Synchrophasor Missing Data Recovery via Data-Driven Filtering

Stavros Konstantinopoulos, *Student Member, IEEE*, Genevieve M. De Mijolla, *Member, IEEE*, Joe H. Chow, *Life Fellow, IEEE*, Hanoch Lev-Ari, *Life Fellow, IEEE*, and Meng Wang, *Member, IEEE*

***Abstract*—To enhance reliability and observability, power systems in North America have installed a significant number of Phasor Measurement Units (PMUs) to monitor dynamic behaviors. For real-time applications, the PMU data are streamed via the Internet from the substations to the phasor data concentrators, in the control centers. The transmission of the PMU data however, is not always reliable and can be subjected to quality issues and losses due to latency and equipment malfunctions. In this paper, a temporal version of the OnLine Algorithm for PMU data processing (OLAP) is proposed to recover the missing data. The algorithm is geared toward prolonged data outages and especially signals exhibiting significant temporal patterns. The method is connected to adaptive filtering and a necessary stability criterion for the algorithm is derived.The method is compared against several low rank and streaming data recovery methods to evaluate its effectiveness.**

***Index Terms*—Synchrophasor, PMU, low rank recovery, adaptive filtering, OLAP.**

I. Introduction

The increasing adaptation of PMU technologies across the Northern American power systems has provided valuable observability into the dynamics of the grid. This presents numerous opportunities for novel control room tools and applications, which enhance the wide area control capabilities of the grid and the operator's situational awareness. For an overview of such applications the interested reader can refer to [1], [2]. PMU applications rely on the time synchronization of the data and the high sampling rate that allows relevant system dynamics to be observable. Application such as dynamical state estimation [3], [4] and modal analysis [5], [6] offer opportunities for more efficient operation and analysis of the grid, but can be significantly hindered if there are significant amounts of missing data from the phasor data streams.

Missing PMU measurements can be a consequence of equipment malfunction or the more commonly, network latency. Data transmission and collection is typically done in a hierarchical manner. The PMU data collected from a substation are transmitted to a local Phasor Data Concentrator (PDC) owned by a transmission company. The collected data are then sent as a single stream to the regional PDC at the independent system operator (ISO) level. Such multi-stage transmission incurs potentially high latency. In order for the high sampling rate of the PMUs to be relevant in real time applications, a measurement is transmitted further up the commutation chain if it is received within a predetermined time window. If not, the data point is deemed as missing and the remainder is sent to the regional PDC.

Recognizing the need for high-quality PMU datasets, many researchers have proposed algorithms for synchrophasor missing data recovery. In [3] the authors incorporate the data recovery problem in the dynamic state estimator. This approach uses an Extended Kalman Filter that incorporates the data outage in the model, such that the problem is transformed to a nonlinear constrained optimization, and is solved by metaheuristics. In [7] the authors utilize Kalman filtering and smoothing and incorporate a quadratic predictor, to provide recovery of the missing data. Learning methods such as neural networks and ensemble learning appear to attract much attention in PMU recovery literature. In [8] the problem is tackled utilizing Gated Recurrent Neural Networks to predict the future states. In [9] graph convolutional recurrent layers, embedding the power network structure into the neural network, are utilized to recover missing data, exploiting spatial and temporal patterns in the data. In [10] generative adversarial learning is utilized to train a generator network that produces data with the same distribution as the PMU data, to fill in missing measurements. In [11] author utilized random vector functional link models, with extreme learning machines, to predict and validate missing measurements, prior to performing dynamic stability assessment. In contrast, the method proposed in this paper requires no offline training and can effectively adapt to changing grid dynamics. In [12] the authors formulate the low-rank matrix recovery problem and solve it utilizing the Alternating Direction Method of Multipliers which can be parallelized and thus is scalable. In addition, they investigate simultaneous missing points across all PMU channels

Manuscript received August 16, 2019; revised December 28, 2019; accepted February 24, 2020. Date of publication April 8, 2020; date of current version August 21, 2020. This work was supported in part by the Engineering Research Center Program of the NSF and the DOE under the Supplement to NSF under Award EEC-1041877, in part by the CURENT Industry Partnership Program, and in part by the Rensselaer Polytechnic Institute. Paper no. TSG-01201-2019. *(Corresponding author: Stavros Konstantinopoulos.)*

Stavros Konstantinopoulos, Joe H. Chow, and Meng Wang are with the Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: konsts@rpi.edu; chowj@rpi.edu; wangm7@rpi.edu).

Genevieve M. De Mijolla is with the Department of Power Economics, General Electric Energy Consulting, Schenectady, NY 12345 USA (e-mail: genevieve.demijolla@ge.com).

Hanoch Lev-Ari is with the Department of Electronics and Communication Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: levari@ece.neu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2020.2986439

and propose an Order and Cut-Column Reshaping Method. In [13] the authors utilize Bayesian estimation in conjunction to the matrix completion formulation for effective recovery of distribution level synchrophasor data. These methods have high computation burdens and thus not suitable for real time. Finally, in [14] block matrix recovery is applied to PMU data by utilizing nonlinear Hankel structures. The method seems to perform well especially during disturbances, but is geared towards processing smaller time windows.

In [15] a spatial OnLine Algorithm for PMU data processing (denoted as OLAP-s) has been proposed, based on low-rank matrix completion ideas such as the Singular Value Thresholding (SVT) method [16]. Given a matrix with incomplete data, low-rank matrix methods aim to recover a full matrix, retaining the same known elements, but with the minimum possible rank. However, the exact matrix recovery with minimal rank problem is NP-hard. Methods such as the SVT introduce convex relaxations to make the solution tractable. OLAP-s translates these concepts in an implementation geared towards online application, being light weight, efficient and aiming to exploit patterns present in the PMU data. In contrast to other established streaming subspace methods such as GROUSE [17] and PETRELS [18], the threshold of OLAP-s is used to adapt the rank of the subspace used for data recovery. In this paper, we adapt the OLAP-s method to a temporal OLAP method, called OLAP-t, by reformulating the method, to exploit time response patterns as captured in the temporal subspace of the data matrix. This allows efficient recovery of irregular periodic events and even total outage of PMU streams can be overcome for short periods of time. An advantage of the OLAP-t is that although temporal characteristics are exploited, spatial relationships are still implicitly captured by the dominant temporal features. Another highlight of the OLAP algorithms is their speed. Since no iterations are necessary, the method is suitable for real time.

The remainder of the paper is organized as follows. Section II contains descriptions of the OLAP method. Section III discusses the OLAP-t's connection to filtering and provides a stability condition and a detailed parameter sensitivity analysis. In Section IV the performance of OLAP in ambient, post disturbance and oscillatory events as captured in real PMU data, will be presented and discussed.

II. Low-Rank Data Recovery and OLAP

This section describes the motivation of low-rank missing data recovery methods and their connection to OLAP [15], [16]. We initially start by defining the matrix of PMU data with dimensions $p \times m$, where p denotes the number of samples contained in a data window and m denotes the number of channels. That is, each row corresponds to measurements from the channels captured at the same time stamp as determined by the GPS clock. To assemble the data matrix we define

$$\alpha_{k,j} = \begin{bmatrix} y_{k-p+1,j} \\ y_{k-p+2,j} \\ \vdots \\ y_{k,j} \end{bmatrix} \tag{1}$$

This vector contains the measurements of the j^{th} channel from time $k - p + 1$ to time k (p data points). We can assemble these vectors as to form a data matrix

$$M_k = \begin{bmatrix} \alpha_{k,1} & \alpha_{k,2} & \dots & \alpha_{k,m} \end{bmatrix} \tag{2}$$

Unless otherwise stated, it is assumed that the M_k matrix has missing data. In this work, the matrices to be recovered will be comprised by voltage and current measurements. Phasor measurements, by nature, present low rank patterns. The power system, although it is subject to non-linearities, tends to present much lower rank behaviors (strong correlation). Voltages and currents are implicitly connected through circuit equations, thus the PMU data usually have much lower numerical ranks than the full rank of the measurement matrix. Thus, low rank matrix recovery approaches are an appropriate choice when tackling such datasets.

A. The Temporal OLAP Algorithm

The OLAP-s algorithm proposed in [15] is a lightweight real-time missing data recovery method, designed specifically for PMU data processing. While other missing data recovery methods that operate in real-time (such as GROUSE [17], PETRELS [18], and MOUSSE [19]), assume a fixed-dimension subspace, the dimension of the subspace can change based on the dynamics of the dataset, which dictate the effective numerical rank. OLAP-s tracks the dimensionality of the matrix being recovered by updating the dominant singular vectors of the matrix at every new sampling time.

For a description of the original OLAP, a reader can refer to [15]. The following algorithm has close connections to the original algorithm which will be highlighted. For the proposed temporal extension (OLAP-t), it is also assumed that there are no missing data points in the initialization. The initialization is accomplished by performing a Singular Value Decomposition (SVD) on an initial data matrix $M_0 \in \mathbb{R}^{p\times m}$, with p being the length of the sliding window of data specified by the user, as

$$M_0 = U\Sigma V^T \tag{3}$$

where $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{m\times m}$ are unitary matrices, with V^T the transpose of V, and $\Sigma \in \mathbb{R}^{p\times m}$ is the diagonal singular value matrix. The initialization algorithm then keeps only the dominant singular values which are greater than $\gamma_{\text{err}} \times \sigma_1$, with σ_1 the largest singular value of M_0, and $\gamma_{\text{err}} \in [0, 1]$ a threshold of relative approximation error set by the user. This process is equivalent to performing a low-rank matrix approximation of M_0

$$M_0^r = U^r\Sigma^r\left(V^r\right)^T \tag{4}$$

by truncating the matrix M_0 to a specific rank r. An important observation is that the left singular vectors in U^r represent the temporal variation of the data and the right singular vectors in V^r represent the spatial relationship between the different PMU channels. Once initialized, as each new set of sample points is received, the OLAP-t algorithm, with M_k being the most current data matrix, proceeds with the following steps:

1) Receive the new data as a row vector $\beta \in \mathbb{R}^{1\times m}$ which may contain missing points to be recovered.

2) Compute β_Ψ and V^r_Ψ, with Ψ the index of the observed entries. Thus β_Ψ is the β vector without its missing entries, and U^r_Ψ is the U^r matrix without the rows corresponding to the indices of Ψ.
3) Use the least-squares method to find x such that $U^r_\Psi x - \beta_\Psi$ is minimized.
4) Compute the missing data entries using $U^r_{\Psi^c}x$, with $U^r_{\Psi^c}$ the dominant singular vector matrix for erasures. Form the new row $\hat{\beta}$ by filling in β with the missing data.
5) Update M_k by dropping the oldest data row and adding the newly computed row $\hat{\beta}$ and denotes the new data matrix as M_{k+1}.
6) Perform SVD of the updated M_{k+1} matrix.
7) Update U^r to the dominant singular vectors corresponding to singular values above $\gamma_{\text{err}} \times \sigma_1$, with σ_1 the largest singular value of the new matrix M_{k+1}. Return to Step 1.

Thus the OLAP-t algorithm fills in the missing data points by using the temporal information in the left singular vectors U. It operates in real-time, efficiently updating the sliding window of data at every new sampling instant, filling in missing data points and updating the dominant singular values of the subspace being considered. The parameters to be chosen are the sliding window length (p) and the numerical rank threshold γ_{err}, which allow this algorithm to capture events with time varying dynamics, making the algorithm more accurate during disturbances, when abrupt rank changes occur. To highlight the differences with OLAP-s, the algorithm can be summarized as

1) For t $= 1, 2, 3 \ldots$ do
2) Receive new data $\beta \in C^{w\times 1}$ with erasures
3) Compute $u^* = \text{argmin}_u ||V^r_\psi u - \beta_\psi||_2$
4) Estimate missing entries from $V^r_\psi u^*$

OLAP-s recovers the missing entries in the row of incoming measurements from the existing values (β_ψ) by minimizing the Frobenius norm between the linear combination of the first r columns and rows ψ (rows with complete data). OLAP-s specifically exploits the spatial patterns in the data (correlation) to estimate the missing points in the measurements. The method demonstrated very good performance in disturbance and ambient data, but can face issues in complete spatial information outage (no measurements in a time step) and when long temporal patterns are present in the data but due to the use of only the V matrix, may not be captured.

III. The Temporal OLAP as a Filter

The process developed in Section II also describes an AutoRegressive (AR) process where each of the past measurements in the window, is linearly combined to produce the estimate. From the previous formulation, it is evident that the only possible row that a missing entry can exist is the latest row of measurements. Thus we partition $U \in C^{p\times w}$ as mentioned in the previous section

$$U = \begin{bmatrix} U_1 & \multirow{2}{*}{U_0} \\ U_2 & \end{bmatrix} \tag{5}$$

where $U_0 \in C^{p\times w-r}$ denotes the less significant singular vectors, and $U_2 \in C^{1\times r}$ is the last row of the first r left singular vectors. $U_1 \in C^{p-1\times r}$ captures the one-step back temporal singular vectors (corresponding to known measurements). If the $p-1$ past measurements in step k are

$$Y = \begin{bmatrix} y_{k-p+1} \\ \vdots \\ y_{k-1} \end{bmatrix} \tag{6}$$

then

$$\hat{y}_k = U_2\left(U_1^T U_1\right)^{-1} U_1^T Y \tag{7}$$

Thus OLAP solves a least-squares minimization problem with a closed from solution as seen in (7). So there are clearly defined coefficients, mapping all the previous measurements in the estimate. These coefficients are calculated as

$$c = U_2\left(U_1^T U_1\right)^{-1} U_1^T = \begin{bmatrix} c_1 & c_2 & \ldots & c_{p-1} \end{bmatrix} \tag{8}$$

The missing data recovery can be obtained from the prediction equation

$$\hat{y_k} = c_1 y_{k-p+1} + \cdots + c_{p-1} y_{k-1} \tag{9}$$

The predictions will be utilized to fill in potential missing data in the synchrophasor measurements. Compared to a simple AR process used for prediction based on past measurements, OLAP-t exploits the low-rank temporal behavior of the measurement matrix (i.e., the system). As a result, the algorithm has increased robustness to handle high noise levels while being sensitive enough to rank changes, due to the time-varying subspace being considered. This makes the filter fast in adapting to system dynamics while keeping a high level of recovery accuracy.

A. Filter Stability Bound

Because OLAP-t can be formulated as a filter, this section provides an initial stability bound that can serve as an indication of potential instability. From the singular value decomposition the U matrix is orthonormal, implying $||U_2||_2 \leq 1$. Using the identity and using the partitioning of the U matrix introduced beforehand the following holds

$$\begin{pmatrix} U_1 \\ U_2 \end{pmatrix}^T \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = I_r = \left(U_1^T U_2^T\right)\begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = U_1^T U_1 + U_2^T U_2 \tag{10}$$

it follows that

$$U_1^T U_1 = I - U_2^T U_2 \tag{11}$$

The right hand side of (11) is an identity matrix minus a rank one matrix. The inverse of this matrix has a closed form solution

$$\left(I - U_2^T U_2\right)^{-1} = I + \frac{1}{1 - ||U_2||_2^2} U_2^T U_2 \tag{12}$$

So the coefficients c in (8) take the form of

$$c = \frac{1}{1 - ||U_2||_2^2} U_2 U_1^T \tag{13}$$

Due to the nonlinear structure of the filter, only some basic BIBO criteria can be derived. Assuming that all the previous

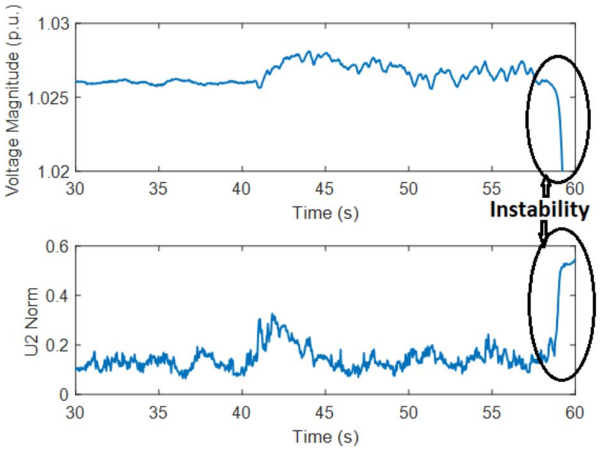

Fig. 1. Filter Instability and $||U_2||_2$.

measurements are present and that they are bounded, the estimate has an upper bound (for channel j)

$$||\hat{y}_{k,j}||_\infty \leq ||c||_1 \max ||Y_j||_\infty \tag{14}$$

where Y_j is the column j of Y containing the channel's past measurements. I.e., the maximum value of any prediction produced by the filter is upper bounded by the maximum value present in the past measurements, stretched by the 1st norm of the coefficient vector. So, c acts as a gain that maps past measurements in the estimate: if the norm c remains bounded, stability can be guaranteed. However, we also can derive a closed form expression for $||c||_2^2$ as

$$\begin{aligned} ||c||_2^2 = cc^T &= \left(\frac{1}{1-||U_2||_2^2}\right)^2 U_2 U_1^T U_1 U_2^T \\ &= \frac{1}{\left(1-||U_2||_2^2\right)^2} U_2\left(I - U_2^T U_2\right) U_2^T \\ &= \frac{||U_2||_2^2}{1-||U_2||_2^2} \end{aligned} \tag{15}$$

For stability $||c||_1 \leq 1$ is needed, thus also $||c||_2 \leq 1$ is implied. This provides a limit of $||U_2||_2 \leq 1/\sqrt{2}$. Fig. 1 shows an unstable case of recovery, and the corresponding magnitude of $||U_2||_2$ is presented. The results indicate that the condition is necessary but not sufficient.

B. Threshold Sensitivity Analysis and Stability

The effective (numerical) rank of the approximation is determined by the rank threshold γ_{err}. A singular value less than γ_{err} is considered as zero. The threshold needed for each PMU signal recovered might differ based on the magnitudes of the signal. An effective way to make the threshold more robust is normalization of the data, so the magnitudes remain in a consistent range. In addition, γ_{err} gives us a sense of the expected error of the approximation of the recovered data. The sensitivity is dictated mainly by three factors: dynamics in the dataset (events happening in the examined window, imply higher rank), preprocessing and normalization of data, and the length of the window. In ambient conditions, the rank is usually robust because most of the low frequency dynamics (slow load and generation movement) are captured in the first few eigenvectors of the subspace. However, the situation changes if disturbances occur in the examined window. Consider the event depicted in Fig. 6. During the event, the less significant singular values, will increase. In Fig. 2 shows the numerical rank for different threshold values.

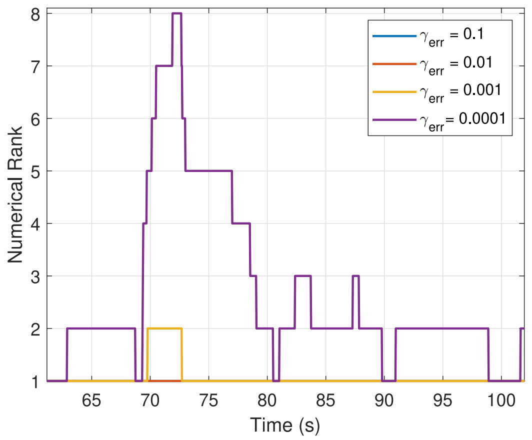

Fig. 2. Numerical Rank of Switching Event for Different Threshold Values.

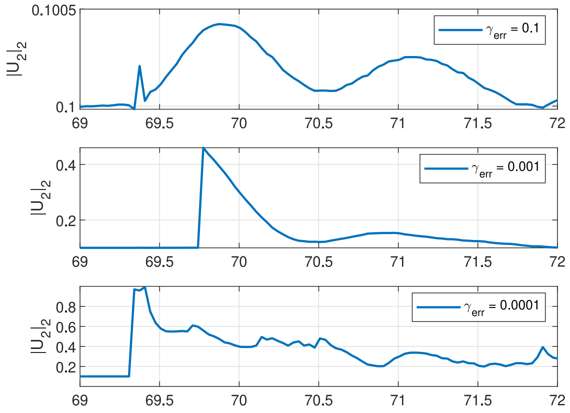

Fig. 3. Euclidean Norm of Last Row of U Matrix.

The selection of the threshold is illustrated by Fig. 3. For a threshold of 0.1 (top plot), $||U_2||_2$ is small and the filter is stable, while there may be some compromise in the accuracy of the recovered data. For a threshold of 0.0001 (bottom plot), the filter becomes unstable, as shown in Fig. 1. In this investigation, the threshold of 0.001 seems to be quite optimal.

The analysis in Fig. 3 uses 100 data point windows. However, when selecting γ_{err} and window length simultaneously, additional considerations are required. Fig. 4, depicts the maximum absolute error, for different window sizes and threshold values. The maximum absolute error was chosen to depict the worst case error.

In Fig. 4, the errors were plotted from a tolerance of 0.00001 to 0.1, since the filters become unstable with smaller tolerances. The dataset used for this test is a capacitor switching event (presence of abrupt change of rank). As one can note, shorter windows achieve in general lower errors if the threshold is tuned accordingly. Fig. 4 indicates that the lowest recovery error can be achieved using a tolerance between 0.001 and 0.01, and a data window of 50 to 100 points.

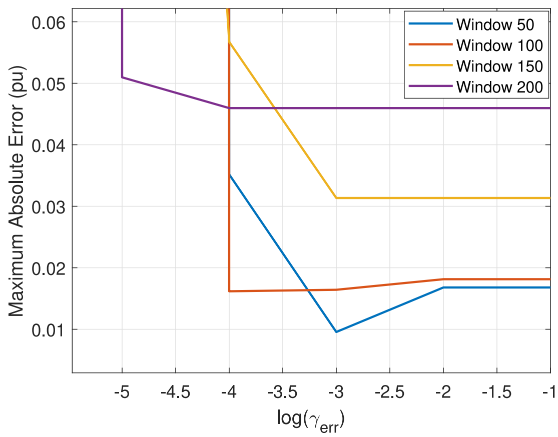

Fig. 4. Maximum Absolute Error for Variable Threshold Values and Window Lengths.

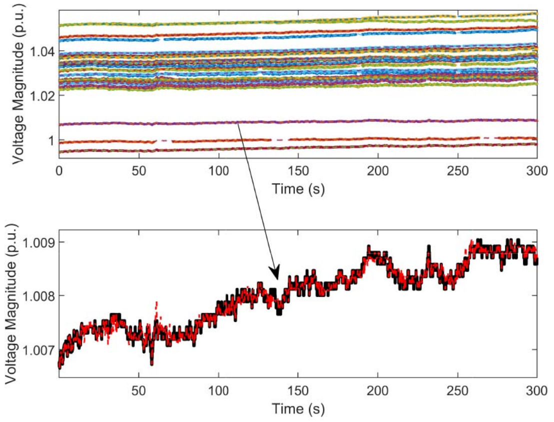

Fig. 5. Ambient Condition, 10 Point Random Outages (Recovered in Dashed Red Line).

TABLE I
ERROR FOR INCREASING CONSECUTIVE POINT OUTAGES

Consecutive Points Drop	Root Mean Square Error (kV)	Mean Abs. Error (kV)	Missing Data (%)	Max Abs. Error (kV)
10	0.0452	0.0334	65.47	0.4251
20	0.0469	0.035	65.12	0.5417
30	0.0402	0.0562	65.31	0.5122
40	0.0394	0.0534	64.98	0.4923
50	0.0437	0.0611	65.33	0.3950
100	0.0476	0.0647	64.38	0.5104
200	0.0703	0.1093	64.20	0.485

TABLE II
AMBIENT CASE COMPARISON OF MEAN ABSOLUTE ERROR IN KV FOR INCREASING CONSECUTIVE POINT OUTAGES (12% MISSING DATA)

Consecutive Points Drop	OLAP-t	OLAP-s	PETRELS	GROUSE	SVT
5	0.0301	0.0298	0.0305	0.2321	0.0350
20	0.0301	0.0297	0.0308	0.3625	0.0355
100	0.0306	0.0302	0.0314	0.2342	0.0377

TABLE III
AMBIENT CASE COMPARISON OF MEAN ABSOLUTE ERROR IN KV FOR INCREASING CONSECUTIVE POINT OUTAGES (30% MISSING DATA)

Consecutive Points Drop	OLAP-t	OLAP-s	PETRELS	GROUSE	SVT
5	0.0312	0.0309	0.0322	0.1817	0.0351
20	0.0315	0.0312	0.0320	0.1554	0.0358
100	0.0327	0.0324	0.0334	0.1384	0.1858

TABLE IV
AMBIENT CASE COMPARISON OF MEAN ABSOLUTE ERROR IN KV FOR INCREASING CONSECUTIVE POINT OUTAGES (50% MISSING DATA)

Consecutive Points Drop	OLAP-t	OLAP-s	PETRELS	GROUSE	SVT
5	0.0323	0.0320	0.0337	0.2148	0.0351
20	0.0329	0.0326	0.0336	0.1547	0.0359
100	0.0345	0.0342	0.0356	0.1840	0.0825

Finally, subtraction of means from the data was examined. DC biases (i.e., nominal values) in normalized PMU data are around 1 pu for voltages and up to 10 pu for current. The presence of such moderate values does not cause rank determination issues even for ambient signals, with variations of typically 1 to 5%. The numerical rank determination was the same with or without the biases.

IV. TESTING ON PMU DATA

The OLAP-t algorithm has been applied to a large number of PMU data sets. For discussion purposes, these PMU signals are divided into four categories. Note that the number of channels (PMUs) analyzed is 139 in the first three cases and 39 in the last. The PMU data reporting rate is 30 Hz.

A. Ambient Condition Testing

Ambient conditions refer to a power system operating without large disruptions, such as a generator trip. Over 90% of the PMU data are ambient data. Random isolated measurement outages are a common occurrence but hardly challenging for most data recovery methods. However, long windows of consecutive missing points can be a more appropriate and illustrative way to evaluate a recovery algorithm. In order to demonstrate the effectiveness of OLAP-t, random outages will be introduced in the data, with increasing length. For the first test, the window of OLAP-t is set to 200 measurement points, the probability of occurrence to 10% and the consecutive points lost to 10. The introduced missing points were 65.72% of the total measurements in the dataset. Fig. 5 shows the data recovery for the ambient case. One of the challenges of such cases is that quantization, as the recovery produces only a smoothed version of the signal. This, however, will increase the error metrics as presented in Table I. So, the errors cannot be attributed to failure of OLAP to recover the signals. The errors across all considered cases are kept to below 0.01 pu (345 kV voltage base) even for prolonged outages.

In order to assess the effectiveness of OLAP-t, the results will be compared in the same data set against the original OLAP-s, PETRELS and GROUSE. In addition, the well-known SVT algorithm will be used as a base line. Note that, SVT is normally not meant for streaming recovery, so the data are segmented into 500 (optimal recovery after testing parameters) time step blocks and then recovered individually. For the test, we increase the chance of outage occurrence from 10-50% and for 5, 20 and 100 consecutive drops.

TABLE V

AMBIENT CASE: COMPARISON OF AVERAGE COMPUTATION TIME PER WINDOW IN MS FOR INCREASING CONSECUTIVE POINT OUTAGES

Missing Data (%)	OLAP-t	OLAP-s	PETRELS	GROUSE	SVT
12	1.8840	2.3772	35.6210	2.2057	145.5313
30	2.0288	2.3731	37.7949	2.2287	207.8846
50	2.1331	2.3747	34.2083	2.2441	294.7143

GROUSE in general seems to have the highest error among the compared methods. On the other hand, PETRELS seems to have considerably lower error and is comparable with the SVT, OLAP-s and OLAP-t. Note also that in this dataset, there is a window of 3 data points that none of the methods aside from OLAP-t recovered. The reason is that all the compared methods require at least 1 measurement present in a time step to produce an estimate. Since, no ground truth was available the errors were not considered in the above calculations. However, this highlights the usefulness of OLAP-t, as it overcomes the shortcoming of methods with similar recovery philosophies and avoids the potential need to switch methods if such cases occur. The next issue is computation time, which is summarized in Table V. OLAP-t is the fastest in average processing time, recovering each window within around 2 ms. PETRELS has the same accuracy, but takes 17 times as much computation to produce comparable results. GROUSE on the other hand, has comparable speed but lacks the same accuracy level. Most methods do not present significant increase in computation time with the number of outages. These tests were performed on a PC with an Intel i7 CPU at 3.6 GHz and 32 GB of RAM.

B. Post Disturbance Recovery

As a second case, the recovery of missing data right after a capacitor switching event will be examined. This case is of particular interest, since some methods face difficulties with abrupt changes in the signal. In Fig. 6 the voltage profile after the event and the recovery results can be noted. It is worth pointing out that the missing points were not artificially introduced but were already present in the original data set. The recovered signal can be noted in red and the original in blue. The recovery result for OLAP-t was successful since the response is still low rank and captured in the temporal singular vectors.

For the same dataset, similar randomized drop testing can be performed. In this case, the introduced number of outages is limited to 5 points with increasing chance of occurrence. As shown in Table VI, OLAP-t performs well against the other methods. The relative errors remain about the same with the ambient dataset tests, with GROUSE having the highest errors. The errors on average tend to increase with the percentage of missing data. OLAP-t seems to have the smallest increase across all the methods (except SVT which remains roughly constant).

C. Recovery of Signal With Irregular Periodicity

In this section, oscillatory and periodic events will be used to evaluate the OLAP-t algorithm. Initially, an event that was examined in [20] will be presented to illustrate the capabilities of OLAP-t in recovering signals with periodic behavior. The voltage profiles during the event can be noted in Fig. 7. The examined channel is the one indicated in red. The same behavior was observed in a few other channels but with smaller amplitude variation. The channels with similar oscillatory patterns were highlighted in magenta. In Fig. 7 the artificial outage of 5 seconds is indicated. The outage was introduced at the second period of the oscillation event with the intention of recovering the periodic pattern. Fig. 8 shows the recovery of the data outage for increasing window sizes of OLAP-t. The recovery can be deemed successful with windows far less than the period of the oscillation event (around 600 data points). However, smaller windows required more tuning of the threshold parameter to produce satisfactory results, while the window sizes of 400 or more points seemed to be more robust towards small threshold variations. Smaller windows usually result in features of the signals to be captured at higher rank SVs thus increasing the threshold sensitivity. In general, due to the low rankness of the event, i.e., the correlated behavior of the voltages in the system, there is enough information from the left singular vectors to recreate the major temporal behavior of the signal. The reconstruction also has two additional characteristics. Initially, the recovered data appear to be smoother, since the noise subspace of the signals is discarded and second, there is an one-sample lag in the reconstruction, which can be rectified in post-processing.

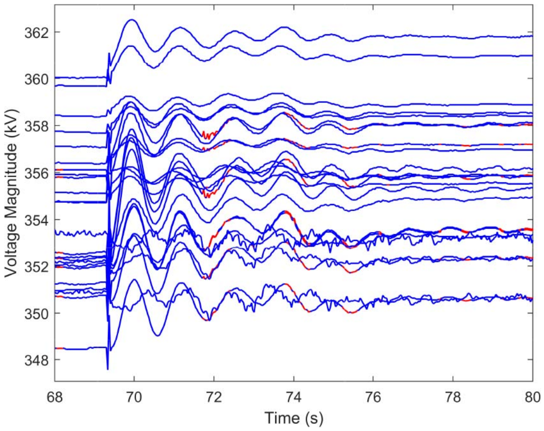

Fig. 6. Post Switching Event Recovery (Recovered in Red).

TABLE VI

SWITCHING EVENT COMPARISON OF MEAN ABSOLUTE ERROR IN KV FOR INCREASING CONSECUTIVE POINT OUTAGES

Missing Data(%)	OLAP-t	OLAP-s	PETRELS	GROUSE	SVT
10	0.0406	0.0386	0.0446	0.2077	0.0769
20	0.0418	0.0401	0.0469	0.2092	0.0773
30	0.0428	0.0421	0.0488	0.2042	0.0773
40	0.0440	0.0453	0.0501	0.2325	0.0772
50	0.0452	0.0496	0.0524	0.2132	0.0771

D. Recovery of Data With Prolonged Outages: Forced Oscillation Event

Further testing was performed on signals exhibiting forced oscillatory signatures. These signals, however, did not exhibit

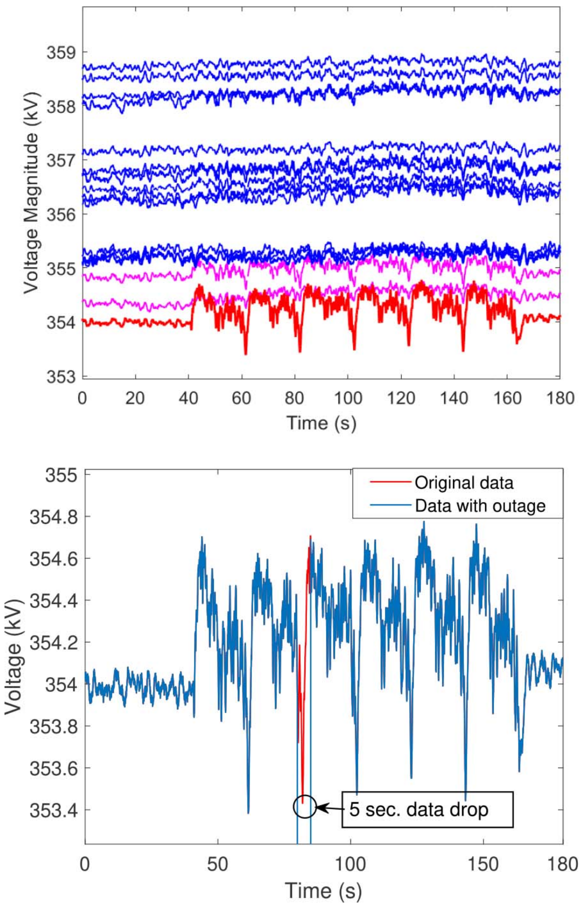

Fig. 7. Periodic Event and 150-Point Artificial Outage.

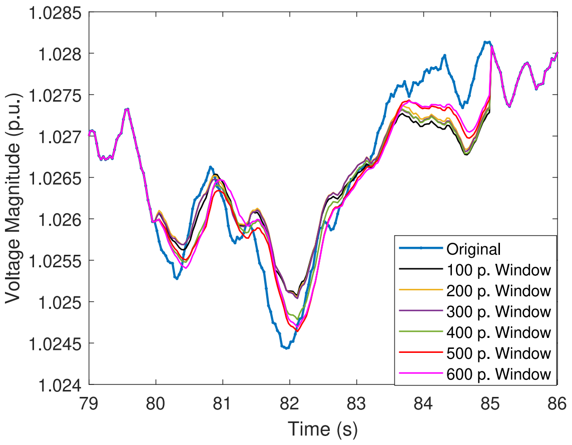

Fig. 8. Consecutive 5 sec. Outage Recovery.

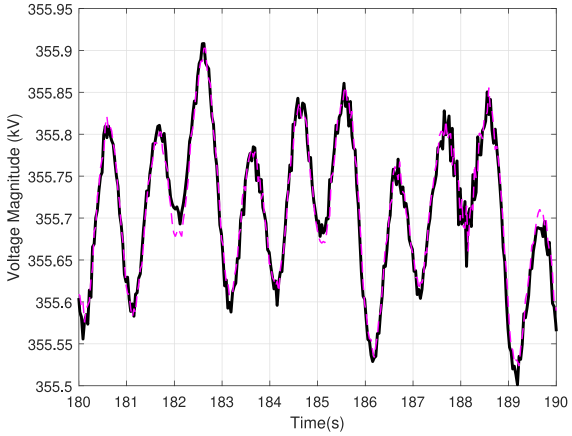

Fig. 9. Voltage Magnitude 1000-Point Outage: 1 Hz Case, 1 Channel Outage, 10sec. Segment.

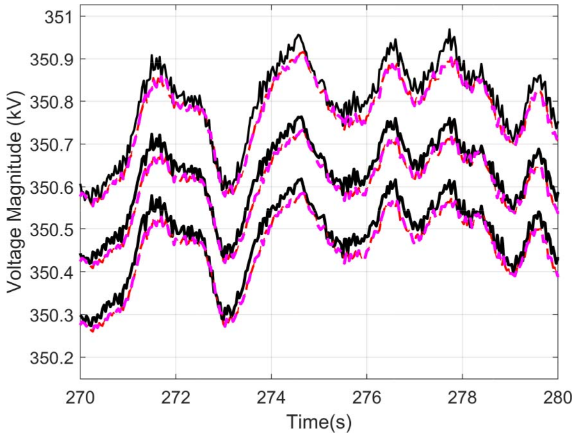

Fig. 10. Voltage Magnitude 1000-Point Outage: 1 Hz Case, 20 Channels Outage, 10 sec. Segment.

such irregular patterns as the signal in Fig. 7 and are much closer to sinusoids. The event used for testing, is a Forced Oscillation (FO) event as recorded by the PMUs across the New England (NE) power system. The event persisted for over 5 minutes and had a modal component around 1 Hz. The focus of this test will be prolonged and simultaneous outages across numerous channels. This is an illustrative way to stress test the ability of OLAP-t to reproduce the temporal patterns in the data, when significant portions are missing. In the following plots, the recovered signals are depicted by magenta dashed lines. For this section, 1000 data points will be dropped for a single and then for 20 channels (approximately 50% of available 345 kV channels).

In order to have a better appreciation of the recovery, in Fig. 9 a 10 sec. segment is presented. The recovery is almost exact for most of the missing points. The deviations from the ground truth are observed at the extrema of the oscillation, where OLAP-t seems to slightly overshoot. However, the errors are small compared to the amount of temporal information that was preserved.

Similarly, for the multiple channel drop case, three channels out of the 20 affected are shown in Fig. 10. Once again, the results seems to be very close to the original with a very small DC bias observed (less than 0.01 kV on average).

As it can be noted in Fig. 11, voltage angle reconstruction is even better, with almost exact recovery. For current magnitudes the results can be noted in Fig. 12. The recovery is of comparable accuracy to that of the voltages. For all the test cases presented, the window used for recovery is 300 points or 10 seconds of data. In general, the results involving voltage magnitudes and angles did not seem to pose any issues for OLAP-t and the recovery was successful across. The algorithm seemed to face some slight issues for prolonged current

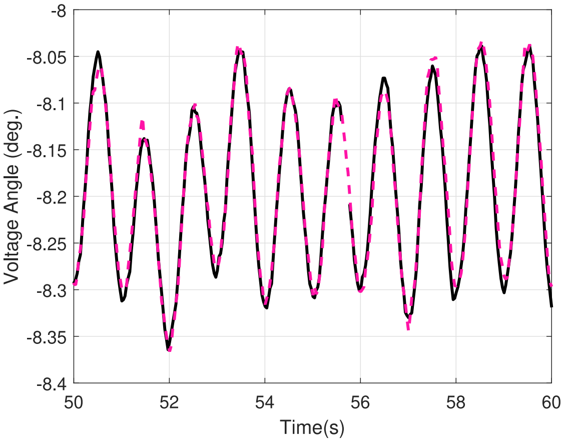

Fig. 11. Voltage Angle 1000-Point Outage: 1 Hz Case, 1 Channel Outage.

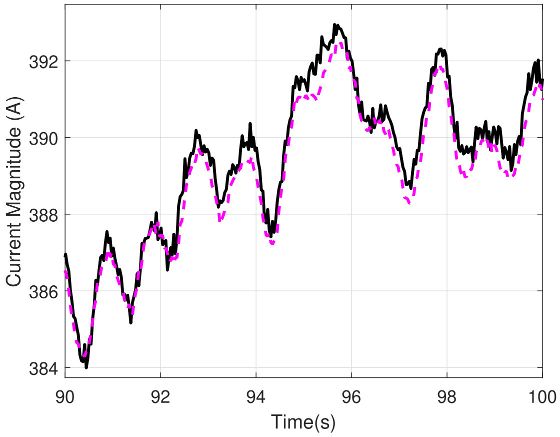

Fig. 12. Current Magnitude 1000-Point Outage: 1 Hz Case, 1 Channel Outage, 10 sec. Segment.

angle outages. In the sequel, a discussion on the recovery of current angles and magnitudes cases will be presented.

For the one channel outage of current magnitudes as it can be noted in Fig. 12 the recovery has very small errors, however, the characteristic signature of the signal is still preserved. However, for the 20 channel outage case, some channels exhibit higher error in the recovery. An illustrative example is shown in Fig. 13, in which the oscillation is preserved, but there seems to be a tracking error of the overall trend. As a result, the ground truth and the recovered signal seem to drift apart. However, the error is relatively small, although at some peaks of the signal it reaches 2 A.

The last signal that was tested for recovery purposes was the current angles. This signal seems to be the hardest to recover out of the four. Some possible reasons were the absence of low-rank behaviour in large parts of the data matrix, and the erratic behaviour of some current angles potentially due to control actions. Given the above difficulties, the recovery in general could be deemed satisfactory but did not seem to have as low errors as the one observed for voltage angles. Figs. 14 shows the successful recovery of the affected channel. In the 20 channel drop case, Fig. 15 presents one of the channels that exhibits a slight loss of the trend, that was observed in the current magnitudes as well (channel with worse deviation presented). Despite the deterioration of errors in current angle recovery, the error is normally less than 0.01° meeting C37.118 error requirements [21].

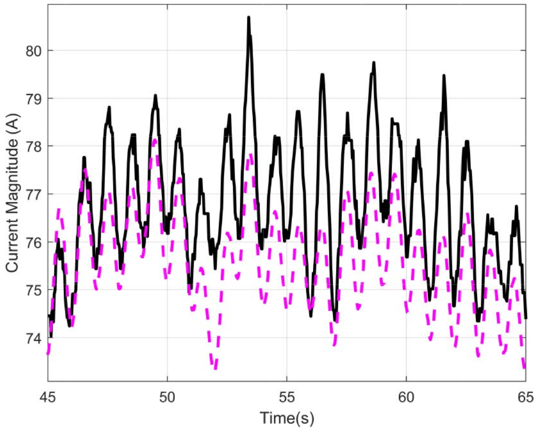

Fig. 13. Current Magnitude 1000-Point Outage: 1 Hz Case, 20 Channels Outage, Loss of Signal's Trend.

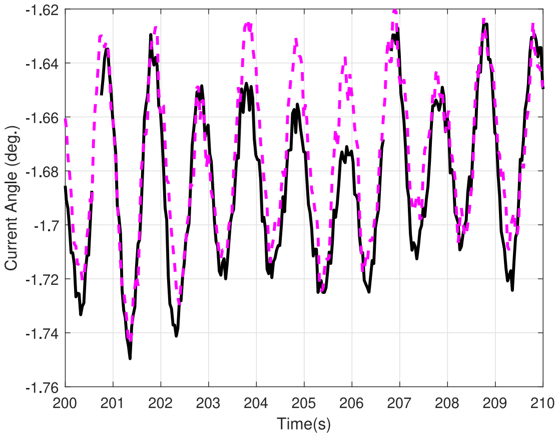

Fig. 14. Current Angle 1000-Point Outage 1 Hz Case 1 Channel Outage.

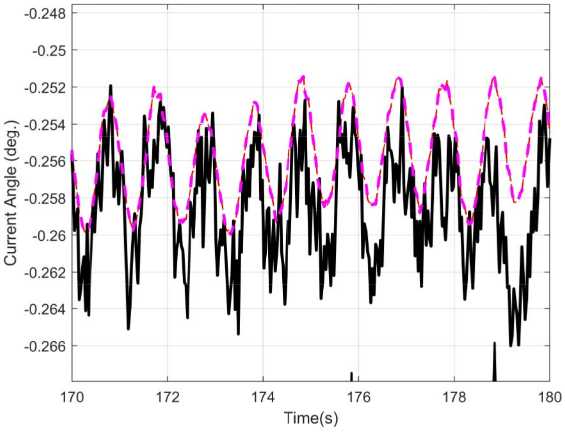

Fig. 15. Current Angle 1000-Point Outage 1 Hz Case 20 Channels Outage.

Table VII summarizes the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for all the cases. As expected, as more data points are missing simultaneously, errors increase. In the voltage cases, the increases can be

TABLE VII
ERROR FOR SIMULTANEOUS OUTAGE FOR FORCED OSCILLATION CASE

Case/ Signal Type	**1 Channel Mean Absolute Error**	**1 Channel Root Mean Square Error**	**20 Channel Mean Absolute Error**	**20 Channel Root Mean Square Error**
V$_m$ (kV)	0.0186	0.0238	0.0174	0.0229
V$_a$ (deg.)	0.0258	0.0245	0.0186	0.0232
I$_m$ (A)	0.3572	0.4605	0.6328	0.8809
I$_a$ (deg.)	0.0241	0.0219	0.2052	0.6513

deemed as minor, however, for currents, the increases are much more apparent. The MAE increased by 0.28 A and the RMSE by 0.42 A for current magnitudes. These deviations can be tolerable since the oscillatory behavior was mostly kept intact and the case of outage was extreme. In contrast, for multiple outage case of current angles, the errors were 10 or 30 times higher, indicating that the recovery is less accurate. Further investigation is required, in order to robustly recover signals exhibiting higher rank patterns, during prolonged missing segments.

Finally, it should be noted that the channels to be dropped, were chosen at random. Spatially correlated outages will be investigated in our future work.

V. Conclusion

In this paper, the temporal variation of the OLAP-s algorithm was presented and tested in terms of filter stability and recovery of PMU data for different situations. This method is based on low-rank matrix recovery concepts and formulated as a lightweight data-driven filter that aims to recover PMU data in real time. The algorithm was tested in ambient and switching event conditions, against well-known subspace methods and achieved comparable accuracy while being faster. The tests ranged from random outages of data to 10-200 consecutive point. Then, the method was then tested on data exhibiting periodicity, where simultaneous data outages across multiple channels and intentionally placed outages were introduced, to test if the method could perform as well. The OLAP-t method seems to perform very well across all these cases and for the majority of the PMU signals (voltage and currents). Some slight issues were observed in recovering current angles as the abrupt changes and the generally higher rank of the data posed some trend-following issues. Potential fixes to this problem could be further division of the signals in correlated subgroups (i.e., dropping the rank) and attempting recovery again. This will be one of the topics of future investigation and improvement, in order to make the procedure more robust. However, the shortcomings of the method are far outweighed by its efficiency in recovering signals with temporal behaviors and more importantly, prolonged missing data segments.

Acknowledgment

The authors thank New York Power Authority and ISO-NE for providing PMU data sets.

References

[1] A. G. Phadke and J. S. Thorp, *Synchronized Phasor Measurements and Their Applications*. New York, NY, USA: Springer, 2008.

[2] P. W. Sauer, M. A. Pai, and J. H. Chow, *Power System Dynamics and Stability: With Synchrophasor Measurement and Power System Toolbox 2e*, New York, NY, USA: Wiley, 2017.

[3] N. Zhou, D. Meng, Z. Huang, and G. Welch, “Dynamic state estimation of a synchronous machine using PMU data: A comparative study,” *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 450–460, Jan. 2015.

[4] L. Hu, Z. Wang, I. Rahman, and X. Liu, “A constrained optimization approach to dynamic state estimation for power systems including PMU and missing measurements,” *IEEE Trans. Control Syst. Technol.*, vol. 24, no. 2, pp. 703–710, Mar. 2016.

[5] U. Agrawal and J. W. Pierre, “Detection of periodic forced oscillations in power systems incorporating harmonic information,” *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 782–790, Jan. 2019.

[6] D. J. Trudnowski and J. W. Pierre, “Overview of algorithms for estimating swing modes from measured responses,” in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Calgary, AB, Canada, 2009, pp. 1–8.

[7] K. D. Jones, A. Pal, and J. S. Thorp, “Methodology for performing synchrophasor data conditioning and validation,” *IEEE Trans. Power Syst.*, vol. 30, no. 3, pp. 1121–1130, May 2015.

[8] J. J. Q. Yu, A. Y. S. Lam, D. J. Hill, Y. Hou, and V. O. K. Li, “Delay aware power system synchrophasor recovery and prediction framework,” *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3732–3742, Jul. 2019.

[9] J. J. Q. Yu, D. J. Hill, V. O. K. Li, and Y. Hou, “Synchrophasor recovery and prediction: A graph-based deep learning approach,” *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7348–7359, Oct. 2019.

[10] C. Ren and Y. Xu, “A fully data-driven method based on generative adversarial networks for power system dynamic security assessment with missing data,” *IEEE Trans. Power Syst.*, vol. 34, no. 6, pp. 5044–5052, Nov. 2019.

[11] Q. Li, Y. Xu, C. Ren, and J. Zhao, “A hybrid data-driven method for online power system dynamic security assessment with incomplete PMU measurements,” in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Atlanta, GA, USA, Aug. 2019, pp. 1–5.

[12] M. Liao, D. Shi, Z. Yu, Z. Yi, Z. Wang, and Y. Xiang, “An alternating direction method of multipliers based approach for PMU data recovery,” *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4554–4565, Jul. 2019.

[13] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, “Robust recovery of missing data in electricity distribution systems,” *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4057–4067, Jul. 2019.

[14] Y. Hao, M. Wang, and J. H. Chow, “Modeless streaming synchrophasor data recovery in nonlinear systems,” *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1166–1177, Mar. 2020, doi: 10.1109/TPWRS.2019.2939559.

[15] 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 Syst.*, vol. 31, no. 2, pp. 1006–1013, Mar. 2016.

[16] J.-F. Cai, E. J. Candes, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jun. 2010.

[17] L. Balzano, R. Nowak, and B. Recht, “Online identification and tracking of subspaces from highly incomplete information,” in *Proc. 48th Annu. Allerton Conf. Commun. Control Comput.*, Allerton, IL, USA, 2010, pp. 704–711.

[18] Y. Chi, Y. C. Eldar, and R. Calderbank, “PETRELS: Parallel subspace estimation and tracking by recursive least squares from partial observations,” *IEEE Trans. Signal Process.*, vol. 61, no. 23, pp. 5947–5959, Dec. 2013.

[19] Y. Xie, J. Huang, and R. Willett, “Change-point detection for high-dimensional time series with missing data,” *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 12–27, Feb. 2013.

[20] G. M. De Mijolla, S. Konstantinopoulos, P. Gao, J. H. Chow, and M. Wang, “An evaluation of algorithms for synchrophasor missing data recovery,” in *Proc. Power Syst. Comp. Conf. (PSCC)*, Dublin, Ireland, 2018, pp. 1–6.

[21] *IEEE Standard for Synchrophasors for Power Systems*, Standard IEEE Standard C37.118-2005, Mar. 2006.

Stavros Konstantinopoulos (Student Member, IEEE) received the B.Sc. degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 2015, and the M.Sc. degree in electrical and systems engineering from Rensselaer Polytechnic Institute, Troy, MI, USA, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include power system dynamics and control, impact of renewable generation integration, and synchronized phasor data applications.

Genevieve M. De Mijolla (Member, IEEE) received the M.S. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, MI, USA. She is currently working as a Senior Engineer for General Electric Energy Consulting. Her primary interests include production cost simulation analysis, renewable integration, and energy storage valuation.

Joe H. Chow (Life Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA. After working in the General Electric Power System Business, Schenectady, NY, USA. He joined Rensselaer Polytechnic Institute in 1987, where he is a Institute Professor of electrical, computer, and systems engineering. His research interests include power system dynamics and control and synchronized phasor data. He is a member of the U.S. National Academy of Engineering.

Hanoch Lev-Ari (Life Fellow, IEEE) received the B.S. (*summa cum laude*) and M.S. degrees in electrical engineering from the Technion—Israel Institute of Technology, Haifa, Israel, in 1971 and 1978, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1984. He is currently a Professor with the Department of Electrical and Computer Engineering, Northeastern University, where he was also the Director of the Communications and Digital Signal Processing Center from 1994 to 1996. His present interests include adaptive filtering under the non-stationary regime, dynamic time-frequency analysis, and multirate/multisensor networked state estimation; with applications to identification of time-variant systems, customized dynamic phasors, dynamic power decomposition, and adaptive power flow control in polyphase power systems. He served as an Associate Editor for *Circuits, Systems and Signal Processing* and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS. He is a member of SIAM.

Meng Wang (Member, IEEE) received the B.S. and M.S. degrees from Tsinghua University, China, in 2005 and 2007, respectively, and the Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2012. She is an Associate Professor with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA. Her research interests include high-dimensional data analytics, machine learning, power systems monitoring, and synchrophasor technologies.