Missing Data Recovery by Exploiting Low-dimensionality in Power System Synchrophasor Measurements

Pengzhi Gao, *Student Member, IEEE,* Meng Wang, *Member, IEEE,* Scott G. Ghiocel, *Member, IEEE,*
Joe H. Chow, *Fellow, IEEE,* Bruce Fardanesh, *Fellow, IEEE,* and George Stefopoulos, *Member, IEEE.*

***Abstract*—This paper presents a new framework of recovering missing synchrophasor measurements (erasures). Leveraging the approximate low-rank property of phasor measurement unit (PMU) data, we connect the problem of recovering PMU data erasures with recent advances in *low-rank matrix completion* methods. Since the existing analysis for matrix completion methods assumes an independent-erasure model that does not capture the correlations in PMU erasures, we propose two models to characterize the temporal and the channel correlations in PMU erasures and provide theoretical guarantees of a matrix completion method in recovering correlated erasures in both models. We also propose an online algorithm that can fill in the missing PMU measurements for real-time applications. Numerical experiments on actual PMU data are conducted to verify the effectiveness of the proposed methods.**

***Index Terms*—low rank, matrix completion, missing data, phasor measurement unit, event detection.**

I. Introduction

PHASOR Measurement Units (PMUs) provide synchronized phasor measurements of remote points in the power system at a much faster sampling rate than that in the traditional Supervisory Control and Data Acquisition (SCADA) system. These synchrophasor measurements can improve the accuracy of offline applications such as post event analysis [3], [16] and system identification [19], [35], as well as real-time data-driven analytics such as state estimation [12], [14], [27], [28], [32], [34] and disturbance identification [8].

State estimation computes the system state based on the obtained measurements. The performance of the state estimator largely depends on the availability and the quality of the measurements. Data losses can happen in an unpredictable way during the communication between PMUs and the phasor data concentrator at the central operator. Losing measurements makes the system unobservable and degrades the performance of the state estimator. Therefore, it is important to develop methods that can recover missing measurements so as to enable the full functionality of the state estimator.

Because PMU measurements are sampled at synchronized time instants, and the measurements of nearby PMUs are correlated through the power system topology, the high dimensional PMU data exhibits a coherence property [8], [10], [15]. If measurements of multiple PMU channels are represented by a matrix with each row representing the measurements of one channel across time, then the matrix only contains a small number of significant singular values (i.e., approximately low-rank). The central idea of this paper is to recover the missing points by leveraging the low-rank property of the PMU data.

Low-rank matrix completion has been widely applied in collaborative filtering [1], computer vision [7], [31], remote sensing [30], and system identification [21]. It is demonstrated that all entries in an $n_1 \times n_2$ $(n = \max(n_1, n_2))$ matrix of rank r $(r \ll n_1, n_2)$ can be efficiently recovered even if only $O(rn \log^2 n)$[1] randomly selected entries of the matrix are observed ([5], [6], [17], [29]). Missing data recovery can be achieved by solving a convex optimization problem [13]. Other matrix completion algorithms have been proposed and analyzed, such as singular value thresholding (SVT) [4], ADMiRa [20], singular value projection (SVP) [18], and information cascading matrix completion (ICMC) [23]. One key assumption of the existing theoretical analysis of matrix completion methods is that each erasure (missing entry) happens independently of others. The independent erasure model does not adequately describe the PMU data erasures, which usually exhibit temporal and spatial correlations. The theoretical analysis of matrix completion performance when the erasures are correlated is still an open problem.

The above low-rank matrix completion methods are block-processing methods, and the recovered measurements can be used for offline applications such as model validation and post-event analysis. Real-time applications such as state estimation and disturbance identification require the development of online data recovery methods that can fill in the missing points immediately after sampling. Methods for online subspace tracking with missing data (e.g.,[2], [8], [9], [22]) have been developed. They fill in the missing points by exploiting the low-dimensionality of the data and usually assume the dimensionality is known and fixed. The dimensionality of block PMU measurements, however, can change significantly when the system experiences a disturbance. Therefore, an online method for recovering missing PMU data should also track the dimensionality of the block PMU data.

P. Gao, M. Wang, S.G. Ghiocel, and J.H. Chow are with the Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY. Email: {gaop, wangm7, ghiocs2, chowj}@rpi.edu.

B. Fardanesh and G. Stefopoulos are with New York Power Authority, White Plains, NY. Email: {Bruce.Fardanesh, George.Stefopoulos}@nypa.gov.

Partial and preliminary results have appeared in [15].

[1] We use the big O notation $g(n) \in O(f(n))$ if as n goes to infinity, $g(n) \leq C \cdot f(n)$ eventually holds for some positive constant C.

The main contributions of this paper are threefold. (1) We connect missing PMU data recovery with the recent advance in low-rank matrix completion methods. We demonstrate the effectiveness of some matrix completion methods through numerical experiments on actual PMU data from the Central New York Power System. (2) The existing analysis of matrix completion methods is based on independent erasure model and does not apply to correlated erasures in PMU data. We demonstrate through theoretical analysis that even when the erasures are correlated, ICMC can recover the missing entries if $O(n^{2-\frac{1}{r+1}}r^{\frac{1}{r+1}}\log^{\frac{1}{r+1}}n)$ entries are observed. (3) We develop an online missing data recovery method that is adaptive to power system disturbances.

The rest of the paper is organized as follows. We demonstrate the low-rank property of PMU data and introduce matrix completion methods in Section II. We provide the theoretical guarantees of recovering correlated erasures in Section III. In Section IV, we propose an online algorithm for recovering missing PMU data. We also discuss the application of low-rank methods to PMU data compression. Section V records the numerical experiments, and Section VI concludes the paper.

II. Low-rank Matrix Completion

A. *Low-rank property of PMU data*

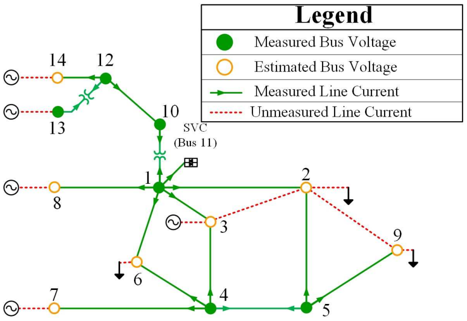

Fig. 1. Six PMUs in the Central NY Power System

The PMU data used here are from six multi-channel PMUs deployed in the Central New York (NY) Power System (Fig. 1). Each PMU measures the voltage phasor at the corresponding bus and the current phasors on its incident lines at a rate of thirty samples per second. The six PMUs measure thirty-seven voltage and current phasors in total.

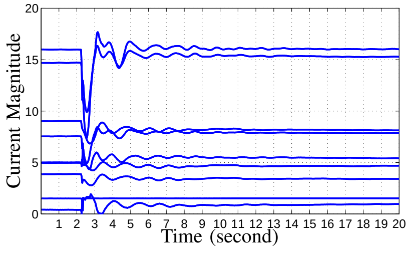

Fig. 2. Current magnitudes of PMU data (9 current phasors out of 37 phasors)

Fig. 2 shows the current magnitudes of PMU measurements. A disturbance occurs around $t = 2.5$ s. Let the complex matrix $M \in \mathbb{C}^{600\times 37}$ contain the PMU data in 20 seconds (30 samples/second). Each column corresponds to a sequence of measurements of one PMU channel such as a voltage or current phasor. Phasors are represented by complex numbers in rectangular form. Each row corresponds to the PMU measurements at the same sampling instant. Fig. 3 shows the singular values of M. The largest ten singular values are 894.5942, 36.8319, 20.7160, 8.3400, 3.0771, 2.4758, 1.9705, 1.3543, 0.5930 and 0.2470. We can approximate M by a rank-eight matrix with a negligible error. The low-dimensionality of PMU measurements has also been observed in [8], [10], [15].

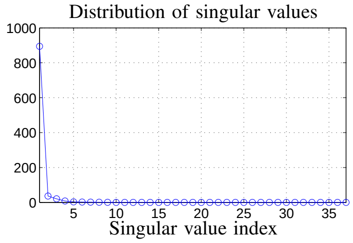

Fig. 3. Singular values of a 600 × 37 PMU data matrix.

B. *Low-rank matrix completion*

A low-rank matrix completion problem aims to identify a low-rank matrix $M \in \mathbb{C}^{n_1\times n_2}$ from its partially observed entries. One approach is to solve a convex program [13]

$$\min_{X\in\mathbb{C}^{n_1\times n_2}} \|X\|_* \quad \text{s.t.} \quad P_\Omega(X) = P_\Omega(M), \tag{1}$$

where the nuclear norm $\|X\|_*$ is the sum of the singular values of X, Ω is the set of the locations (i, j) of the observed entries, and $P_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ denotes the projection of a matrix onto the pairs of indices in Ω :

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

The solution to (1) is the original matrix M under certain conditions ([5], [6], [17], [29]).

Singular value thresholding (SVT) [4] is an algorithm that approximately solves (1) by solving a modified version

$$\min_{X\in\mathbb{C}^{n_1\times n_2}} \lambda\|X\|_* + \frac{1}{2}\|X\|_F^2 \quad \text{s.t.} \quad P_\Omega(X) = P_\Omega(M), \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm, and λ is a fixed parameter.

Another matrix completion method, called Information Cascading Matrix Completion (ICMC) [23], solves for $X \in \mathbb{C}^{n_1\times r}$ and $Y \in \mathbb{C}^{n_2\times r}$ (the rank r of M is given) such that

$$P_\Omega(XY^T) = P_\Omega(M), \tag{3}$$

and uses XY^T as an estimate of M. M is assumed to be non-degenerate, i.e., any r rows of X or Y are linearly independent.

ICMC assumes that any r rows of X or Y are linearly independent and progressively determines the entries of X and Y. Construct a bipartite graph with left vertices $U = \{\mu_1, ..., \mu_{n_1}\}$ and right vertices $V = \{\nu_1, ..., \nu_{n_2}\}$. There is an edge between μ_i and ν_j if and only if the matrix entry M_{ij} is observed. A vertex μ_i in U (or ν_j in V) is called "infected" if the ith row of X (or the jth row of Y) is determined. $L \subseteq U$ and $R \subseteq V$ are defined as the sets of infected vertices.

At initialization, ICMC chooses a set of r rows of X (or Y) and fix them to be an $r \times r$ identity matrix I_r. The corresponding vertices are marked as infected. ICMC then repeats the the following two steps in each iteration:

1. For every uninfected ν_j in V, if ν_j is connected to at least r vertices in L, compute the jth row of Y by solving r linearly independent equations. Mark ν_j as infected and add it to R.
2. For every uninfected μ_i in U, if μ_i is connected to at least r vertices in R, compute the ith row of X by solving r linearly independent equations. Mark μ_i as infected and add it to L.

ICMC fully determines X and Y if $L = U$ and $R = V$ in the end, or it declares recovery failure otherwise.

III. Matrix completion with correlated erasures

We divide PMU data into blocks and recover missing data in each block using low-rank matrix completion methods. Existing analysis of matrix completion methods assumes that erasures are independent of each other. Here we propose two models to characterize the correlations in PMU data erasures and analyze the performance of ICMC in recovering correlated erasures. We discuss the two models separately to simplify the analysis. Note that these models are used for performance analysis and not required by ICMC for data recovery.

Let $M \in \mathbb{C}^{n_1 \times n_2}$ denote the data matrix with each column representing consecutive measurements of one PMU channel. Since measurements are usually noisy, M is non-degenerate (verified in our simulation as well) as required by ICMC. Define $n = \max(n_1, n_2)$ and $\rho = \min(n_1, n_2)/n$. We assume ρ is a constant in $(0, 1]$. Let r denote the rank of M, and we assume r is much less than n.[2]

A. Temporal correlation in missing PMU data

We model the temporal correlation among the erasures as follows. For each entry in the jth column of M, τ_j consecutive measurements starting from this entry are erased with probability $1 - p_j$ for some integer $\tau_j \geq 1$ and constant $p_j \in [0, 1]$. Thus, each entry in the jth column (except the first $\tau_j - 1$ entries) is observed with probability $p_j^{\tau_j}$. Note that the erasure of each entry depends on the erasure of its previous $\tau_j - 1$ samples. When $\tau_j = 1$ for all j, this model reduces to the case that erasures are independent.

We apply the ICMC algorithm to recover M starting from a set R_0 with $|R_0| = r$ such that R_0 contains the highest-weight vertices in V. Define $p = \min_j p_j$ and $\tau = \max_j \tau_j$. We have

Theorem 1. *For every positive γ, there exists a positive constant $c(\gamma)$ such that if*

$$p \geq \left(\frac{c(\gamma)\tau r \log n}{n}\right)^{\frac{1}{\tau(r+1)}} \tag{4}$$

holds, then ICMC correctly recovers M with probability at least $1 - n^{-\gamma}$.

Proof: The proof uses the same idea of the proof of Theorem 4.1 in [23] with some modifications to address the correlations in erasures. Fix an uninfected v in V. For each t in U, define I_{tv} as the indicator function that is 1 if t is connected to all vertices in R_0 and v. Note that u will be infected after two steps by ICMC starting from R_0 if $\sum_{t \in U} I_{tv} \geq r$.

Pick a set $U' = \{\mu_\tau, \mu_{2\tau}, ..., \mu_{\lfloor n_1/\tau \rfloor \tau}\} \subseteq U$. Although I_{tv}'s ($t \in U$) are dependent random variables, I_{tv}'s for $t \in U'$ are independent variables taking values in $\{0, 1\}$, and

$$\Pr(I_{tv} = 1) = \prod_{j \in R_0 \cup \{v\}} p_j^{\tau_j} \geq p^{\tau(r+1)}.$$

Then we have[3]

$$E[\sum_{t \in U'} I_{tv}] = \sum_{t \in U'} \Pr(I_{tv} = 1) \geq \frac{n_1 p^{\tau(r+1)}}{\tau} \geq c(\gamma)\rho r \log n,$$

where the second inequality follows from (4). Then from the Chernoff bound [24],

$$\begin{aligned} \Pr(\sum_{t \in U'} I_{tv} < (1-\delta)E[\sum_{t \in U'} I_{tv}]) &\leq \exp(-\delta^2 E[\sum_{t \in U'} I_{tv}]/2) \\ &\leq \exp(-c(\gamma)\delta^2 \rho r \log n/2). \end{aligned} \tag{5}$$

For any given γ, we pick constant $c(\gamma)$ and $\delta \in (0, 1)$ such that $(1-\delta)c(\gamma)\rho > 1$ and $c(\gamma)\delta^2\rho/2 \geq \gamma + 2$. Then we have

$$(1-\delta)E[\sum_{t \in U'} I_{tv}] \geq (1-\delta)c(\gamma)\rho r \log n \geq r \tag{6}$$

and

$$c(\gamma)\delta^2 \rho r/2 \geq \gamma + 2 \tag{7}$$

Combining (5)-(7), we have

$$\Pr(\sum_{t \in U} I_{tv} < r) \leq \Pr(\sum_{t \in U'} I_{tv} < r) \leq e^{-(\gamma+2)\log n} \leq \frac{1}{2n^{\gamma+1}},$$

where the first inequality holds since I_{tv}'s are nonnegative random variables. By taking a union bound over all $v \notin R_0$, we obtain with probability at least $1 - \frac{1}{2n^\gamma}$, that all vertices in V will be infected after two steps.

After V is infected, a vertex t in U will be infected in the following step if the number of edges incident to t, denoted by S_t, is at least r. S_t is a sum of n_2 independent r.v.'s S_{tj} taking values in $\{0, 1\}$ where $\Pr(S_{tj} = 1) = p_j^{\tau_j}$. Thus,

$$\begin{aligned} \Pr(S_t < r) < \Pr(S_t < (1-\delta)E[S_t]) &\leq \exp(-\delta^2 n_2 p^\tau/2) \\ &\leq \exp(-(\gamma+2)\log n) \leq \frac{1}{2n^{\gamma+1}}, \end{aligned}$$

where the first and the third inequalities follow from our choice δ and $c(\gamma)$, and the second inequality follows from the Chernoff bound. By taking a union bound over all $t \in U$, we conclude that with probability at least $1 - n^{-\gamma}$, ICMC will infect all the vertices in U and V. ∎

Theorem 1 indicates that if the temporal correlation of the erasures satisfies certain condition, ICMC can correctly recover all the missing entries in M. The expected number of obtained PMU measurements is $Cn^{2-\frac{1}{r+1}}(\tau r \log n)^{\frac{1}{r+1}}$ for some constant C. If r and τ are constants, and n is sufficiently large[4], the required number of samples is much less than the matrix dimension $n_1 n_2$.

[2] Strictly speaking, M is only approximately low-rank and can be viewed as the summation of a low-rank matrix plus noise. We assume M is strictly low-rank for notational simplicity, and the results can be extended to approximate low-rank matrices with simple modifications.

[3] We drop $\lfloor \cdot \rfloor$ in later part for notational simplicity.

[4] Although theoretical results require a large n, we obtain good practical recovery performance as long as n is not too small, see Section V.

When the erasures are independent of each other, $O(nr\log^2 n)$ measurements are sufficient to recover the remaining missing entries (e.g., [5], [17]). Although our bound of the required number of observations is larger than the existing bound when the erasures are independent, this is the first theoretical guarantee of low-rank matrix completion methods for correlated erasures to the best of our knowledge.

B. Channel correlations in missing PMU data

Let $M \in \mathbb{C}^{n_1\times n_2}$ contain the data of S multi-channel PMUs and d_i denote the number of measurement channels of PMU i. The first d_1 columns of M correspond to PMU 1, and the last d_S columns correspond to PMU S. We assume $d_{\max} = \max_i d_i$ is a constant.

We model the channel correlation among the erasures as follows. At each sampling instant, all d_i measurements of PMU i are erased simultaneously with probability $1-q_i$ for some $q_i \in [0,1]$. If simultaneous erasure does not happen, the measurement of the jth channel is erased independently with probability $1-p_{ij}$. We assume $p_{i1} \geq p_{i2} \geq p_{id_i}$ without loss of generality. Thus, at each sampling instant, the measurement from the jth channel of PMU i is observed with probability $p_{ij}q_i$, and the erasures in the same PMU are correlated.

We apply ICMC to recover M starting from a set $L_0 \subset U$ with $|L_0| = r$ that contains the highest-weight vertices in U.

Theorem 2. *For every $\gamma > 0$, there exists a constant $c(\gamma)$ such that if both*

$$p_{i1}q_i \geq \left(\frac{c(\gamma)d_{\max}r\log n}{n}\right)^{\frac{1}{r+1}}, \quad \forall i, \tag{8}$$

and

$$p_{ij}q_i \geq \frac{c(\gamma)r\log n}{n}, \quad \forall i,j, \tag{9}$$

hold, then ICMC correctly recovers M with probability at least $1-n^{-\gamma}$.

Proof: Since the arguments are similar to the proof of Theorem 1, we skip the details. Fix $t \in U$ such that $t \notin L_0$. For each $v \in V$, define $\hat{I}_{tv}$ as the indicator function that is 1 if v is connected to all vertices in L_0 and t. Then t will be infected after two steps starting from L_0 if $\sum_{v\in V}\hat{I}_{tv} \geq r$.

Pick the set $V' = \{\nu_1, \nu_{d_1+1}, \nu_{d_1+d_2+1}, ..., \nu_{\sum_{i=1}^{S-1}d_i+1}\} \subseteq V$. Although $\hat{I}_{tv}$'s ($v \in V$) are dependent random variables, $\hat{I}_{tv}$'s for $t \in V'$ are independent variables taking values in $\{0,1\}$. From (8) and the Chernoff bound, one can derive that

$$\Pr\left(\sum_{v\in V'}\hat{I}_{tv} < r\right) \leq \frac{1}{2n^{\gamma+1}}.$$

Then from the union bound, U will be infected after two steps with probability at least $1-\frac{1}{2n^{\gamma}}$.

From (9) and the Chernoff bound, with probability at least $1-\frac{1}{2n^{\gamma+1}}$, each fixed column of M has at least r observations. Then with probability at least $1-\frac{1}{2n^{\gamma+1}}$, V will be infected in the following step after U is infected. ■

Theorem 2 indicates that when the erasures have channel correlations, ICMC can recover the missing entries if $Cn^{2-\frac{1}{r+1}}(r\log n)^{\frac{1}{r+1}}$ entries are observed for some constant C. In the special case that $p_{ij} = p$ and $q_j = q$ for all i and j, when we fix the expected number of observations (by fixing pq) but vary the channel correlation (by changing q), the recovery performance changes with channel correlation even though the number of observations is the same (see Section V). Theorem 2 does not capture this dependence since it considers the worst-case channel correlation.

IV. Online Method for PMU Erasure Estimation and Event Detection

We have discussed the theoretical guarantee of matrix completion methods on recovering correlated erasures in Section III. The recovered measurements could be used for offline applications like system identification and past event analysis. We then propose an online algorithm for PMU data processing (OLAP) that can fill in the missing data for real-time applications like state estimation and disturbance detection.

Low-dimensional structures exist in the high-dimensional datasets of other applications such as image and video processing, Internet traffic monitoring, etc. Some methods have been developed to track the low-dimensional structure with incomplete data. Examples include Grassmannian Rank-One Update Subspace Estimation (GROUSE) [2], Parallel Estimation and Tracking by Recursive Least Squares (PETRELS) [9] when the low-dimensional is represented by a linear subspace, as well as Multiscale Online Union of Subsets Estimation (MOUSSE) [33] when the low-dimensional structure is represented by a union of subspaces. These methods assume the dimension of the subspace is known and fixed, since the low-dimensional structure is slow-varying in the aforementioned applications. The PMU measurements, however, might change significantly at the onset of a disturbance. OLAP is built upon the recent art in subspace tracking and has an additional component that can track the dimensionality of the PMU data.

OLAP continuously updates the dominant singular values and singular vectors of a matrix that contains PMU measurements in the past w sampling instants. At the new sampling instant, OLAP fills in the missing points based on the stored singular vectors. It also declares a disturbance if the proposed event indicator becomes larger than a pre-specified threshold.

We maintain a matrix $M_{\mathrm{w}} \in \mathbb{C}^{w\times n_2}$ that contains the PMU measurements in the past w instants. We compute the singular value decomposition (SVD) of M_{w}^T,

$$M_{\mathrm{w}}^{\mathrm{T}} = U\Sigma V^{\dagger},$$

where $U \in \mathbb{C}^{n_2\times n_2}$ and $V \in \mathbb{C}^{w\times w}$ are unitary matrices, $V^{\dagger}$ is the complex conjugate of V, and $\Sigma \in \mathbb{R}^{n_2\times w}$ has singular values on its diagonal. We fix constant $\gamma_{\mathrm{err}} \in [0,1]$ as the threshold of the relative approximation error. Let U^r and V^r denote r dominant left and right singular vectors. Σ^r represents the diagonal matrix with r dominant singular values. We pick the smallest integer r such that

$$\|M_{\mathrm{w}}^{\mathrm{T}} - U^r\Sigma^r(V^r)^{\dagger}\|_2/\|M_{\mathrm{w}}^{\mathrm{T}}\|_2 \leq \gamma_{\mathrm{err}}$$

holds, and we use this integer as the approximate rank of M_{w}. This process is summarized in Subroutine 1.

Let $\boldsymbol{\beta} \in \mathbb{C}^{n_2\times 1}$ denote the new measurements (including erasures) obtained at the current sampling instant. Let Ψ

denote the support set of the observed entries, and let Ψ^c denote the set of erasures. Let U^r_Ψ and $\boldsymbol{\beta}_\Psi$ be the submatrices of U^r and $\boldsymbol{\beta}$ with row indices in Ψ. We compute

$$\boldsymbol{v}^* = \arg\min_{\boldsymbol{v}\in\mathbb{C}^{r\times 1}} \|U^r_\Psi \boldsymbol{v} - \boldsymbol{\beta}_\Psi\|_2,$$

and use $U^r_{\Psi^c}\boldsymbol{v}^*$ as the estimate of the missing data points. We update the matrix M_{w} by dropping the oldest data row and adding the new data row $\boldsymbol{\beta}^{\mathrm{T}}$ and repeat the process.

We define coefficient $\eta_t := \sigma_2/\sigma_1$, where σ_1 and σ_2 are the largest and second largest singular values of M_{w} at instant t, respectively. OLAP computes η_t at each instant and keeps each η for the past $\bar{w}$ instants. It then computes the average relative change of η from time t to time k, i.e.,

$$\zeta_{\mathrm{t,k}} := \frac{\eta_t - \eta_k}{\eta_k(t-k)}.$$

OLAP declares disturbance at instant t if $\zeta_{\mathrm{t,k}}$ becomes larger than a pre-determined threshold θ for some $k \in [t-\bar{w}, t)$. Note that θ represents the relative change of η and takes the value from 0 to 100%. OLAP is summarized in Algorithm 2.

```
Subroutine 1 $U^r$ = AppRank($U$, $\Sigma$, $V$, $\gamma_{\mathrm{err}}$)
Input: $U \in \mathbb{C}^{n_2\times n_2}$, $\Sigma \in \mathbb{C}^{n_2\times w}$, $V \in \mathbb{C}^{w\times w}$, relative
    approximation error $\gamma_{\mathrm{err}}$.
Find the smallest $r$ that $\frac{\|U\Sigma V^\dagger - U^r\Sigma^r(V^r)^\dagger\|_2}{\|U\Sigma V^\dagger\|_2} \leq \gamma_{\mathrm{err}}$ holds.
Return: $r$ dominant left singular vectors $U^r$.
```

```
Algorithm 2 Online algorithm for PMU data processing
Input: Approximate rank $r$ and $U^r$ obtained from the initial
    data, threshold coefficients $\gamma_{\mathrm{err}}$ and $\theta$ in $[0,1]$.
for $t = 1, 2, 3, ...,$ do
    Receive new data $\boldsymbol{\beta} \in C^{n_2\times 1}$ with erasures.
    Compute $\boldsymbol{v}^* = \arg\min_{\boldsymbol{v}} \|U^r_\Psi\boldsymbol{v} - \boldsymbol{\beta}_\Psi\|_2$.
    Fill in the missing entries of $\boldsymbol{\beta}$ with $U^r_{\Psi^c}\boldsymbol{v}^*$.
    Update $M_{\mathrm{w}}$ by dropping the oldest row and adding $\boldsymbol{\beta}^{\mathrm{T}}$.
    Compute SVD $M^{\mathrm{T}}_{\mathrm{w}} = U\Sigma V^\dagger$ and coefficient $\eta_t = \frac{\sigma_2}{\sigma_1}$.
    Compute $U^r$ = AppRank($U, \Sigma, V, \gamma_{\mathrm{err}}$).
    if $\max_{t-\bar{w}\leq k<t}\zeta_{\mathrm{t,k}} > \theta$ and no event is declared yet
    then
        Declare that an abnormal event happens.
    end if
end for
```

We remark that the low-rank property of the PMU measurements can be leveraged for lossy data compression as well. The fast sampling rate (30 Hz or more) and the increasing deployment of PMUs introduce a challenge to data storage. For example, the Tennessee Valley Authority with 120 PMUs installed throughout the eastern half of North America needs to manage 36 GB data per day [10]. Various methods (e.g., [11], [25]) have been developed to compress data in individual channels. By only storing the dominant singular values and vectors, we compress the data both inside a channel and across channels. The SVD-based method for PMU data compression has also been discussed in [26].

Consider $M \in \mathbb{C}^{n_1\times n_2}$ as the PMU data matrix. Suppose the approximate rank of M is k, and we store the k largest singular values and the corresponding singular vectors. The compression ratio λ for M is

$$\alpha = k(n_1 + n_2 + 1)/(n_1 n_2),$$

which can be very small when k is small.

V. Simulation

We present the numerical experiments on four PMU datasets from the Central New York (NY) Power System (Fig. 1). The current magnitudes of these 20-second PMU datasets are shown in Fig. 4. The disturbances are caused by generator trips. We first compare the performance of SVT and ICMC in Section V-A. We then test OLAP and compare it with SVT and ICMC in Section V-B. We also evaluate the data compression performance in Section V-C. The recovery performance is measured by the relative recovery error $\|M - M_{\mathrm{rec}}\|_F/\|M\|_F$, where matrix M represents the actual PMU data, and M_{rec} represents the recovered data. The average erasure rate p_{avg} is the percentage of missing entries. We remark that when p_{avg} is high, ICMC sometimes cannot converge and would declare recovery failure. In these cases, we simply use a zero matrix as an estimate of the original matrix, and the resulting relative recovery error is one.

A. Filling in temporal-correlated and channel-correlated erasures by SVT and ICMC

We delete some PMU measurements based on the correlated erasure models introduced in Section III. For temporal-correlated erasures, we fix p_{avg} and vary τ. The temporal correlation increases when τ increases. For spatial-correlated erasures, we fix p_{avg} and vary q. When q decreases, the channel correlation increases. In the recovery algorithms, the parameter λ is set to be $8(\sqrt{n_1 n_2})$ in SVT where $n_1 \times n_2$ is the dimension of the data matrix. The rank parameter r is set to be 8 in ICMC. All results are averaged over 100 runs.

We use the first 5-second PMU data and 1-second of PMU data ($t = 2 \sim 3$ s) in dataset #1 for temporal and channel correlations respectively. Figs. 5 and 6 show the relative recovery error of SVT and ICMC for temporally correlated erasures. The recovery performance of SVT is generally better than ICMC. The recovery error of ICMC, however, decreases when the temporal correlation of the erasures increases, ICMC can recover about 35% of erasures with high temporal correlation.

Figs. 7 and 8 show the recovery performance of the SVT and ICMC for spatially correlated erasures. Note that p_{avg} cannot be smaller than $1-q$. The recovery error of SVT increases when the channel correlation increases, while the recovery performance of ICMC has the opposite trend. On a computer with 3.4 GHz Intel Core i7, the average run time of SVT is 0.9 second, while ICMC only needs 0.09 second.

We next run ICMC on four datasets for temporal correlations. We use the first 5-second of four PMU datasets and fix $\tau = 5$. We apply an interpolation method as a benchmark, which estimates the consecutive erasures with the same value by averaging the two nearest observed data in that channel. Table I shows the recovery performance of ICMC for temporal-correlated erasures. The recovery performance of

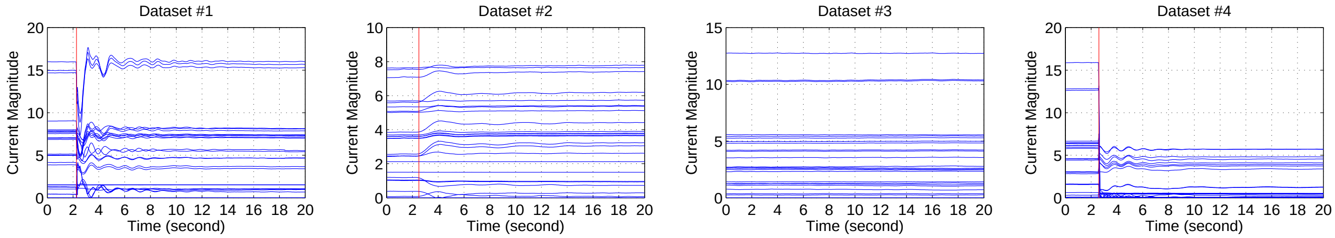

Fig. 4. Current magnitudes in four datasets. The red lines indicate the starting points of some disturbance detected by OLAP

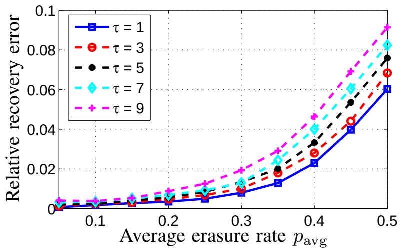

Fig. 5. Relative recovery error of SVT for various value of τ

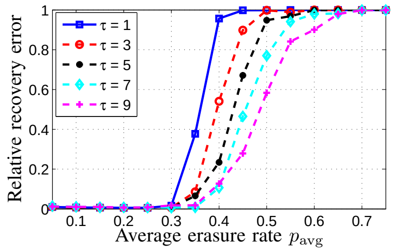

Fig. 6. Relative recovery error of ICMC for various value of τ

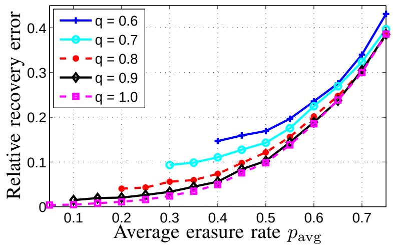

Fig. 7. Relative recovery error of SVT for various value of q

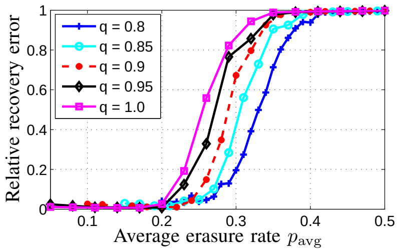

Fig. 8. Relative recovery error of ICMC for various value of q

ICMC is generally better than the interpolation method when p_{avg} is below 0.4. We also compare ICMC and SVT on dataset #1 with the interpolation method for channel correlations. We use 1-second PMU data ($t = 2 \sim 3$ s) in dataset #1 and fix $q = 0.95$. From Table II, we can see that ICMC achieves the best performance among the three algorithms when p_{avg} is below 0.20. SVT performances better than the interpolation method when p_{avg} is below 0.55.

B. Recovery of uniformly random erasures and event detection by OLAP algorithm

We use 20-second actual PMU data to evaluate the performance of OLAP algorithm. The window size w is set to be 30 samples. The relative approximation error threshold γ_{err} is set to be 1%. Note that the performance of OLAP depends on the choice of parameters. We do not attempt to optimize the parameters here. We assume that there is no erasure in the first second. Fig. 9 shows the estimates of the erasures made by the SVT, ICMC and OLAP algorithms for the current magnitudes of PMU dataset #1 ($t = 2 \sim 5$ s) when $p_{\text{avg}} = 0.25$. Table III shows the recovery performance of OLAP on four 20-second datasets with p_{avg} ranging from 0.05 to 0.40. The average run time of OLAP is 0.42 millisecond in each sampling instant on the same computer. Each result is averaged over 100 runs. We can see that OLAP still has good recovery performance in the presence of disturbances.

We set $\bar{w}$ and θ to be 5 (samples) and 20% respectively for event detection. We consider the situation when p_{avg} equals 0.15. Fig. 10 shows the coefficient η along the process of OLAP for dataset #1. When the abnormal event happens, the value of η increases. Table IV shows the starting time of the abnormal event identified by OLAP for four PMU datasets. Notice that dataset #3 only contains ambient data, so no abnormal event is detected. Fig. 4 records the starting time determined by OLAP for four datasets. We can see that OLAP algorithm achieves good detection performance for the abnormal event.

We compare the performance of SVT, ICMC and OLAP on 1-second PMU data ($t = 1 \sim 2$ s in dataset #1). In ICMC, r is set to be 8. In OLAP γ_{err} is 1%, and window size w is 30 (samples). Fig. 11 shows the relative recovery error of the SVT, ICMC and OLAP algorithms. SVT and ICMC perform slightly better than OLAP. This is not surprising since the former two are offline methods while OLAP is an online algorithm which does not use future data.

C. Data compression of PMU data by SVD-based method

We evaluate the compression ratio by the SVD-based compression method. We divide the 20-second data into twenty segments and apply Subroutine 1 to determine the approximate rank of each sub-matrix. The relative approximation error threshold γ_{err} is set to be 1%. Then only the dominant singular values and the corresponding singular vectors are stored. Fig.

TABLE I

Recovery error of ICMC and interpolation method (shown in parentheses) for temporally correlated erasures

p_{avg}	Dataset #1	Dataset #2	Dataset #3	Dataset #4
0.05	0.0056 (0.0108)	0.0006	0.0003	0.0406
0.10	0.0061 (0.0174)	0.0006	0.0003	0.0428
0.15	0.0073 (0.0267)	0.0007	0.0004	0.0516
0.20	0.0074 (0.0296)	0.0008	0.0004	0.0738
0.25	0.0077 (0.0377)	0.0008	0.0004	0.1458
0.30	0.0079 (0.0440)	0.0009	0.0005	0.1697
0.35	0.0084 (0.0531)	0.1005	0.0005	0.3109
0.40	0.3023 (0.0613)	0.1009	0.0009	0.5604

TABLE II

Recovery error of ICMC, SVT and interpolation method for spatially correlated erasures on dataset #1

p_{avg}	ICMC	SVT	Interpolation method
0.05	0.0033	0.0041	0.0114
0.10	0.0079	0.0085	0.0218
0.15	0.0098	0.0110	0.0300
0.20	0.0193	0.0127	0.0398
0.25	0.0315	0.0160	0.0474
0.30	0.1577	0.0240	0.0562
0.35	0.6650	0.0253	0.0680
0.40	0.9939	0.0383	0.0805
0.45	0.9956	0.0608	0.0902
0.50	0.9977	0.0971	0.1058
0.55	0.9984	0.1335	0.1143
0.60	0.9985	0.1845	0.1326
0.65	1.0000	0.2278	0.1521

TABLE III

Relative recovery error of OLAP on four datasets

p_{avg}	Dataset #1	Dataset #2	Dataset #3	Dataset #4
0.05	0.0082	0.0026	0.0006	0.0186
0.10	0.0089	0.0039	0.0009	0.0235
0.15	0.0119	0.0053	0.0011	0.0442
0.20	0.0145	0.0060	0.0013	0.0494
0.25	0.0153	0.0066	0.0014	0.0726
0.30	0.0191	0.0075	0.0017	0.0779
0.35	0.0204	0.0082	0.0018	0.1137
0.40	0.0227	0.0091	0.0019	0.1189

TABLE IV

Event detection for four PMU datasets by OLAP

Time	Dataset #1	Dataset #2	Dataset #3	Dataset #4
Start	2.3s	3.2s	No disturbance	2.6s

12 shows the compression ratio of twenty sub-matrices for four datasets. The compression ratio for ambient data is 6%. Even during the disturbances in dataset #1 and #4, the compression ratio is still about 30%.

VI. Conclusion and Discussions

We formulate the missing data recovery problem into a low-rank matrix completion problem and demonstrate the effectiveness of matrix-completion methods on recovering missing PMU measurements. We propose two models to characterize the correlations in PMU data erasures and provide the theoretical guarantee for the recovery of correlated erasures. We also propose an online missing data recovery method that can be easily extended to data compression and event detection.

Our theoretical bound of the required number of samples to recover a low-rank matrix when the erasures are correlated is higher than the existing bound under the independent erasure model. One interesting direction for future work is to improve the bound for correlated erasures. We have not considered bad data in this paper. We are currently developing methods that can identify and correct the bad data in PMU measurements.

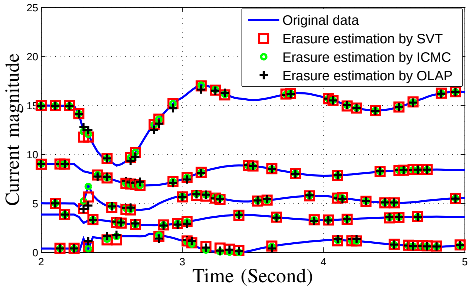

Fig. 9. PMU dataset #1 ($t = 2 \sim 5s$) and erasure estimations made by SVT, ICMC and OLAP algorithms

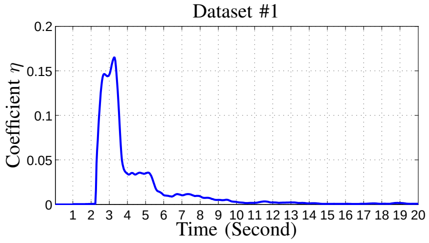

Fig. 10. Coefficient η along the process of OLAP for dataset #1

Acknowledgement

We thank New York Power Authority for providing the PMU data. This research is supported in part by the ERC Program of NSF and DoE under NSF Award EEC-1041877 and the CURENT Industry Partnership Program, and in part by NYSERDA Grants #28815 and #36653.

References

[1] "The netflix prize," June 2006. [Online]. Available: http://www.netflixprize.com/

[2] L. Balzano, R. Nowak, and B. Recht, "Online identification and tracking of subspaces from highly incomplete information," *Allerton Conf. Commun., Control Comput.*, pp. 704–711, 2010.

[3] J. Burnett, R.O., M. Butts, T. W. Cease, V. Centeno, G. Michel, R. J. Murphy, and A. Phadke, "Synchronized phasor measurements of a power system event," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1643–1650, 1994.

[4] 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.

[5] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053 –2080, 2010.

[6] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[7] 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.*, vol. 26, no. 8, pp. 1051–1063, 2004.

[8] Y. Chen, L. Xie, and P. Kumar, "Dimensionality reduction and early event detection using online synchrophasor data," in *Proc. IEEE Power and Energy Society General Meeting*, 2013, pp. 1–5.

[9] 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, 2013.

[10] 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)*, 2012, pp. 1–7.

[11] S. Das and P. S. N. Rao, "Principal component analysis based compression scheme for power system steady state operational data," *IEEE PES Innovative Smart Grid Technologies-India*, 2011.

[12] E. Farantatos, G. Stefopoulos, G. Cokkinides, and A. Meliopoulos, "PMU-based dynamic state estimation for electric power systems," in *Proc. IEEE Power Energy Society General Meeting*, 2009, pp. 1–8.

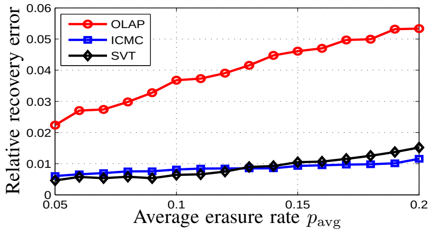

Fig. 11. Comparison of SVT, ICMC, and OLAP

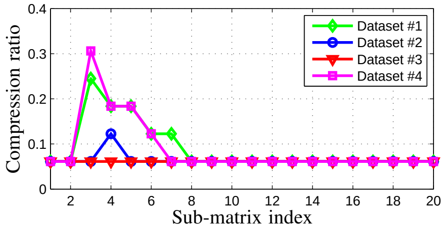

Fig. 12. Compression ratio of sub-matrices when γ_{err} is 1%

[13] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, Stanford University, 2002.

[14] F. Gao, J. Thorp, A. Pal, and S. Gao, "Dynamic state prediction based on Auto-Regressive (AR) model using PMU data," in *Proc. IEEE Power and Energy Conference at Illinois (PECI)*, 2012, pp. 1–5.

[15] P. Gao, M. Wang, S. Ghiocel, and J. H. Chow, "Modeless reconstruction of missing synchrophasor measurements," in *Proc. IEEE PES General Meeting*, 2014.

[16] R. Gardner, J. Wang, and Y. Liu, "Power system event location analysis using wide-area measurements," in *Proc. IEEE PES General Meeting*, 2006, pp. 1–7.

[17] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[18] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Advances in Neural Information Processing Systems*, 2010, pp. 937–945.

[19] I. Kamwa and L. Gerin-Lajoie, "State-space system identification-toward MIMO models for modal analysis and optimization of bulk power systems," *IEEE Trans. Power Syst.*, vol. 15, no. 1, pp. 326–335, 2000.

[20] K. Lee and Y. Bresler, "ADMiRA: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4402–4416, 2010.

[21] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.

[22] M. Mardani, G. Mateos, and G. B. Giannakis, "Dynamic anomalography: Tracking network anomalies via sparsity and low rank," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 1, pp. 50–66, 2013.

[23] R. Meka, P. Jain, and I. S. Dhillon, "Matrix completion from power-law distributed samples," in *Advances in Neural Information Processing Systems*, 2009, pp. 1258–1266.

[24] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[25] J. Ning, J. Wang, W. Gao, and C. Liu, "A wavelet based data compression technique for smart grid," *IEEE Trans. Smart Grid*, vol. 2, no. 1, 2011.

[26] T. Overbye, P. Sauer, C. DeMacro, B. Lesieutre, and M. Venkatasubramanian, "Using PMU data to increase situational awareness," *PSERC Final Project Report*, 2010.

[27] A. Phadke, J. Thorp, and K. Karimi, "State estimlatjon with phasor measurements," *IEEE Trans. Power Syst.*, vol. 1, no. 1, pp. 233–238, 1986.

[28] A. Phadke, J. Thorp, R. Nuqui, and M. Zhou, "Recent developments in state estimation with phasor measurements," in *Proc. IEEE Power Systems Conference and Exposition*, 2009, pp. 1–7.

[29] B. Recht, "A simpler approach to matrix completion," *The Journal of Machine Learning Research*, pp. 3413–3430, 2011.

[30] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[31] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[32] L. Vanfretti, J. Chow, S. Sarawgi, and B. Fardanesh, "A phasor-data-based state estimator incorporating phase bias correction," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 111–119, 2011.

[33] 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, 2013.

[34] M. Zhao and A. Abur, "Multiarea state estimation using synchronized phasor measurements," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 611–617, 2005.

[35] N. Zhou, J. Pierre, and J. Hauer, "Initial results in power system identification from injected probing signals using a subspace method," *IEEE Trans. Power Syst.*, vol. 21, no. 3, pp. 1296–1302, 2006.

Pengzhi Gao (S'14) received the B.E. degree from Xidian University, Xian, China, in 2011 and the M.S. degree in electrical engineering from University of Pennsylvania, Philadelphia, PA, in 2013.

He is pursuing the Ph.D. degree in electrical engineering at Rensselaer Polytechnic Institute, Troy, NY. His research interests include signal processing, compressive sensing, low-rank matrix recovery, and power networks.

Meng Wang (M'12) received the Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2012.

She is an Assistant Professor in the department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. Her research interests include high dimensional data analysis and their applications in power systems monitoring and network inference.

Scott G. Ghiocel (S'08) received the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2013.

He is a Postdoctoral Research Fellow at Rensselaer Polytechnic Institute. His research interests include synchronized phasor measurements, voltage stability, and power system dynamics.

Joe H. Chow (F'92) received the M.S. and Ph.D. degrees from the University of Illinois, Urbana-Champaign, Urbana, IL, USA.

After working in the General Electric power system business in Schenectady, NY, USA, he joined Rensselaer Polytechnic Institute, Troy, NY, USA, in 1987, where he is a Professor of Electrical, Computer, and Systems Engineering. His research interests include multivariable control, power system dynamics and control, FACTS controllers, and synchronized phasor data.

Bruce Fardanesh (F'13) received his Doctor of Engineering degree in Electrical Engineering from Cleveland State University in 1985.

He joined New York Power Authority in 1991, where he is the Chief Electrical Engineer. His research areas of interest are power system analysis, modeling, dynamics, operation, and control.

George Stefopoulos (M'08) received his Ph.D. degree in Electrical Engineering from the Georgia Institute of Technology in 2009.

He is a Research and Technology Development Engineer with the New York Power Authority. His research interests include power system state estimation, synchrophasor technology applications, and modeling and simulation of power systems.