Recent Results of PMU Data Analytics by Exploiting Low-dimensional Structures

Meng Wang, Joe H. Chow, Pengzhi Gao, Yingshuai Hao, Wenting Li, Ren Wang

Abstract—The increasingly denser coverage of PMUs enables dynamic visibility into power systems. The large amounts of data obtained by PMUs impose significant challenges to data management and information extraction. The spatial-temporal blocks of PMU exhibit low-dimensional structures despite the high ambient dimensions. The low-dimensionality can be exploited to enable and simplify multiple PMU data management tasks under the same framework. This paper summarizes our recent results on data privacy enhancement, fast event identification, and further development of missing data recovery.

Random noise and quantization are applied to the measurements before transmission to compress data and enhance data privacy. To maintain the data accuracy for the operator, we develop novel methods to recover the original data from quantized measurements even when partial measurements are corrupted. A real-time event identification method is also proposed, based on the new idea of characterizing an event by the low-dimensional subspace spanned by the dominant singular vectors of the data matrix. Additional structures in the PMU data besides low-rankness are exploited to enhance the accuracy of missing data recovery. All the methods are evaluated on synthetic and historical PMU datasets.

Index Terms—phasor measurement units, data analytics, low rank, spatial-temporal data

I. INTRODUCTION

D ATA scarcity has been a major issue for power system monitoring. The current Supervisory Control and Data Acquisition (SCADA) systems typically provide measurements every 2-4 seconds, offering only a steady state view into the power system behavior. After the American Recovery and Reinvestment Act of 2009 [27], two thousand multichannel phasor measurement units (PMUs)[22] have now been installed North America [19]. PMUs can directly measure bus voltage phasors and line current phasors at synchronized time instants. With data rates of 30 or 60 samples per second, these PMUs produce terabytes of data daily, offering dynamic visibility into the power system.

This vast wealth of data is revolutionizing the operations of power systems. Conventional practice requires accurate modeling of power systems to compensate for the lack of measurements. The current abundance of data enables the development of data-oriented model-free methods in power system monitoring. These methods can improve the accuracy and reduce the latency of real-time situational awareness, which is of vital importance for building a reliable and efficient power grid.

As the coverage of PMUs in power systems becomes denser, it is natural to collect data from PMUs in electrically close area over a certain period of time and process this spatialtemporal blocks of PMU simultaneously. Interestingly, the PMU data block exhibits low-dimensional structures despite its high ambient dimension. The low-dimensional structure can be exploited to enable and simplify a variety of PMU data management tasks such as data compression, error correction, and feature extraction. We have obtained promising results on missing data recovery, detection of cyber data attacks by exploiting this structure. This paper reports our recent progress under this framework on further exploiting this structure to enhance PMU data privacy, identify system events, and enhancement and implementation of missing data recovery.

The remainder of this paper is organized as follows. Section II explains the low-dimensional property of PMU data. Section III discusses our proposed method to enhance the privacy of individual users without sacrificing the data accuracy at the operator. Section IV introduces our proposed method for fast event identification by representing the events by low-dimensional subspaces of the PMU data matrix. Section V records our further development of missing data recovery.

II. LOW-DIMENSIONAL STRUCTURES OF PMU DATA BLOCKS

We illustrate the low-rank property using the data from PMUs deployed in six substations in Central New York power system. The system diagram is shown in Fig. 1. Six PMUs records 11 voltage phasors and 26 line current phasors at a rate of 30 samples per second.



Fig. 1. Six PMUs in the Central Power System

Fig. 2 shows the voltage magnitudes of PMU measurements. A disturbance occurs at $t \approx 2.5$ second. Let a 37 by 600 complex matrix L contain the voltage and current phasors in 20 seconds. Each row corresponds to time series measurements in one channel. Each column corresponds to the PMU measurement at the same time instant. Singular value decomposition (SVD) is applied to matrix L,

$$L = U\Sigma V^H, \tag{1}$$

The authors are with the Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY.



Fig. 2. Recorded voltage magnitudes of PMU data

where Σ is the matrix with singular values on the main diagonal, and U and V are the left and right singular vectors, respectively, V^H denotes the conjugate transpose of V.

The singular values are shown in Fig. 3. The largest singular values is 894.5942, while the ninth largest one is 0.5930. We could use a rank-8 matrix to approximate L with a negligible error.



Fig. 3. Singular values of the PMU data matrix in decreasing order

III. DATA PRIVACY

Due to the mismatch with communication networks that were not designed to carry high-speed PMU data, data losses and data quality degradations happen quite often, especially in the Eastern interconnection [23]. On the other hand, data privacy in smart grids is receiving increasing attentions. To the best of our knowledge, only [28] considered protecting the privacy of PMU data through encryption.

In [12], we proposed a simple signal processing approach to achieve data compression and data privatization simultaneously for PMU data. Random noise are added to the measurements to protect data privacy, and quantization is applied afterwards to reduce the amount of information to transmit. A similar idea appeared recently for sensor networks [31]. Quantization can protect the data privacy, and adding noise is known to enhance the privacy level of individual users, measured by the differential privacy [8]. The privacy of each PMU is enhanced because an intruder can only observe highly quantized values even if it eavesdrops the data communication from the PMUs to the operator. Such data privacy is, however, achieved at a cost of reduced data accuracy. Our major contribution is a data recovery method using the quantized measurements from multiple PMUs such that the recovery error compared with the actual data diminishes when the number of PMUs increases. Therefore, the reduced data transmission, privacy enhancement of individual utilities, and the information accuracy for the central operator are achieved simultaneously.

Let $L^*, C^* \in \mathbb{R}^{m \times n}$ denote the actual data and the sparse additive errors in the measurements. Let $M^* = L^* + C^*$ denote the measurements that are partially erroneous. Let $N \in \mathbb{R}^{m \times n}$ denote the noise matrix which has i.i.d. entries with known cumulative distribution function $\Phi(z)$. Given a positive constant K, let [K] denote the set $\{1, ..., K\}$. A Klevel quantized noisy measurement Y_{ij} satisfies

$$Y_{ij} = \mathcal{Q}(L_{ij}^* + C_{ij}^* + N_{ij}), \ \forall (i,j),$$
(2)

where the operator Q maps a real number to one of the K labels. Given quantization boundaries $\omega_0 < \omega_1 < ... < \omega_K$,

$$Q(x) = l \quad \text{if } \omega_{l-1} < x \le \omega_l, \ l \in [K]. \tag{3}$$

(

$$Y_{ij} = l$$
 with probability $f_l(M_{ij}^*), \ \forall (i,j),$ (4)

where $\sum_{l=1}^{K} f_l(M_{ij}^*) = 1$, and $f_l(M_{ij}^*) = P(Y_{ij} = l | M_{ij}^*) = \Phi(\omega_l - M_{ij}^*) - \Phi(\omega_{l-1} - M_{ij}^*).$ (5)

The data recovery problem is stated as follows.

(P1) Given quantized observations Y, noise distribution Φ , and $\omega_l, l \in [K]$, can we recover the actual data L^* ?

To solve (P1), we proposed to estimate the unknown (L^*, C^*) using a constrained maximum likelihood approach. The negative log-likelihood function is given by

$$F_Y(X) = -\sum_{i=1}^m \sum_{j=1}^n \sum_{l=1}^K (\mathbf{1}_{[Y_{ij}=l]} \log(f_l(X_{ij}))), \quad (6)$$

where $\mathbf{1}_{[A]}$ denotes the indicator function that takes value '1' if A is true and value '0' otherwise. (6) is a convex function in X when the function f_l is log-concave in X_{ij} . We estimate (L^*, C^*) by (\hat{L}, \hat{C}) , where

$$(\hat{L}, \hat{C}) = \arg\min_{L, C} -\sum_{i, j} \sum_{l=1}^{K} \mathbf{1}_{[Y_{ij}=l]} \log(f_l(L_{ij} + C_{ij})),$$

s.t. $L + C \in \mathcal{S}_f,$ (7)

and the feasible set S_f is defined as

$$S_f := \{ X \in \mathbb{R}^{m \times n} : X = L + C, \|L\|_{\infty} \le \alpha, \|C\|_{\infty} \le \alpha,$$
$$\operatorname{rank}(L) \le r, \sum_{ij} \mathbf{1}_{[C_{ij} \neq 0]} \le s \}.$$
(8)

(7) is nonconvex due to the nonconvexity of S_f .

The recovery accuracy of the global minimizer of (7) is discussed in [12]. We provided the information theoretical limit of solving (P1) and demonstrated that the recovery accuracy of the global minimizer of (7) is order-wise optimal. We also proposed an approximate algorithm to solve (7). The algorithm alternatively updates L^* and C^* while fixing the other matrix.

We test our method on actual PMU dataset shown in Fig. 2. We set K = 5 with quantization boundaries $\omega_0 = -\infty$, $\omega_1 = -0.3$, $\omega_2 = 0$, $\omega_3 = 0.5$, $\omega_4 = 1.2$ and $\omega_5 = \infty$. Fig. 4 shows the original data, the quantized value with 5% of corruptions, and the recovered data of channels 4 and 20

from time 0 to 3.5 second. One can see that the details of the time series are masked in the quantized measurements. The recovered data are noisier than the actual data because we add noise to each channel with a noise level comparable to the signal level before quantization. Still, the overall trend of the time series are correctly recovered. If needed, the noise in the recovered data could be removed by applying a low-pass filter. Fig. 5 compares our method with simply using the mean value of each bin as the estimate and an existing data recovery method QRPCA [16]. Our method performs the best.



Fig. 4. Original, quantized, and recovered data of channels 4 and 20



Fig. 5. Relative recovery error on actual PMU data

IV. DISTURBANCE IDENTIFICATION

Faults in power systems can potentially lead to cascading failures if the operator does not take the right action in time. Various methods have been developed to identify events like generator trips [2], line outages [3], [26], [32], and oscillations [7], [18], [29]. Most of the event identification methods are model-based (see e.g., [24], [26], [32]), while data-driven event identification methods are developed recently, e.g., [1], [4], [5], [13], [15], [25], [30]. Data-drive approaches have the advantage that no modeling is needed and attract much attention with the increasing quantity of PMU data. The limitations of these approaches include the high requirement on the memory and storage, and the computational complexity, especially for massive datasets.

We proposed a novel data-driven fast identification method in [17]. The central idea is to represent a spatial-temporal block of PMU data by its low-dimensional row subspace. These subspaces can be easily obtained through Singular Value Decomposition. We showed that the subspace characterizes the dominant system dynamics and is robust to initial conditions of the events. Compared with directly using time series measurements, the low-dimensional subspace is a compact representation of the system events. For example, Fig. 6 shows the voltage magnitudes of two line trip events at different locations in New England. Although they look different due to different initial conditions, the angle of their corresponding subspaces of the data in one second after the event is only 6.04° . A small subspace angle indicates the similarity of two subspaces.



Fig. 6. Voltage magnitudes of three types of events in New England.

We proposed to represent different events by their corresponding low-dimensional subspaces of the data matrices. The proposed event identification method constructs a dictionary of these subspaces of different events offline. The size of the dictionary is significantly reduced compared with using PMU data directly due to the compactness of the subspace and its robustness to initial conditions. Once an event happens, the subspace of the data in 1-5 seconds after the event starting point would be computed and compared with the dictionary. The event would be identified as a type of event based on the most similar event in the dictionary. The event is located based on the PMU channel that has the most significant change after the event.

We validated this method on 380 simulated events, including load change, line trip and three-phase short circuit events, in the IEEE 16-generator 68-bus test system. Table I demonstrates 3 sample test cases. The minimum angles between the row subspace of the event and dictionary types (DT) are recorded. One can see that the minimum subspace angle is indeed achieved between an event and a dictionary atom with the same event type. Table II shows identification performance of these 380 cases, where Identification Accuracy Rate (IAR), Approximate Location Accuracy Rate (ALAR) and Exact Location Accuracy Rate (ELAR) are explained in details in [17]. In addition, we also verified our method on the historical datasets in New England. We tested on 21 events including load change, line trip and short circuit events with 100% IAR.

TABLE I MINIMUM SUBSPACE ANGLES BETWEEN A TEST CASE AND DICTIONARY ATOMS WITH DIFFERENT INITIAL CONDITIONS

DT	Load Change	Line Trip	Short Circuit
Load Change 1	1.03	10.63	18.50
Line Trip 1-30	11.80	0.67	14.32
Short Circuit 5	21.18	11.75	1.97

 TABLE II

 Identification results of 380 cases with different initial conditions

Type of event	IAR %	ELAR %	ALAR %
Line Trip	100	85	94 (among 3 buses)
Short Circuit	100	77	90 (among top 3 buses)
Load Change	100	46	90 (among top 5 buses)

V. MISSING DATA RECOVERY

Due to the early deployment of older PMUs and the mismatch with communication networks that were not designed to carry high-speed PMU data, data losses and data quality degradations are quite often, especially in the Eastern interconnection [23]. The unreliability of PMU data is a bottleneck for incorporating them into control-room operations. Missing PMU data recovery is an important pre-processing step for the subsequent applications.

In our paper [11], we proposed data recovery algorithms based on the low-rank property of the PMU data. Our developed online missing data recovery algorithm, termed as OLAP, can recover the missing points in streaming data. We also implemented OLAP as an action adaptor on openPDC [20] to facilitate the testing and potential adoption from utilities and ISOs. The openPDC is an open-source PMU data stream processor that is available for the users to build new functionalities transparently. Fig. 7 shows the interface of the OLAP algorithm running on openPDC.



Fig. 7. Software interface of our OLAP missing data recovery method implemented on openPDC.

To further enhance the data recovery accuracy, we exploited other spatial and temporal correlations in the PMU data matrix besides the low-rankness. In [10], we proposed a new model, termed as "union and sums of subspaces", to characterize the structures of the measurements under multiple disturbances. Each row of the PMU matrix represents the observations of a bus voltage or a line current across time. One disturbance may affect nearby voltages and currents, and the corresponding rows belong to the same low-dimensional subspace. Quantities affected by multiple overlapping disturbances belong to the sum of subspaces.

We proposed a new method to recover missing entries under this model. Its performance was evaluated in the IEEE 39bus New England Power System (Fig. 8). Assume two events happen consecutively due to the sudden drop of the active power generations of two generators. Generator 32 drops at t = 0.07 s, and generator 33 drops at t = 0.6 s. Fig. 9 shows the active power of generator 32 and 33 and the voltage magnitudes at bus 2, 6 and 12. We assume sixteen PMUs are installed in the power system. Each PMU measures the voltage phasor at the corresponding bus. We delete some measurements uniformly at random for each channel and test the recovery performance of Singular Value Thresholding (SVT) [6], a high-rank matrix completion (HRMC) method in [9], and our method. As shown in Fig. 10, our method outperforms SVT and HRMC when p_{avg} is below 0.2.



Fig. 8. IEEE 39 New England Power System [14][21]. We assume that sixteen PMUs are installed at bus 2, 4, 6, 8, 10, 12, 16, 18, 20, 22, 26, 33, 36, 37, 38 and 39.



Fig. 9. (a) Active power of generator 32 and 33; (b) Voltage magnitudes at bus 2, 6 and 12.



Fig. 10. Relative recovery errors of SVT, HRMC, and our method on the simulated PMU data.

We recently demonstrated that the temporal correlations in the data can be represented by the low-rank property of the Hankel matrix. Let vector $y_t \in \mathbb{C}^m$ denote the PMU data at instant t, where m is the number of PMU channels, then the PMU data from time 1 to t is represented by

$$Y(1,t) = \begin{bmatrix} y_1 & y_2 & \cdots & y_t \end{bmatrix}.$$
(9)

A Hankel matrix is defined as

$$Y_{H}^{k}(1,t) = \begin{bmatrix} y_{1} & y_{2} & \cdots & y_{t-k+1} \\ y_{2} & y_{3} & \cdots & y_{t-k+2} \\ \vdots & \vdots & \ddots & \ddots \\ y_{k} & y_{k+1} & \cdots & y_{t} \end{bmatrix},$$
(10)

where k is called a pencil parameter. In this case, the size of the Hankel matrix constructed with $y_1 \sim y_t$ is $km \times (t-k+1)$.

Leveraging the low-rank property of the Hankel matrix, we developed a new data recovery algorithm, referred to as OLAP-H. The advantage of OLAP-H over OLAP is that it can recover missing points even under extreme conditions when simultaneous and consecutive data losses happen. OLAP-H is also more robust to measurement noise.

Fig. 11 shows the recorded voltage phasors of five PMUs provided by Korea Electric Power Corporation. If 10% measurements are erased randomly, an example of the recovered data with OLAP method is shown in Fig. 12, while the recovered data with OLAP-H method is shown in Fig. 13. Compared with the recovered data by OLAP, no obvious spikes exist in the recovered data by OLAP-H.



Fig. 11. Recorded voltage magnitudes of five PMUs



Fig. 12. Recovered data with OLAP method



Fig. 13. Recovered data with the developed method

VI. CONCLUSION AND DISCUSSIONS

This paper describes a framework for PMU data management and analysis by exploiting the low-dimensional structures of spatial-temporal PMU data blocks. Exploiting this lowdimensional structure can enable and simplify multiple data processing tasks such as the reconstruction of missing points and the correction of bad measurements and cyber data attacks. This paper reports our recent work in this direction to recover data from highly quantized and erroneous measurements, identify disturbance in real time, and recover data losses. It is interesting to explore the application of this structure in other tasks such as model identification and oscillation detection.

ACKNOWLEDGEMENT

This research is supported in part by expired and current grants including NYSERDA-36653, EPRI-00-10005909, NSF ECCS-1508875 and the ERC Program of NSF and DoE under the supplement to NSF Award EEC-1041877 and the CURENT Industry Partnership Program. We thank New York Power Authority, ISO New England, Korea Electric Power Corporation, and Tennessee Valley Authority for providing historical PMU datasets.

REFERENCES

- R. Agrawal and D. Thukaram, "Identification of fault location in power distribution system with distributed generation using support vector machines," in *Proc. IEEE PES Innovative Smart Grid Technologies* (*ISGT*), 2013, pp. 1–6.
- [2] J. N. Bank, R. M. Gardner, J. K. Wang, A. J. Arana, and Y. Liu, "Generator trip identification using wide-area measurements and historical data analysis," in 2006 IEEE PES Power Systems Conference and Exposition, 2006, pp. 1677–1681.
- [3] P. Bhui and N. Senroy, "Online identification of tripped line for transient stability assessment," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 2214–2224, 2016.
- [4] T. S. Bi, X. N. Song, J. T. Wu, and Q. X. Yang, "Novel method for disturbance identification in power systems," in 2006 IEEE Power Engineering Society General Meeting, 2006.
- [5] A. Bykhovsky and J. H. Chow, "Power system disturbance identification from recorded dynamic data at the northfield substation," *International Journal of Electrical Power & Energy Systems*, vol. 25, no. 10, pp. 787 – 795, 2003.
- [6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
- [7] M. L. Crow and A. Singh, "The matrix pencil for power system modal extraction," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 501–502, 2005.
- [8] C. Dwork and A. Smith, "Differential privacy for statistics: What we know and what we want to learn," *Journal of Privacy and Confidentiality*, vol. 1, no. 2, p. 2, 2010.
- [9] B. Eriksson, L. Balzano, and R. Nowak, "High-rank matrix completion," in *Proc. International Conference on Artificial Intelligence and Statistics*, 2012, pp. 373–381.
- [10] P. Gao, M. Wang, J. H. Chow, M. Berger, and L. M. Seversky, "Matrix completion with columns in union and sums of subspaces," in *Proc. IEEE Global Conference on Signal and Information Processing* (*GlobalSIP*) 2015, 2015, pp. 785–789.
- [11] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," *IEEE Trans. Power Systems*, vol. 31, no. 2, pp. 1006 – 1013, 2016.
- [12] P. Gao, R. Wang, M. Wang, and J. H. Chow, "Low-rank matrix recovery from quantized and erroneous measurements: Accuracy-preserved data privatization in power grids," in *Proc. Asilomar Conference on Signals*, *Systems, and Computers*, 2016.
- [13] A. K. Ghosh and D. L. Lubkeman, "The classification of power system disturbance waveforms using a neural network approach," *IEEE Trans. Power Del.*, vol. 10, no. 1, pp. 109–115, 1995.

- [14] IEEE 10 Generator 39 Bus System, http://sys.elec.kitami-it.ac.jp/ueda/ demo/WebPF/39-New-England.pdf.
- [15] T. Y. Ji, Q. H. Wu, L. Jiang, and W. H. Tang, "Disturbance detection, location and classification in phase space," *IET Generation, Transmission Distribution*, vol. 5, no. 2, pp. 257–265, 2011.
- [16] A. S. Lan, C. Studer, and R. G. Baraniuk, "Matrix recovery from quantized and corrupted measurements," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4973–4977.
- [17] W. Li, M. Wang, and J. H. Chow, "Fast event identification through subspace characterization of pmu data in power systems," accepted to IEEE Power and Energy Society (PES) General Meeting, 2017.
- [18] G. Liu, J. Quintero, and V. Venkatasubramanian, "Oscillation monitoring system based on wide area synchrophasors in power systems," in Bulk Power System Dynamics and Control-VII. Revitalizing Operational Reliability, 2007 iREP Symposium, 2007, pp. 1–13.
- [19] North America Synchrophasor Initiative(NASPI), https: //www.naspi.org/Badger/content/File/FileService.aspx?fileid= 49CC0BEB3E3C36F3BF9C7930E0FFFD1B.
- [20] openPDC, http://openpdc.codeplex.com.
- [21] A. Pai, Energy function analysis for power system stability. Springer, 1989.
- [22] A. Phadke and J. Thorp, Synchronized phasor measurements and their applications. Springer, 2008.
- [23] A. Silverstein and J. E. Dagle, "Successes and challenges for synchrophasor technology: An update from the north american synchrophasor initiative," in *Proc. Hawaii International Conference on System Science (HICSS)*, 2012, pp. 2091–2095.
- [24] M. J. Smith and K. Wedeward, "Event detection and location in electric power systems using constrained optimization," in *Proc. IEEE Power Energy Society General Meeting*, 2009, pp. 1–6.
- [25] Y. Song, W. Wang, Z. Zhang, H. Qi, and Y. Liu, "Multiple event analysis for large-scale power systems through cluster-based sparse coding," in *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2015, pp. 301–306.
 [26] J. E. Tate and T. J. Overbye, "Line outage detection using phasor angle
- [26] J. E. Tate and T. J. Overbye, "Line outage detection using phasor angle measurements," *IEEE Trans. Power Syst.*, vol. 23, no. 4, pp. 1644–1652, 2008.
- [27] The American Recovery and Reinvestment Act, https://www.whitehouse. gov/recovery/about, 2009.
- [28] Y. Tong, J. Sun, and K. Sun, "Privacy-preserving spectral estimation in smart grid," in *Proc. IEEE International Conference on Smart Grid Communications(SmartGridComm)*, 2015, pp. 43–48.
- [29] D. Trudnowski, J. Johnson, and J. Hauer, "Making prony analysis more accurate using multiple signals," *IEEE Trans. Power Syst.*, vol. 14, no. 1, pp. 226–231, 1999.
- [30] W. Wang, L. He, P. Markham, H. Qi, Y. Liu, Q. C. Cao, and L. M. Tolbert, "Multiple event detection and recognition through sparse unmixing for high-resolution situational awareness in power grid," *IEEE Trans. Smart Grid*, vol. 5, no. 4, pp. 1654–1664, 2014.
- [31] S. Xiong, A. D. Sarwate, and N. B. Mandayam, "Randomized requantization with local differential privacy," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2189–2193.
- [32] H. Zhu and G. B. Giannakis, "Sparse overcomplete representations for efficient identification of power line outages," *IEEE Trans. Power Syst.*, vol. 27, no. 4, pp. 2215–2224, 2012.