Low-Rank Block Sparse Decomposition Algorithm for Anomaly Detection in Networks

Masoumeh Azghani* and Sumei Sun†
* Sahand University of Technology, Tabriz, Iran
† Institute for Infocomm Research, Singapore
E-mail: mazghani@sut.ac.ir; sunsm@i2r.a-star.edu.sg

***Abstract*— In this paper, a method is suggested for the anomaly detection in wireless networks. The main problem that is addressed is to detect the malfunctioning sub-graphs in the network which bring about anomalies with block sparse structure. The proposed algorithm is detecting the anomalies considering the low-rank property of the data matrix and the block-sparsity of the outlier. Hence, the problem boils down to a compressed block sparse plus low rank decomposition that is solved with the aid of the ADMM technique. The simulation results indicate that the suggested method surpasses the other technique especially for higher block-sparsity rates.**

***Index Terms*—Compressed Sensing, Low rank minimization, Anomaly detection.**

I. Introduction

Anomalies may occur in the networks due to the hackers, node or link failures which would deteriorate the throughput of the network. Hence, the anomaly detection is essential to guarantee the acceptable performance of the networks. Different anomaly detection schemes have been suggested in the literature [1]. The statistical techniques include wavelet analysis [2], covariance matrix analysis [3], principle component analysis [4] and Kalman filtering [5]. Some discrete algorithms such as heavy hitter detection and heavy change detection have also been suggested for the anomaly detection problem. Another group of the techniques are based on machine learning schemes. Adaptive learning and thresholding scheme and clustering based method are the examples of such techniques. The most recent anomaly detection techniques are leveraging some properties of the signal such as sparsity and low-rank. A signal is said to be sparse if it has a very few number of non-zero entries. Most of the communicational and natural signals possess some kind of the sparsity property which is widely utilized in various applications [6]–[8]. A low-rank signal has very few number of non-zero singular values. This property can also play an important role in solving different problems ranging from image/video processing [9] to wireless communications and biomedical engineering [10]. Sparse coding for anomaly detection has been suggested in [11]. In [12], the robust matrix factorization has been exploited to detect the anomalies. Sparse approximation theory has been used in [13] for anomaly detection in smart grids. The low rank sparse decomposition is addressed in [14]. In [15], a reweighted low-rank and reweighted sparse technique is suggested to decompose the sparse and low-rank components. The proposed method in [16] solves the compressed low-rank sparse problem by minimizing the convex relaxation of the cost function. In this paper, the network anomaly detection is modeled as the compressed low-rank and block sparse decomposition. The application of this modeling is to detect the mal-functioning sub-graphs in a wireless network. Since some sub-graphs of the network are assumed to mal-function, the corresponding outlier matrix would be block sparse. The data matrix, however, is low rank since the data pattern does not change too much over the time. An algorithm has been suggested to recover the low rank data matrix from the measurement matrix and eliminate the block sparse outlier. The simulation results confirm that the proposed technique outperforms the other anomaly detection techniques. The superiority of the suggested technique becomes magnificent when the block size (sub-graph size) increases. In such case, the sparsity of the outlier matrix decreases which deteriorates the efficiency of the sparsity-based techniques.

The rest of the paper is organized as follows: the proposed modeling together with the suggested decomposition algorithm are illustrated in Section II. The simulation results and relevant discussions are given in Section III. Section IV concludes the paper.

II. The compressed low-rank block sparse decomposition problem

Suppose that $\mathbf{R}$ is a routing matrix of size $L \times F$, $\mathbf{Z}$ is a clean traffic matrix of $F \times T$ and $\mathbf{A}$ is the outlier matrix of the same size which includes the network anomalies. L, F, and T indicate for the number of links, flows, and time slots, respectively. The measurement matrix $\mathbf{Y}$ can be modeled as [16]:

$$\begin{aligned}\mathbf{Y} =& \mathbf{R}(\mathbf{A}+\mathbf{Z}) \\ =& \mathbf{R}\mathbf{A}+\mathbf{X}\end{aligned} \tag{1}$$

where $\mathbf{X}$, which is the multiplication of the clean traffic matrix and the routing matrix, is called the data matrix. The problem to be solved here is to detect a number of mal-functioning sub-graphs in a network. Hence, the outlier matrix $\mathbf{A}$ would be block-sparse. The routing matrix is low-rank since there are repeated traffic patterns over the time and flow. Hence, the data matrix $\mathbf{X}$ would also be low-rank. In order to decompose the

low-rank data and the block-sparse outlier matrix, we define the following optimization problem.

$$\begin{aligned} &\min \|\mathbf{X}\|_* + \lambda \|\mathbf{A}\|_{2,1} \\ &\text{s.t.} \quad \mathbf{Y} = \mathbf{R}\mathbf{A} + \mathbf{X} \end{aligned} \tag{2}$$

Suppose that the matrix $\mathbf{A}$ consists of P blocks:

$$\mathbf{A} = [\mathbf{A}[1], \mathbf{A}[2], \mathbf{A}[3], \cdots, \mathbf{A}[P]] \tag{3}$$

The mixed $L_{1,2}$ norm can be calculated as:

$$\|\mathbf{A}\|_{2,1} = \sum_{l=1}^{P} \|\mathbf{A}[l]\|_2 \tag{4}$$

We apply the Alternating Direction Method of Multipliers (ADMM) [17] to solve (2). Being multiplied by the matrix $\mathbf{R}$, the entries of $\mathbf{A}$ are coupled which renders the problem more complicated. In order to address that issue, we introduce an auxiliary variable $\mathbf{B}$ to the problem as:

$$\begin{aligned} &\min \|\mathbf{X}\|_* + \lambda \|\mathbf{A}\|_{2,1} \\ &\text{s.t.} \quad \mathbf{Y} = \mathbf{R}\mathbf{B} + \mathbf{X} \\ &\text{s.t.} \quad \mathbf{A} = \mathbf{B} \end{aligned} \tag{5}$$

The quadratic augmented Lagrangian function would be obtained as:

$$\begin{aligned} &L(\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbf{M}, \mathbf{N}) \\ &= \|\mathbf{X}\|_* + \lambda \|\mathbf{A}\|_{2,1} + Tr(\mathbf{M}^T(\mathbf{Y} - \mathbf{R}\mathbf{B} - \mathbf{X})) \\ &+ Tr(\mathbf{N}^T(\mathbf{A} - \mathbf{B})) + (c/2)\|\mathbf{A} - \mathbf{B}\|_F^2 \\ &+ (c/2)\|\mathbf{Y} - \mathbf{R}\mathbf{B} - \mathbf{X}\|_F^2 \end{aligned} \tag{6}$$

The matrices $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{B}$ are the primal variables, while $\mathbf{M}$ and $\mathbf{N}$ are the dual matrices. The matrix $\mathbf{X}$ is updated as:

$$\begin{aligned} \mathbf{X}^k &= \text{argmin}_{\mathbf{X}} L(\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbf{M}, \mathbf{N}) \\ &= \text{argmin}_{\mathbf{X}} \|\mathbf{X}\|_* + \frac{c}{2}\|\mathbf{X} - \mathbf{Y} + \mathbf{R}\mathbf{B} - \frac{\mathbf{M}}{c}\|_F^2 \\ &= D(\mathbf{Y} - \mathbf{R}\mathbf{B} + \frac{\mathbf{M}}{c}, \frac{1}{c}) \end{aligned} \tag{7}$$

The function $D(\mathbf{x}, \tau)$ is defined as follows:

$$D(\mathbf{x}, \tau) = \mathbf{U} * F(\mathbf{S}, \tau) * \mathbf{V}^T \tag{8}$$

where the singular value decomposition of $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{U} * \mathbf{S} * \mathbf{V}^T \tag{9}$$

and the shrinkage function $F(\mathbf{x}, \tau)$ is defined as:

$$F(\mathbf{x}, \tau) = \begin{cases} \mathbf{x} - \tau & \text{if} \quad \mathbf{x} > \tau \\ \mathbf{x} + \tau & \text{if} \quad \mathbf{x} < -\tau \end{cases} \tag{10}$$

The second primal variable $\mathbf{A}$ is updated as follows:

$$\begin{aligned} \mathbf{A}^k &= \text{argmin}_{\mathbf{A}} L(\mathbf{X}, \mathbf{A}, \mathbf{B}, \mathbf{M}, \mathbf{N}) \\ &= \text{argmin}_{\mathbf{A}} \|\mathbf{A}\|_{2,1} + \frac{c}{2\lambda}\|\mathbf{A} - \mathbf{B} + \mathbf{N}/c\|_F^2 \end{aligned} \tag{11}$$

This problem can be decomposed as:

$$\mathbf{A}^k = \text{argmin}_{\mathbf{A}} \sum_{l=1}^{P} \|\mathbf{A}[l]\|_2 + \frac{c}{2\lambda}\|\mathbf{A}[l] - \mathbf{B}[l] + \mathbf{N}[l]/c\|_F^2 \tag{12}$$

In order to solve (12), we consider the point that the cost function is convex and non-smooth. Hence, the optimal solution is achieved when the sub-gradient equals zero. The sub gradient of the $L_{1,2}$ norm is derived in the following lemma.

Lemma 1. *For the function $f(\mathbf{A}) = \|\mathbf{A}\|_{2,1}$, the sub-gradient is derived as:*

$$\frac{\partial}{\partial \mathbf{A}[l]} f(\mathbf{A}) = \begin{cases} [-1, 1] & \|\mathbf{A}[l]\|_2 = 0 \\ \frac{\mathbf{A}[l]}{\|\mathbf{A}[l]\|_2} & \|\mathbf{A}[l]\|_2 \neq 0 \end{cases} \tag{13}$$

Proof: This lemma can be proved using the definition of the sub-gradient. The sub-gradient of the function $f(\mathbf{x})$ in x_0 is the vector ξ defined as [18]:

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \xi^T(\mathbf{x} - \mathbf{x}_0) \tag{14}$$

Here, we have: $f(\mathbf{A}) = \|\mathbf{A}\|_{2,1} = \sum_{l=1}^{P} \|\mathbf{A}[l]\|_2$. For the case of $\|\mathbf{A}[l]\|_2 \neq 0$, the function is smooth and differentiable, so we get:

$$\frac{\partial}{\partial \mathbf{A}[l]} f(\mathbf{A}) = \frac{\mathbf{A}[l]}{\|\mathbf{A}[l]\|_2} \tag{15}$$

For the case of $\|\mathbf{A}[l]\|_2 = 0$, according to the definition of sub-gradient, we have:

$$\|\mathbf{A}[l]\|_2 \geq Tr(\xi^T \mathbf{A}[l]) \tag{16}$$

Then, we would have: $|\xi_{i,j}| \leq 1$ which completes the proof. ■

The solution of (12) is characterized in the following Theorem.

Theorem 1.1. *The optimal solution of the optimization problem in* (12) *is characterized as:*

$$\mathbf{A}[l] = Th(\mathbf{Q}[l](1 - \frac{1}{\rho\|\mathbf{Q}[l]\|_2}), \theta) \tag{17}$$

where $\mathbf{Q}[l] = \mathbf{B}[l] - \mathbf{N}[l]/c$, $\theta > \rho$, $\rho = \lambda/c$, and $Th(., \theta)$ is the thresholding function defined as:

$$Th(\mathbf{z}[l], \theta) = \begin{cases} 0 & \|\mathbf{Q}[l]\|_2 \leq \theta \\ \mathbf{z}[l] & \|\mathbf{Q}[l]\|_2 > \theta \end{cases} \tag{18}$$

Proof: Let the function $g(.)$ be defined as the cost function in (II).

$$g(\mathbf{A}[l]) = \|\mathbf{A}[l]\|_2 + \frac{\rho}{2}\|\mathbf{A}[l] - \mathbf{B}[l] + \mathbf{N}[l]/c\|_F^2 \tag{19}$$

The sub-gradient of the cost function is obtained as:

$$\frac{\partial}{\partial \mathbf{A}[l]} g(\mathbf{A}) = \begin{cases} \frac{\mathbf{A}[l]}{\|\mathbf{A}[l]\|_2} + (1/\rho)(\mathbf{A}[l] - \mathbf{Q}[l]) & \|\mathbf{Q}[l]\|_2 \neq 0 \\ \tau + (1/\rho)(\mathbf{A}[l] - \mathbf{Q}[l]) & \|\mathbf{Q}[l]\|_2 = 0 \end{cases} \tag{20}$$

Setting the sub-gradient equal to zero, we would have:

$$\text{A[l]}=\begin{cases}\mathbf{Q}[l](1-\frac{1}{\rho\|\mathbf{Q}[l]\|_2}) & \|\mathbf{Q}[l]\|_2 \neq 0,(1/\rho)\\ \mathbf{Q}[l]-\tau/\rho & \|\mathbf{Q}[l]\|_2=0\end{cases} \quad (21)$$

$\|\mathbf{Q}[l]\|_2 = 0$ implies that $\mathbf{Q}[l] = 0$. In order to have a more stable algorithm, we introduce a thresholding operator as:

$$\mathbf{A}[l] = \begin{cases}\mathbf{Q}[l](1-\frac{1}{\rho\|\mathbf{Q}[l]\|_2}) & \|\mathbf{Q}[l]\|_2 \geq \theta\\ 0 & \|\mathbf{Q}[l]\|_2 < \theta\end{cases} \quad (22)$$

where we have selected $\tau = 0$. ■

The last primal variable $\mathbf{B}$ is updated according to:

$$\begin{aligned}&\mathbf{B}^k = \text{argmin}_{\mathbf{B}} L(\mathbf{X},\mathbf{A},\mathbf{B},\mathbf{M},\mathbf{N}) \quad (23)\\ &=(\mathbf{R}^T\mathbf{R}+\mathbf{I})^{-1}(\mathbf{R}^T(\mathbf{Y}-\mathbf{X}^k)+\mathbf{A}^k+1/c\mathbf{N}^{k-1}\\ &+1/c\mathbf{R}^T\mathbf{M}^{k-1})\end{aligned}$$

The dual variables, $\mathbf{M}$ and $\mathbf{N}$, are also straightforwardly updated according to the ADMM technqiue. The proposed algorithm for the solution of (2) is illustrated in Algorithm 1.

Algorithm 1 The proposed algorithm

input:
A routing matrix $\mathbf{R} \in \mathbb{R}^{m\times n}$.
A measurement matrix $\mathbf{Y} \in \mathbb{R}^m$.
The maximum number of iterations $iter_{max}$.
The number of blocks L.
The threshold value θ.
output:
The recovered matrix $\widehat{\mathbf{X}} \in \mathbb{R}^n$ of the original signal.
procedure THE PROPOSED ALGORITHM($\mathbf{y}, \mathbf{x}$)
    $\mathbf{X}^0 \leftarrow 0$
    $\epsilon \leftarrow 10^{-4}$
    $\mathbf{A}^0 \leftarrow 0$
    for k=1:$iter_{max}$ **do**
        $[\mathbf{U},\mathbf{S},\mathbf{V}] \leftarrow svd(\mathbf{Y}-\mathbf{R}\mathbf{B}^{k-1}+(1/c)\mathbf{M}^{k-1})$
        $X^k \leftarrow \mathbf{U} * F(\mathbf{S},1/c) * \mathbf{V}^T$

        for l=1:L **do**
            $\mathbf{Q}^k[l] \leftarrow \mathbf{B}^k[l]-\mathbf{N}^k[l]/c$
            $\mathbf{A}^k[l] \leftarrow Th(\mathbf{Q}^k[l](1-\frac{1}{\rho\|\mathbf{Q}^k[l]\|_2}),\theta)$
        end for
        $\mathbf{B}^k \leftarrow (\mathbf{R}^T\mathbf{R}+\mathbf{I})^{-1}(\mathbf{R}^T(\mathbf{Y}-\mathbf{X}^k)+\mathbf{A}^k+1/c\mathbf{N}^{k-1}+1/c\mathbf{R}^T\mathbf{M}^{k-1})$
        $\mathbf{N}^k \leftarrow \mathbf{N}^{k-1}+c(\mathbf{A}^k-\mathbf{B}^k)$
        $\mathbf{M}^k \leftarrow \mathbf{M}^{k-1}+c(\mathbf{Y}-\mathbf{X}^k-\mathbf{R}\mathbf{B}^k)$
    end for
    return $\widehat{\mathbf{X}} \leftarrow \mathbf{X}^{iter_{max}}$
end procedure

III. SIMULATION RESULTS

In this section, the simulation results are reported. The dimensions are $L = 105$, $F = 210$, and $T = 420$. The block sparsity of the matrix $\mathbf{A}$ which shows the percentage of malfunctioning sub-graphs is s. The parameters of the proposed algorithm are set as: $iter_{max} = 100$, $\rho = 0.1$, $\theta = 20$ and $\lambda = 1.5$. The matrix $\mathbf{R}$ is a random binary matrix with 50% ones. s percent of the blocks of A are selected randomly to be non-zero. The entries of the non-zero blocks are generated randomly according to $N(0,1)$. In the first scenario, the phase transition diagram of the proposed method is depicted in Figure 1. In this figure, the relative error of the recovered signal matrix $\mathbf{X}$, calculated as $e_r = \|\mathbf{X}_0-\mathbf{X}\|_F/\|\mathbf{X}_0\|_F$, is depicted for various values of rank r and block sparsity rate s. The darker color indicates lower value of relative error and better reconstruction region.

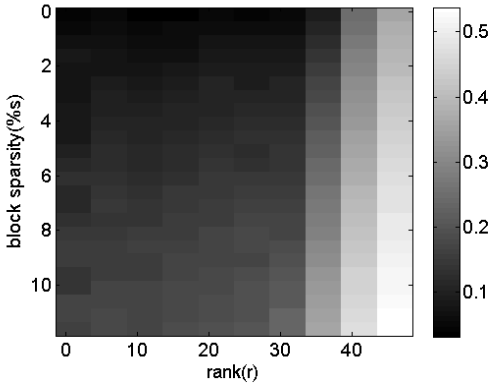

Fig. 1: The relative error of $\mathbf{X}_0$, $e_r = \|\mathbf{X}_0-\mathbf{X}\|_F/\|\mathbf{X}_0\|_F$, for various values of r and s where $L = 105$, $F = 210$, and $T = 420$.

expected, the performance of the algorithm is better for lower rank and lower block sparsity rate. Hence, the upper left corner of the figure represents for the lowest recovery error and the lower right corner is related to the highest error.

The next scenario is to investigate the performance of the method for various sizes of the matrix $\mathbf{R}$. Different parameters for assessing the recovery performance of the proposed method for 3 values of L are given in Table I.

TABLE I: Recovery performance by varying the size of $\mathbf{R}$ for $r = 10$ and $s = 0.05$

L	$\|\mathbf{A}_0\|_{2,1}$	$\|\hat{\mathbf{A}}\|_{2,1}$	$\|\mathbf{X}_0\|_*$	$\|\hat{\mathbf{X}}\|_*$	mse($\mathbf{A}$)	mse($\mathbf{X}$)
F	728.74	727.48	4568	4567	5.9e-06	5.6e-05
F/2	742.75	687.35	3238	3148	2.2e-04	1.9e-04
F/4	735.99	488.91	2293	2214	3.0e-04	3.2e-04

According to this table, as the value of L increases, the estimation of the rank of the nuclear norm of the data matrix and the $L_{1,2}$ norm of the outlier matrix becomes more accurate and the estimation error of both the data matrix and the outlier

matrix decreases. Hence, the algorithm behaves better as the matrix $\mathbf{R}$ approaches the square form.

In order to have a comparison between the proposed method and the algorithm in [16], we plot the relative error of the data matrix $\mathbf{X}$ as well as that of the outlier matrix $\mathbf{A}$ versus block sparsity in the Figures 2 and **??**, respectively. The rank of the data matrix is fixed to $r = 11$ in these cases.

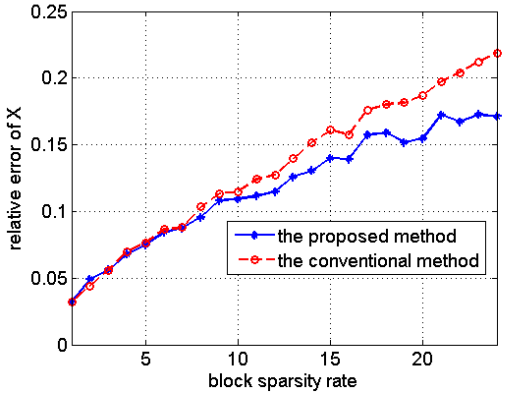

Fig. 2: The relative error of X versus block sparsity for both methods for $r = 11$, $L = 210$,$F = 420$, $T = 512$.

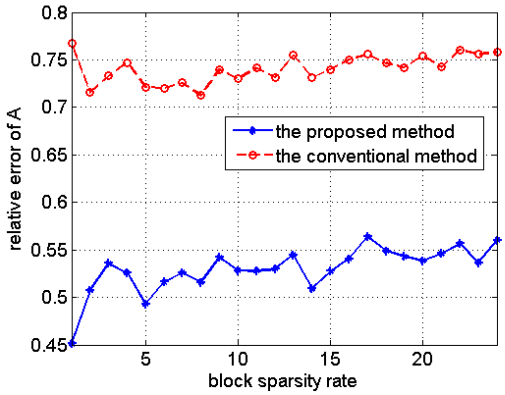

Fig. 3: The relative error of A versus block sparsity for both methods for $r = 11$, $L = 210$,$F = 420$, $T = 512$.

As can be seen from these two figures, the proposed method performs better than the other algorithm in terms of the recovery error of both the data matrix and the outlier matrix. Moreover, it is obvious from the two figures that increasing the block sparsity, the relative error increases too.

IV. Conclusion

This paper discusses the anomaly detection in wireless networks. The special kind of the anomaly considered in this paper is related to the case that a sub-graph of the network is mal-functioning. The proposed method detects such anomalies and recovers the corresponding data. The point to be exploited here is that the outlier matrix of the network graph would be block-sparse, the corresponding blocks of the mal-functioning sub-graphs would be non-zero, while the others would be zero. The low-rank property of the data matrix and the block-sparsity of the outlier matrix are leveraged to separate the useful data from the noisy outliers.

References

[1] M. Thottan, G. Liu, and C. Ji, "Anomaly detection approaches for communication networks," in *Algorithms for Next Generation Networks*, pp. 239–261, Springer, 2010.

[2] V. Alarcon-Aquino and J. A. Barria, "Anomaly detection in communication networks using wavelets," *IEE Proceedings-Communications*, vol. 148, no. 6, pp. 355–362, 2001.

[3] D. S. Yeung, S. Jin, and X. Wang, "Covariance-matrix modeling and detecting various flooding attacks," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 37, no. 2, pp. 157–169, 2007.

[4] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," in *ACM SIGCOMM Computer Communication Review*, vol. 35, pp. 217–228, ACM, 2005.

[5] A. Soule, K. Salamatian, and N. Taft, "Combining filtering and statistical methods for anomaly detection," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, pp. 31–31, USENIX Association, 2005.

[6] M. Azghani and F. Marvasti, "Sparse signal processing," in *New Perspectives on Approximation and Sampling Theory*, pp. 189–213, Springer, 2014.

[7] M. Azghani, P. Kosmas, and F. Marvasti, "Microwave medical imaging based on sparsity and an iterative method with adaptive thresholding," *IEEE Transactions on Medical Imaging*, vol. 34, no. 2, pp. 357–365, 2015.

[8] M. Azghani, M. Karimi, and F. Marvasti, "Multi-hypothesis compressed video sensing technique," *IEEE Transactions on Circuits and Systems for Video Technology*, 2014.

[9] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1791–1798, IEEE, 2010.

[10] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015.

[11] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, IEEE, 2013.

[12] L. Xiong, X. Chen, and J. Schneider, "Direct robust matrix factorizatoin for anomaly detection," in *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pp. 844–853, IEEE, 2011.

[13] M. Levorato and U. Mitra, "Fast anomaly detection in smartgrids via sparse approximation theory," in *2012 IEEE 7th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 5–8, IEEE, 2012.

[14] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *preprint*, vol. 12, 2009.

[15] Y. Peng, J. Suo, Q. Dai, and W. Xu, "Reweighted low-rank matrix recovery and its application in image restoration," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2418–2430, 2014.

[16] M. Mardani, G. Mateos, and G. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies," *IEEE Transactions on Information Theory*, vol. 59, pp. 5186–5205, Aug 2013.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.