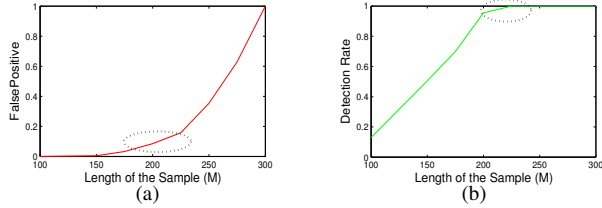


Figure 4. Optimal selection for M .

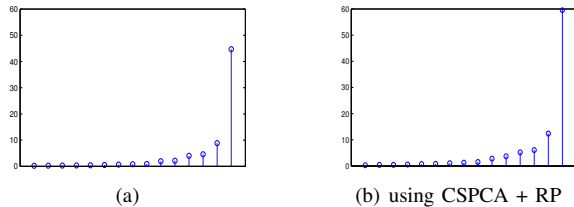


Figure 5. Eigenvalue plot of \mathbf{X}_{Train} data

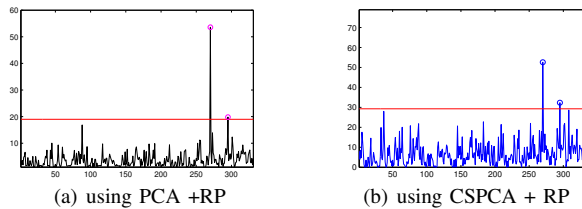
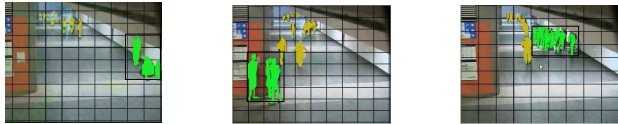


Figure 6. Residual plot of \mathbf{X}_{Test1} data



(a) Leaning on the wall (b) Hanging out in groups (c) Group Loitering

Figure 7. Anomaly detection in Public Surveillance Data

imposed by limited bandwidth available in large-scale networks. The framework is based on the CS theory and provides an effective solution for anomaly detection for both the case when the number of sensors or the number of data instances exceed 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 CS is capable of capturing most information with few measurements. We show that spectral methods used for

volume anomaly detection can be directly applied to the CS data with guarantees on performance and we demonstrate the effectiveness of the framework using both real and synthetic datasets.

REFERENCES

- [1] D. Achlioptas *et al.* Database-friendly random projections. In *Proc. PODS*, pp. 274–281, 2001.
- [2] IF. Akyildiz *et al.* A survey on sensor networks. *IEEE Communications Magazine*, pp. 102–114, 2002.
- [3] V. Barnett and T. Lewis. Outliers in statistical data. *Chichester, New York*, 1984.
- [4] M. Breunig *et al.* LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104, 2000.
- [5] E. Candes, A. Romberg and T. Tao. Near optimal signal recovery from random projections: Universal encoding strategies. *IEEE Trans. Info. Theory*, 2006.
- [6] V. Chandola *et al.* Anomaly detection: A survey. *ACM Computing Surveys*, 2009.
- [7] D. Donoho. Compressed sensing. *IEEE Trans. Info. Theory*, volume 52, pp. 1289–1306, 2006.
- [8] P. Drineas *et al.* Fast Monte Carlo Algorithms for Matrices II: Computing a Low-Rank Approximation to a Matrix. *SIAM Journal of Computing*, 36(1):158, 2006.
- [9] R. Fujimaki *et al.* Anomaly Detection Support Vector Machine and Its Application to Fault Diagnosis. In *Proc. ICDM*, pp. 797–802, 2008.
- [10] L. Huang *et al.* In-Network PCA and Anomaly Detection. In *Proc. NIPS*, pp. 617–624, 2007.
- [11] L. Huang *et al.* Communication-efficient on-line detection of network-wide anomalies. In *Proc. INFOCOM*, pp. 134–142, 2007.
- [12] E. Jackson and G. Mudholkar. Control procedures for residuals associated with principal component analysis. *Technometrics*, 21(3):341–349, 1979.
- [13] D. Janakiram *et al.* Outlier detection in wireless sensor networks using Bayesian belief networks. In *Proc. Communication System Software and Middleware*, pp. 1–6, 2006.
- [14] A. Lakhina *et al.* Diagonising network-wide traffic anomalies. In *Proc. ACM SIGCOMM*, 2004.
- [15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. IJCAI*, volume 81, pp. 674–679, 1981.
- [16] M. Mahoney and P. Chan. Learning rules for anomaly detection of hostile network traffic. In *Proc. ICDM*, pp. 601–604, 2003.
- [17] J. Niebles *et al.* Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 79(3):299–318, 2008.
- [18] D. Pham, B. Saha, M. Larazescu, and S. Venkatesh. Scalable Network-Wide Anomaly Detection Using Compressed Data. *Technical report*, Department of Computing, Curtin University of Technology, 2009. (available at www.impc.cs.curtin.edu.au/pubs/reports.php).
- [19] S. Vempala. *The Random Projection Method*. SIAM, 2004.
- [20] J. Yin *et al.* Clustering Distributed Time Series in Sensor Network. In *Proc. ICDM*, 678–687, 2008.