

# A Low-Rank Matrix Approach for the Analysis of Large Amounts of Power System Synchrophasor Data

**Meng Wang, Joe H. Chow, Pengzhi Gao, Xinyu Tony Jiang, Yu Xia, Scott G. Ghiocel**

Rensselaer Polytechnic Institute (RPI)

wangm7, chowj, gaop, jiangx4, xiay4, ghiocs2@rpi.edu

**Bruce Fardanesh, George Stefanopoulos**

New York Power Authority (NYPA)

Bruce.Fardanesh, George.Stefopoulos@nypa.gov

**Yutaka Kokai, Nao Saito**

Hitachi America

Yutaka.Kokai, Nao.Saito@hal.hitachi.com

**Micahel Razanousky**

New York State Energy Research and Development Authority (NYSERDA)

Michael.Razanousky@nyserda.ny.gov

**Abstract**—With the installation of many new multi-channel phasor measurement units (PMUs), utilities and power grid operators are collecting an unprecedented amount of high-sampling rate bus frequency, bus voltage phasor, and line current phasor data with accurate time stamps. The data owners are interested in efficient algorithms to process and extract as much information as possible from such data for real-time and off-line analysis. Traditional data analysis typically analyze one channel of PMU data at a time, and then combine the results from the individual analysis to arrive at some conclusions. In this paper, a spatial-temporal framework for efficient processing of blocks of PMU data is proposed. A key property of these PMU data matrices is that they are low rank. Using this property, various data management issues such as data compression, missing data recovery, data substitution detection, and disturbance triggering and location can be processed using singular-value based algorithms and convex programming. These functions are illustrated using some historical data from the Central New York power system.

**Index Terms**—Power system dynamics, synchrophasor, phasor measurement unit, low-rank matrix, data completion algorithms, singular values

## I. INTRODUCTION

THIS paper proposes a framework to facilitate efficient management of and information extraction from large amount of high-sampling rate synchronized phasor measurements from large power systems [33]. After the DOE smart grid investment program started in 2009, the North American power system will have close to 2,000 phasor measurement units (PMUs) in a few years [16]. Most of these PMUs have multiple channels. At a phasor data transmission rate of 30, 50, or 60 samples per second, collectively the PMU data accumulation is in the range of several terabits per day [32]. Such collections of data will be futile if they are

not supported by efficient data management and information extraction methods.

At this initial stage of PMU network deployment, the common concerns include ensuring the quality of PMU data like filling in missing data, and using the PMU data for studying the transient impact of disturbances. Currently, these tasks are mostly accomplished by analyzing the channels individually and then combining the results if necessary [6], [21], [42]. Such single-channel analysis can become less efficient with a denser coverage of PMUs of a power grid. Furthermore the existing techniques are task specific and lack a common framework, as they were developed when the coverage of PMU data was quite sparse.

A denser coverage by PMUs allows the analysis of data collectively from PMUs located in electrically close regions and in distinct control regions. Furthermore, the affinity and diversity of these clusters of PMU data can be discovered by processing them over a period of time. This idea of processing such spatial-temporal blocks of PMU data, in our opinion, will make information extraction more efficiently. This block-data analysis approach based on low-rank matrices has been used in other disciplines such as movie ratings [1], computer vision [12], [39], machine learning [2], [3], remote sensing [37], and system identification [30]. In power systems, such blocks of PMU data have been used to study the propagation of electromechanical oscillations after a disturbance [15], [35] and to detect voltage stability using a singular value approach [28]. This paper uses the spatial-temporal block data approach to provide a common framework to launch a set of similar algorithms to manage and extract information from the PMU data. Some of these algorithms have already been developed for other disciplines.

The remainder of this paper is organized as follows. Section II provides some background and insights for this approach.

X. T. Jiang is now at Quanta Technology, and S. Ghiocel is now at Exponent

Section III uses a singular value decomposition approach for data compression. Section IV discusses the use of convex programming algorithms to recover missing PMU data. Section V describes a convex programming formulation to detect PMU data substitution. Section VI utilizes singular values to detect the occurrence of disturbances. Illustrations of these low-rank matrix approaches using historical New York PMU data are provided in these sections.

## II. LOW-RANK MATRIX ANALYSIS OF SPATIAL-TEMPORAL PMU DATA BLOCKS

Power system measured data capture the bus voltages and line currents of a geographical disperse network over time. In power grid control centers, the SCADA data collected at every 4 to 5-second time spans can be viewed as discrete data, missing the system dynamic response if a disturbance happens between two successive sets of SCADA data points. However, system dynamics up to half of the PMU data transmission frequency (Nyquist frequency) can be captured by PMU data. The increased visibility of the power system dynamics comes with its own burden, namely, how should the data be processed efficiently so that the relevant information could be readily extracted and used.

As the PMU data collection routinely yields hundreds of gigabits of data per day, single-channel analysis may no longer be optimal. An alternative is to consider the PMUs in electrically close area as a group and analyze this cluster of PMU data over a fixed period of time (say 5 to 20 seconds) simultaneously as a block. Figure 1 depicts the PMU data arranged as a matrix with the rows indicating the PMUs and its channels, and the columns are ordered time-tagged data, with the most recent data in the last columns. The PMUs can also be grouped by the regions that they are located in, such as New York, New England, and PJM. Although not shown in Figure 1, the New York PMUs can be further divided into smaller, electrically close regions such as Central, Western, and Northern New York. Figure 1 also depicts some of the typical data issues in PMU data collection:

- 1) Missing data refer to data that have not arrived at the phasor data concentrator (PDC) in time to be collected. The reasons for this data drop may include PMU malfunction, insufficient network bandwidth, congestion, and data dropout due to malicious reasons. A means to readily recover or reconstruct the missing data will be of great interest.
- 2) Bad data refer to those data from PMUs with instrumentation errors, such as polarity and scaling errors. Such data normally can be detected from state estimation algorithms using data from the same time instant. This paper will not deal with such data correction.
- 3) Cyber attack, although unlikely, can impact PMU data in many ways. One of the more sophisticated cyber attacks involves data substitution that cannot be detected by state estimators [25]. It is highly desirable to find a reliable means to detect such cyber attacks.
- 4) Disturbance triggering, propagation, and location are important information for control center operation. In

particular, identifying disturbance events caused by prior disturbances is important as they may be precursors to cascading failures.

This paper looks into the development of techniques to perform these diverse tasks so that large amounts of PMU data can be handled efficiently. It is highly desirable that these algorithms can be based on a common framework.

The starting point of the proposed analysis approach is a basic data block which is a matrix cell of PMU data (Figure 1). Power system is an interconnected network, that is, data measured at various buses will be driven by some underlying system condition. The system operating condition may change, but some consistent relationship between the PMU data from different buses will still be there. Given the PMU data over a period of time  $T$  at a particular bus, most likely the PMU data on neighboring buses will have a similar response. In an extreme case when the power system operating condition does not change in 5 minutes, then the rank of the data matrix cell will be unity for these 5 minutes, as all the PMU data will be constant, that is, all the rows in the cell will be linearly dependent. During disturbances, the rank of the cell will be greater than one, but still small. This low-rank condition of the PMU data cells offers a common framework to develop PMU data analysis schemes.

The study of low-rank matrices has attracted the attention of many researchers in applied mathematics and optimization theory, since the announcement of the Netflix Prize [1]. The singular values of a matrix, which can be computed efficiently, are good indicators of the similarity of the data. More important, efficient low-rank matrix completion and decomposition methods have been developed [8]–[11], [13], [19], [24], [34], [41].

The following sections describe the use of data block cells to process the PMU data measured in Central New York. The system diagram is shown in Figure 2 [22]. There are 6 substations equipped with PMUs measuring frequencies, bus voltage phasors, and line current phasors. The PMU data is recorded at 30 samples per second. Using the phasor-only state estimator [22], the virtual PMU data at 7 additional substations can be calculated. The data from two PMUs in Western New York are also available. Historical PMU data measured during a loss-of-generation event in New England are used for the illustration.

## III. DATA COMPRESSION

For data compression studies, PMU data of 20 seconds long are analyzed, that is, the PMU data matrix has 600 columns. Figure 3 shows the frequency values at 5 buses in Central New York (CNY), and one bus in Western New York (WNY) (plotted in red). Note that this was a loss of generator event as the frequency dropped by 20 mHz. Even though this was a disturbance event in which the system went through a transient response, the frequency variations between the various substations in NY were very similar.

Singular value decomposition is applied to a PMU data matrix cell  $L$  with these six channels for 20 seconds

$$L = U\Sigma V^T \quad (1)$$

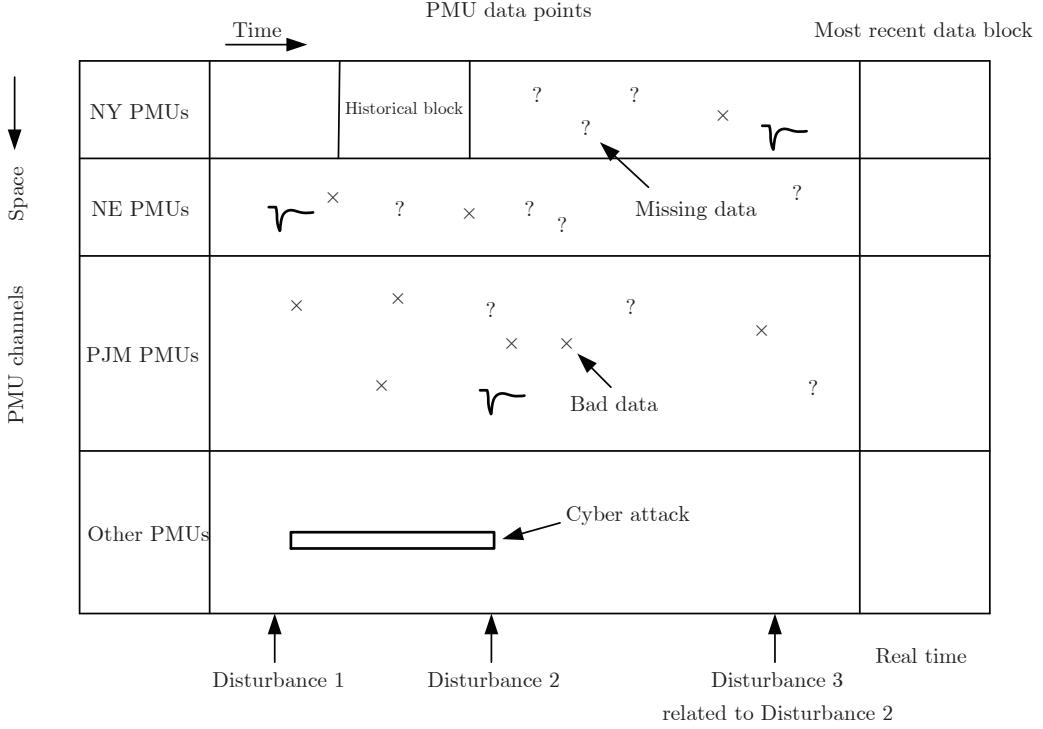


Fig. 1. Spatial-temporal PMU data shown as a matrix

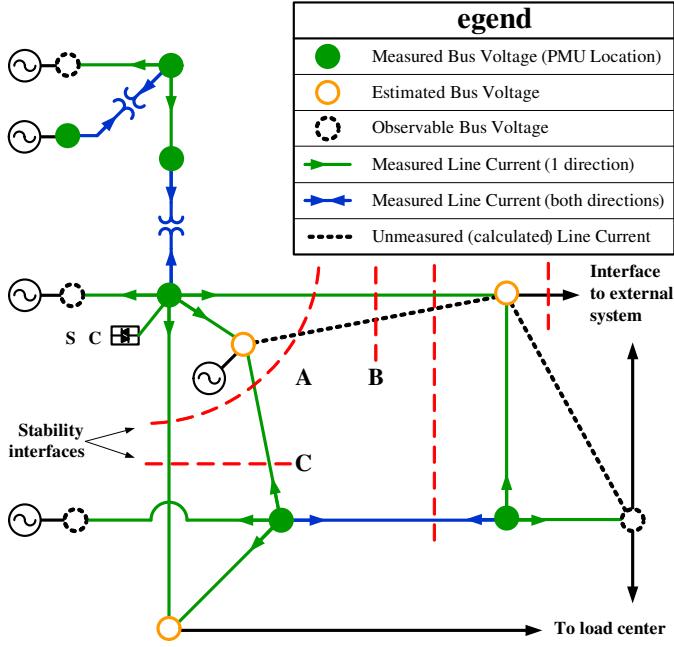


Fig. 2. Locations of PMUs in the Central New York Power System

where  $L$  is 6 by 600,  $\Sigma$  is the matrix with singular values on the main diagonal, and  $U$  and  $V$  are the left and right singular vectors, respectively.

If  $L$  is low rank,  $L$  can be approximated by

$$\hat{L} = \hat{U}\hat{\Sigma}\hat{V}^T \quad (2)$$

where  $\hat{\Sigma}$  contains only the most significant singular values,

and  $\hat{U}$  and  $\hat{V}^T$  are the corresponding singular vectors.

For the responses shown in Figure 3, the singular values of  $L$  are

$$\sigma = [3597.1, 0.086, 0.022, 0.010, 0.0084, 0.0078] \quad (3)$$

The largest singular value, a measure of the norm of the matrix, is 3597.1, which is much larger than the other singular values. The rest of the singular values are much smaller, reflecting the small differences in the time responses at different PMU locations.

Figures 4 and 5 show the result of retaining only the largest singular value and the two largest singular values, respectively, for representing this disturbance time response. With only the largest singular value, Figure 4 shows that the time response of the WNY PMU becomes the same of the CNY PMU responses. By including the second largest singular value, the WNY PMU time response becomes more accurate. The RMS error for keeping various number of singular values is shown in Figure 6. If only two singular values are kept, the data compression is about 66%. Note that this compression is lossy. A similar idea is found in [14], [17], using a principal component analysis of block PMU data. A similar compression can be applied to the voltage and current phasor data.

Note that as expected, the time responses shown after data compression are “filtered” responses, as some of the small spikes in the original data have been removed. It should be noted that the filtered data may remove the small amplitude oscillations in ambient conditions as observed in [27] and render the compressed data not suitable for such analysis.

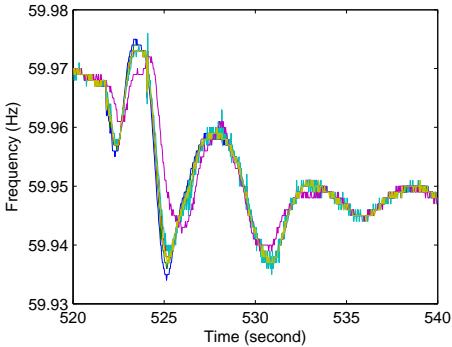


Fig. 3. Original frequency profile

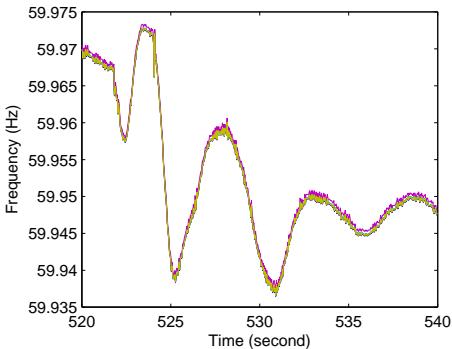


Fig. 4. Frequency profile with one singular value retained

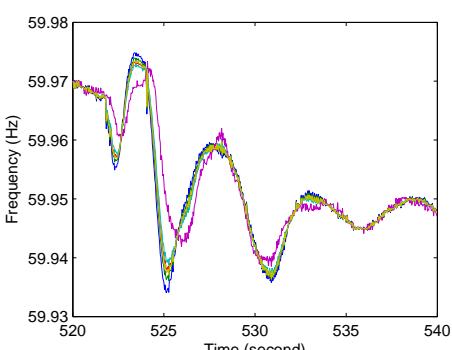


Fig. 5. Frequency profile with two singular values retained

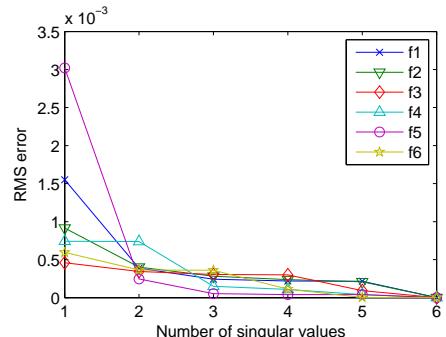


Fig. 6. RMS error from SVD compression

#### IV. MISSING DATA RECOVERY

In our recent paper [20], we proposed data recovery algorithms based on the low-rank property of the PMU data. We verify our methods on actual PMU data in the Central New York Power System (Fig. 2). Six PMUs directly measure 37 voltage and current phasors at 30 samples per second. Figure 7 shows the current magnitudes of the PMU data for a 20-second period. An event occurs around 2.5s. Let the matrix  $L \in \mathbb{C}^{37 \times 600}$  contain the measurements of the voltage phasors and the current phasors. The singular values of  $L$  are shown in Figure 8. The ninth largest singular value is 0.5930, while the largest one is 894.6. We could use a rank-eight matrix to approximate  $L$  by keeping its largest eight singular values and setting the remaining ones to zero. The approximation error is very small. Note that when no event happens, we could design a low-rank matrix with a smaller rank to approximate the actual PMU data. The low-dimensionality of PMU data is also observed in recent work [14], [17].

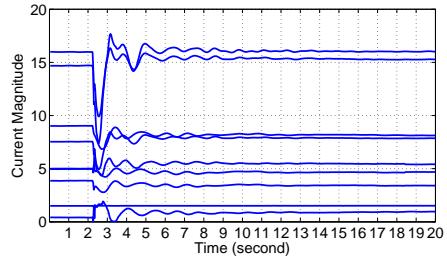


Fig. 7. Current magnitudes of PMU data (9 current phasors out of 37 phasors)

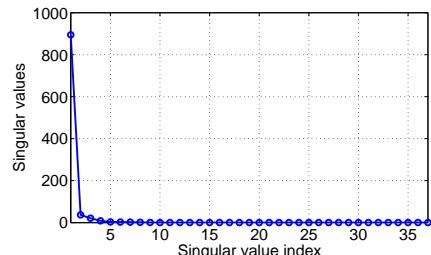


Fig. 8. Singular values of the PMU data matrix in decreasing order

In general, let matrix  $L \in \mathbb{C}^{n_1 \times n_2}$  contain the phasor measurements of multiple PMU channels at multiple synchronized sampling instants. Because  $L$  could be approximated by a low-rank matrix,<sup>1</sup> we formulate the problem of recovering missing PMU measurements as a low-rank matrix completion problem in [20]. The goal is to identify the missing points in a low-rank matrix from other observed entries. One popular matrix completion method is to solve the following nuclear-norm minimization problem [19] and use its solution as an estimate of the original low-rank matrix  $L$ :

$$\min_{X \in \mathbb{C}^{n_1 \times n_2}} \|X\|_* \quad \text{s.t.} \quad P_\Omega(X) = P_\Omega(L) \quad (4)$$

<sup>1</sup>Strictly speaking,  $L$  is approximately low-rank and can be viewed as the summation of a low-rank matrix plus noise. We assume  $L$  is strictly low-rank for notational simplicity in this paper, and the results can be extended to low-rank matrices plus noise with little modifications.

where  $\Omega$  denotes the set of locations of the observed entries, and

$$[P_\Omega(X)]_{ij} = \begin{cases} X_{ij}, & (i, j) \in \Omega \\ 0, & (i, j) \notin \Omega \end{cases}$$

The nuclear norm  $\|X\|_*$  is the sum of the singular values of  $X$ . The optimization (4) is convex and can be solved efficiently [19]. Furthermore, even when only  $O(rn \log^2 n)$  ( $n = \max(n_1, n_2)$ )<sup>2</sup> randomly selected entries of a rank- $r$  matrix  $L$  are observed,  $L$  is guaranteed to be the unique solution to (4) (see, e.g., [9], [10], [24], [34]). Therefore, if the rank  $r$  is much smaller than  $n$ , a small percentage of observed entries are sufficient to fully determine the matrix  $L$ .

The existing theoretical analysis of matrix completion methods critically relies on the assumption that the locations of the missing entries (erasures) are independent of each other. This model of independent erasures, however, does not adequately describe the pattern of missing PMU data points, which usually exhibit temporal correlations (one PMU channel may lose a sequence of measurements for some period of time) and spatial correlations (measurements in multiple channels of a PMUs might be lost simultaneously). In [20], we propose two models to characterize the temporal and the spatial correlations in the locations of missing PMU data. We also provide a theoretical guarantee (Theorem 1 and 2 in [20]) that even though the locations of the missing data points exhibit certain correlation that is unknown to the operator, a low-rank matrix completion can still reconstruct the missing data correctly.

We skip the theory and only show the numerical results here. We use the first 5-second part of the PMU data in Figure 7. The matrix  $L$  is a  $37 \times 150$  complex matrix with each row representing the consecutive measurements of one voltage/current phasor. We remove some entries of  $L$ , and the locations of the missing points are either temporally or spatially correlated. We use the singular value thresholding (SVT) algorithm [7] and information cascading matrix completion (ICMC) algorithm [31] to reconstruct the missing data. The reconstruction performance is measured by the relative recovery error  $\|L - L_{\text{rec}}\|_F / \|L\|_F$ , where  $L_{\text{rec}}$  is the reconstructed matrix, and  $\|L\|_F$  is the Frobenius norm of  $L$ . The average erasure rate  $p_{\text{avg}}$  denotes the ratio of the number of missing entries to the total number of entries.

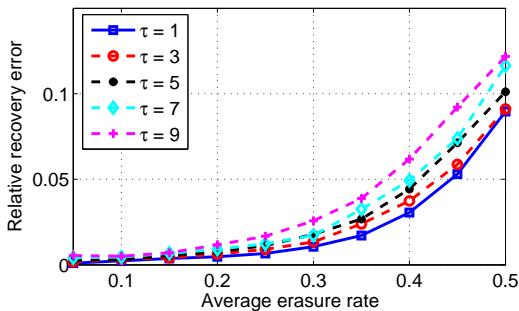


Fig. 9. Relative recovery error of the SVT algorithm

<sup>2</sup> $g(n) \in O(h(n))$  means as  $n$  goes to infinity,  $g(n) \leq ch(n)$  eventually holds for some positive constant  $c$ .

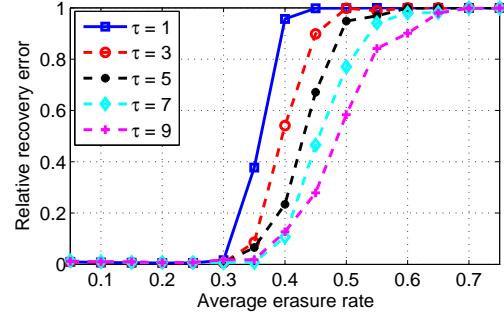


Fig. 10. Relative recovery error of the ICMC algorithm

Figures 9 and 10 (Figures 4 and 5 of [20]) show the relative recovery error of SVT and ICMC when the locations of the erasures are temporally correlated. We simulate the temporally correlated erasures by erasing  $\tau$  consecutive measurements in a selected PMU channel when an instance of data loss occurs. With the same erasure rate, the temporal correlation increases when  $\tau$  increases. Each curve is averaged over 100 runs. With either method, the reconstruction error is generally less than 3% even when about 20~30% of the measurements are missing. We also obtain similar reconstruction performance when the locations of the missing data are spatially correlated. Please refer to [20] for details.

The reconstructed measurements by low-rank matrix completion methods could be used for offline applications such as model validation and post-event analysis. We are also developing an online missing data recovery method for real-time applications such as state estimation and disturbance detection.

## V. DATA SUBSTITUTION

Introducing modern information technologies into power system operations could bring many benefits, but it might also increase the possibility of cyber attacks from malicious intruders. One type of cyber attacks is the data substitution attack where an intruder, aware of the system configuration information, can alter multiple measurements simultaneously in an intelligent way. These data substitutions are carefully designed by the intruder to be consistent with each other so that they cannot be detected by any existing bad data detector that rely on the redundancy of measurements. These cyber data attacks are called “the worst interacting bad data injected by an adversary” [26] and could result in significant errors in the output of state estimators. These data substitution attacks are first studied in [29] and are called “unobservable attacks” in [26], because the removal of affected measurements would make the system unobservable. Note that the proliferation of PMU installation in the power system has limited contribution to the reduction of the risk of data substitution attacks. For example, an intruder can inject an arbitrary error to the estimation of a bus voltage, as long as it can manipulate the voltage measurement of the particular bus and the current measurements of the transmission lines that are incident to that bus.

Existing approaches on state estimation in the presence of data substitutions [4], [18], [26], [29], [36], [38] usually attempt to identify and protect some key PMUs so as to prevent the intruders from injecting unobservable attacks. To the best of our knowledge, only one recent paper [38] addresses the detection of unobservable attacks in Supervisory Control and Data Acquisition (SCADA) system. The detection method proposed in [38] is based on statistical learning and relies on the assumption that the measurements at different time instants are independent and identically distributed samples of random variables. This assumption might not hold when the system is experiencing some disturbances.

In the recent paper [40], we propose a new method that can identify the unobservable cyber data attacks on PMU measurements, even when the system is under disturbance. We consider the scenario that an intruder can inject data to multiple PMU channels continuously so that at each time instant the injection is undetectable to any bad data detectors and would result in a wrong estimate of the system state. Our method can detect occasional and sustained data attacks in both steady state operation and transient condition, as long as the dynamics of the data injections is different from the system dynamics.

We use bus voltage phasors as the state variables. Let  $n$  denote the number of buses, and  $p$  denote the total number of PMU channels that measure voltage/current phasors. Let  $Z^{ij}$  and  $Y^{ij}$  denote the impedance and admittance, respectively, of the transmission line between Bus  $i$  and Bus  $j$ . Let  $S \in \mathbb{C}^{n \times t}$  denote the state variables in  $t$  time instants with each column representing the system state at a given time. We define matrix  $\bar{W} \in \mathbb{C}^{p \times n}$  as follows. If the  $k$ th PMU channel measures the voltage phasor of bus  $j$ ,  $\bar{W}_{kj} = 1$ ; if it measures the current phasor from bus  $i$  to bus  $j$ , then  $\bar{W}_{ki} = 1/Z^{ij} + Y^{ij}/2$ ,  $\bar{W}_{kj} = -1/Z^{ij}$ ;  $\bar{W}_{kj} = 0$  otherwise. We use  $L \in \mathbb{C}^{p \times t}$  to denote the actual PMU measurements (without attack). Then  $L$  and  $S$  are related through

$$L = \bar{W}S \quad (5)$$

Let  $M \in \mathbb{C}^{p \times t}$  denote the obtained measurements (with data attacks). With data substitutions, a state estimator will output  $S + \bar{C}$  as system states across time instead of the true states  $S$ . The matrix  $\bar{C} \in \mathbb{C}^{n \times t}$  denotes the additive errors to the system states due to cyber data attacks, where each column corresponds to the errors at one time instant, and each row corresponds to the errors to one bus voltage across time. Then  $M$  can be represented as

$$M = L + \bar{W}\bar{C} \quad (6)$$

We propose to identify the unobservable data attacks by solving the following convex program

$$\min_{X \in \mathbb{C}^{p \times t}, C \in \mathbb{C}^{n \times t}} \|X\|_* + \nu \sum_{i=1}^n \|C_{i*}\|_2, \quad \text{s.t. } M = X + WC \quad (7)$$

where  $\nu$  is some positive constant. The  $j$ th column of  $W$  satisfies  $W_j = \bar{W}_j/\|\bar{W}_j\|_2$ , and  $\|C_{i*}\|_2$  represents the  $\ell_2$  norm of the  $i$ th row of  $C$ . Note that (7) is convex [5] and thus, can be solved efficiently by solvers such as CVX [23]. Let  $(\hat{X}, \hat{C})$

denote the optimal solution to (7) and the set  $J$  contain the indices of nonzero rows of  $\hat{D} := W\hat{C}$ . We demonstrate in [40] through theoretical analysis that under mild assumptions,  $J$  is indeed the set of PMU channels that are under cyber data attacks.

We verify the proposed method on the data set from the CNY power system. The weight  $\nu$  is fixed to be 1. We first consider the scenario that an intruder attacks the PMUs on Buses 1 and 5, and changes the current phasors from Bus 1 to Bus 2 and from Bus 5 to Bus 2 (denoted by  $I^{12}$  and  $I^{52}$ ) simultaneously, so that the erroneous voltage phasor on Bus 2 as calculated from  $I^{12}$  agrees with that calculated from  $I^{52}$ . Since the voltage and current phasors are represented by complex numbers in our simulation, we generate the additive errors from the complex normal distribution. Two 2-second PMU data sets are tested. One contains ambient data ( $t = 17 \sim 19$  s in Figure 7). The other one contains the data with system dynamic response or abnormal event ( $t = 2 \sim 4$  s in Figure 7).

Because the actual PMU data contain noise, we identify a row of  $\hat{D}$  as nonzero if its  $\ell_2$  norm exceeds some predefined threshold. Figure 11 shows the  $\ell_2$  norm of each row of the resulting  $\hat{D}$  matrix. Our method successfully identifies the two PMU channels that are under data attack (significant  $\ell_2$  norm), and they correspond to the channels that measure  $I^{12}$  and  $I^{52}$ .

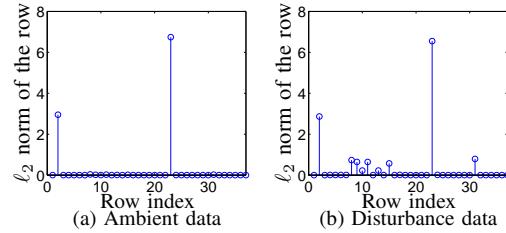


Fig. 11. The  $\ell_2$  norm of each row of  $\hat{D}$  for ambient and disturbed data

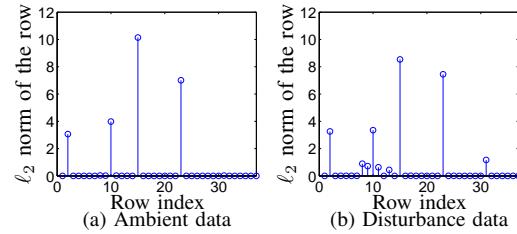


Fig. 12. The  $\ell_2$  norm of each row of  $\hat{D}$  for ambient and disturbed data

We next increase the number of PMU channels under attack. We assume the intruder alters the PMU channels that measure  $I^{12}$ ,  $I^{52}$ ,  $I^{13}$  and  $I^{43}$ , so that the voltage phasor estimate of Buses 2 and 3 are corrupted.

Figure 12 shows the  $\ell_2$  norm of each row of the resulting  $\hat{D}$  matrix. The rows with significant  $\ell_2$  norm also correspond to the exact PMU channels that are under attack.

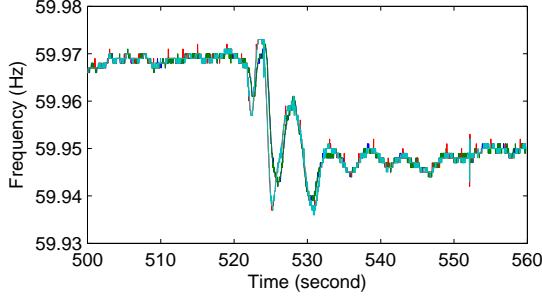


Fig. 13. Frequency profile

## VI. DISTURBANCE TRIGGERING

The rate-of-change-of-frequency ( $df/dt$ ) data channel in a PMU is often used for disturbance triggering indicating the occurrence of a disturbance that significantly affects the flow conditions at the bus. It will be useful to expand this local view of disturbance detection into more comprehensive system-wide disturbance triggering and location schemes. In the low-rank matrix approach, singular values of the PMU data matrix can be used for disturbance detection and triggering. This can be done by moving a window of a certain time length over a PMU data block, and performing SVD on the window each time the window moves forward in real time. Preferably the PMU data block will have measurements of the same type, such as all voltage magnitudes or all bus frequencies. The singular values of each window of data can be viewed as the correlation between the measurements in the different PMU channels. In steady state when the data are highly correlated, the data matrix will have one large singular value<sup>3</sup> (in case when only voltage channel data are used) and the other singular values are very small. During a disturbance, however, the discrepancies between PMU channels become significant, thus causing a few small singular values to become 5 to 10 times bigger. The jump in the values of these singular values provides an opportunity for disturbance triggering.

This analysis is applied to the PMU frequency measurements from the same event in Section III. The time range of PMU data is expanded to 60 seconds long and the number of the PMU data matrix columns thus increases to 1800. Figure 13 shows the frequency traces at two buses in Central New York, and two buses in Western New York. The differences in frequency variations in each individual area were very small. Figure 14 shows the trace of the second largest singular value. SVD was performed on the data block using a window size of 30 samples (1 second), and the window is moved 15 samples each time. The window size and the number of samples to move forward between each SVD calculation can be adjusted to modify the sharpness and resolution of the resulting trace. Figures 13 and 14 clearly show that the second largest singular value rises sharply at the onset of the disturbance. The second largest singular value can therefore be monitored to detect disturbances.

It should be noted that this block processing technique can process PMU data in groups and thus can assess the

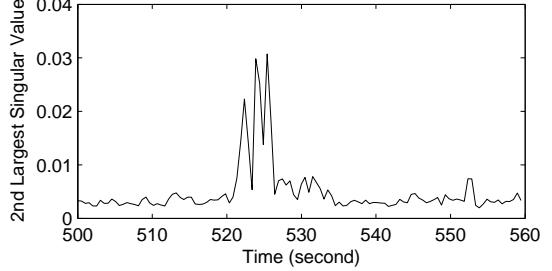


Fig. 14. The second largest singular value

spread of a disturbance rapidly. This will be a more efficient process compared to processing the PMUs on an individual basis. It should be also noted that any intentional system switchings, such as closing a shunt capacitor bank or moving a transformer tap, will result in transients that can be detected by the algorithm. Post processing will be required to sort out these switching conditions.

The singular-value-based disturbance triggering approach can be further used to detect the origin of the disturbance. This can be achieved by first dividing an interconnected power system into the individual regions, for example, the New York power system as a region, and the New England power system as another region. Then a data matrix from selected PMUs is constructed for each region. These matrices are then used to determine the arrival time of disturbance and the severity of the disturbance as seen locally. By examining the local time of arrival of the disturbance and the severity of the disturbance, the direction of the disturbance propagation can be determined and the origin of the disturbance can be located. More details will be reported later.

## VII. CONCLUSIONS

This paper describes a new framework of processing large amounts of synchrophasor data. The central idea is that the PMU data matrix of nearby buses exhibits a consistent relationship such that the matrix has low rank. This property can be exploited to develop methods for data compression, missing data recovery, data substitution attach resolution, and disturbance triggering. These methods are based on singular value decomposition and convex programming. In addition, parallel processing can be applied to obtain the results from the individual blocks, so that the PMU data can be processed efficiently.

## ACKNOWLEDGEMENT

This work was supported in part by the Engineering Research Center Program of the National Science Foundation and the Department of Energy under NSF Award Number EEC-1041877, NYSERDA grants #28815 and #36653, Hitachi America, and NYPA.

## REFERENCES

- [1] Netflix Prize. available: <http://www.netflixprize.com>, 2009.
- [2] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proc. the 24th international conference on Machine learning*, pages 17–24. ACM, 2007.

<sup>3</sup>The largest singular value reflects that the nominal values are non-zero.

- [3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.
- [4] R. B. Bobba, K. M. Rogers, Q. Wang, H. Khurana, K. Nahrstedt, and T. J. Overbye. Detecting false data injection attacks on DC state estimation. In *Proc. the First Workshop on Secure Control Systems (SCS)*, 2010.
- [5] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [6] A. Bykhovsky and J. H. Chow. Dynamic data recording in the new england power system and an event analyzer for the northfield monitor. *Electric Power and Energy Journal*, 25:787–795, 2003.
- [7] J.-F. Cai, E. J Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [8] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.
- [9] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.
- [10] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010.
- [11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [12] P. Chen and D. Suter. Recovering the missing components in a large noisy low-rank matrix: Application to sfm. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1051–1063, 2004.
- [13] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Trans. Inf. Theory*, 59(7):4324–4337, 2013.
- [14] Y. Chen, L. Xie, and P. R. Kumar. Dimensionality reduction and early event detection using online synchrophasor data. *Proceedings of 2013 IEEE PES General Meeting*, pages 1–5, 2013.
- [15] M. L. Crow and A. Singh. The matrix pencil for power system modal extraction. *IEEE Transactions on Power Systems*, 20(1):501–502, 2005.
- [16] J. Dagle. Current PMU locations - map of networked PMUs. available: <https://www.naspi.org/documents/2013>.
- [17] N. Dahal, R. L. King, and V. Madani. Online dimension reduction of synchrophasor data. In *Proc. IEEE PES Transmission and Distribution Conference and Exposition (T&D)*, pages 1–7, 2012.
- [18] G. Dán and H. Sandberg. Stealth attacks and protection schemes for state estimators in power systems. In *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 214–219, 2010.
- [19] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, 2002.
- [20] P. Gao, M. Wang, S.G. Ghiocel, and J. H. Chow. Modeless reconstruction of missing synchrophasor measurements. In *Proc. IEEE PES General Meeting*, 2014.
- [21] R. M. Gardner, Y. Liu, and Z. Zhong. Location determination of power system disturbances based on frequency responses of the system. *US Patent 7,765,034*, 2010.
- [22] S. G. Ghiocel, J. H. Chow, G. Stefopoulos, B. Fardanesh, D. Maragall, M. Razanousky, and D. B. Bertagnolli. Phasor state estimation for synchrophasor data quality improvement and power transfer interface monitoring. *IEEE Transactions on Power Systems*, 29(2):881–888, 2014.
- [23] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. available: <http://cvxr.com/>.
- [24] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory*, 57(3):1548–1566, 2011.
- [25] T. T. Kim and H. V. Poor. Strategic protection against data injection attacks on power grids. *IEEE Transactions on Smart Grid*, 2(2):326–333, 2011.
- [26] O. Kosut, L. Jia, R. J. Thomas, and L. Tong. Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures. In *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 220–225, 2010.
- [27] Dosiek L., N. Zhou, J. Pierre, Z. Huang, and D. Trudnowski. Mode shape estimation algorithms under ambient conditions: A comparative review. *IEEE Transactions on Power Systems*, 28(2):779–787, 2013.
- [28] J. M. Lim and C. L. DeMarco. Model-free voltage stability assessments via singular value analysis of PMU data. In *IREP Symposium*, 2013.
- [29] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1):13, 2011.
- [30] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.
- [31] R. Meka, P. Jain, and I. S Dhillon. Matrix completion from power-law distributed samples. In *Advances in Neural Information Processing Systems*, pages 1258–1266, 2009.
- [32] M. Patel, S. Aivaliotis, E. Ellen, et al. Real-time application of synchrophasors for improving reliability. *NERC Report*, 2010.
- [33] A. G. Phadke and J. S. Thorp. *Synchronized phasor measurements and their applications*. Springer, 2008.
- [34] B. Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, pages 3413–3430, 2011.
- [35] J. J. Sanchez-Gasca and J. H. Chow. Performance comparison of three identification methods for the analysis of electromechanical oscillations. *IEEE Transactions on Power Systems*, 14(3):995–1002, 1999.
- [36] H. Sandberg, A. Teixeira, and K. H Johansson. On security indices for state estimators in power networks. In *Proc. the First Workshop on Secure Control Systems (SCS)*, 2010.
- [37] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, 34(3):276–280, 1986.
- [38] H. Sedghi and E. Jonckheere. Statistical structure learning of smart grid for detection of false data injection. In *Proc. IEEE Power and Energy Society General Meeting (PES)*, pages 1–5, 2013.
- [39] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.
- [40] M. Wang, P. Gao, S.G. Ghiocel, J. H. Chow, B. Fardanesh, G. Stefopoulos, and M. P. Razanousky. Identification of “unobservable” cyber data attacks on power grids. *submitted to IEEE SmartGridComm*, 2014.
- [41] H. Xu, C. Caramanis, and S. Sanghavi. Robust pca via outlier pursuit. *IEEE Trans. Inf. Theory*, 58(5):3047–3064, May 2012.
- [42] Y. Ye, J. Dong, and Y. Liu. Analysis of power system disturbances based on distribution-level phasor measurements. *Proceedings of 2011 IEEE PES General Meeting*, pages 1–7, 2011.