Machine Learning and Big Data Analytics in Power Distribution Systems

Dr. Nanpeng Yu

Department of Electrical and Computer Engineering

Department of Computer Science

(cooperating faculty)

nyu@ece.ucr.edu

951.827.3688

UNIVERSITY OF CALIFORNIA, RIVERSIDE

Team Members and Research Sponsors

- Principal Investigator
  - Dr. Nanpeng Yu
- Ph.D. Students
  - Brandon Foggo (B.S. UCLA), Wei Wang (M.S. University of Michigan)
  - Yuanqi Gao (B.S. UCR), Wenyu Wang (M.S. Iowa State University)
  - Jie Shi (M.S. Southeast University), Farzana Kabir (B.S. BUET)
  - Yinglun Li (M.S. UCR)
- Research Sponsors and Collaborating Organizations

NSF

SOUTHERN CALIFORNIA EDISON®
An EDISON INTERNATIONAL® Company

EPRI ELECTRIC POWER RESEARCH INSTITUTE

PG&E Pacific Gas and Electric Company®

U.S. DEPARTMENT OF ENERGY

NREL NATIONAL RENEWABLE ENERGY LABORATORY

FORTIS BC™

Computing Facilities

- Deep Learning Workstation
  - 4 x NVIDIA RTX 2080
  - 4 x 16 GB Memory
  - 512 GB SSD (OS)
  - 2 x 2TB HDD (Data)

- Oracle Big Data Appliance
  - Number of Nodes: 6
  - Number of Core: 216
  - Hard Drive: 288 TB of 7,200 rpm High Capacity SAS Disks
  - Memory: 768 GB DDR4
  - Hadoop Platform: CDH Enterprise Edition
  - Tools: Hive, Pig, Impala, PySpark, Scala, TensorFlow

Outline

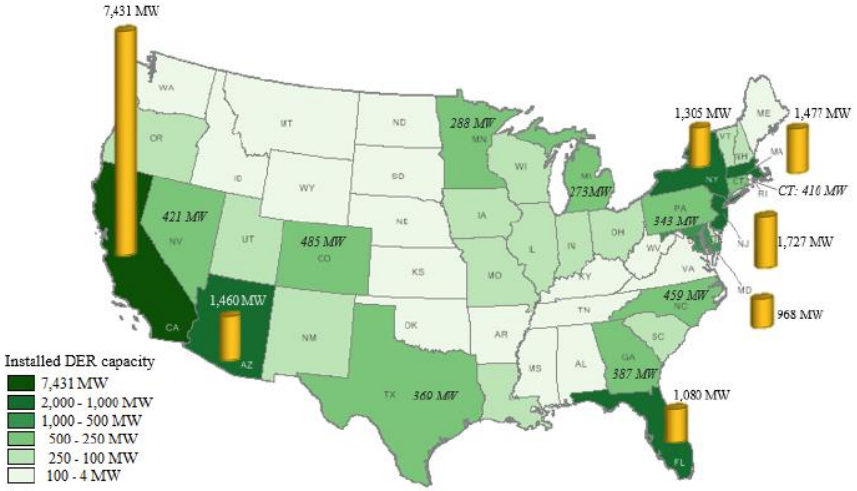
- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

https://intra.ece.ucr.edu/~nyu/Teaching/2019-03-28-IEEE-Big-Data-Seminar

Why focus on distribution systems?

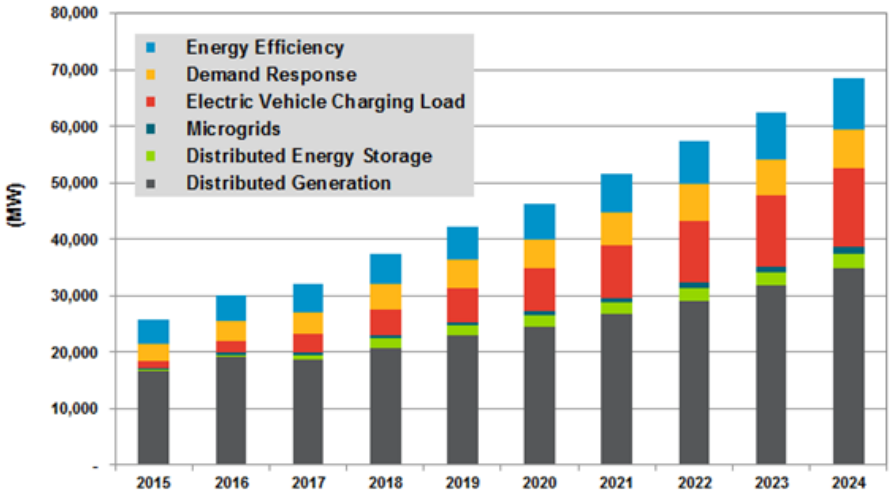
- Increasing penetrations of distributed energy resource (DER) in power distribution systems
  - On a 5-year basis (2015-2019), DER in US is growing almost 3 times faster than central generation (168 GW vs. 57 GW).
  - In 2016, distributed solar PV installations alone represented 12% of new capacity additions.
  - California DER, 7GW in 2017, 12 GW by 2020 (peak load 50 GW)

U.S. DER Deployments



Source: The U.S. EIA and FERC DER Staff Report

Annual Installed DER Power Capacity Additions by DER Technology, United States: 2015-2024



Source: Navigant Report, Take Control of Your Future

The need for advanced modeling, monitoring, and control in distribution systems

- The cold hard facts about modern power distribution systems
  - Modeling
    - Incomplete topology information in the secondary systems
      - Phase connection
      - Transformer-to-customer mapping
    - Even the three-phase load flow results are unreliable!
  - Monitoring
    - Most utilities do not have online three-phase state estimation for their entire distribution network
  - Control
    - Reactive Control
      - System restoration, equipment maintenance
    - Limited Proactive Control
      - Volt-VAR control, CVR, network reconfiguration

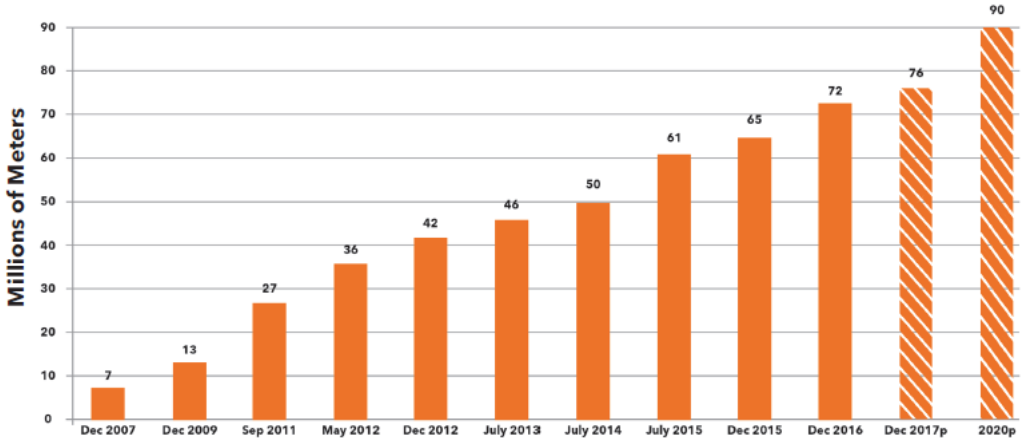
Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

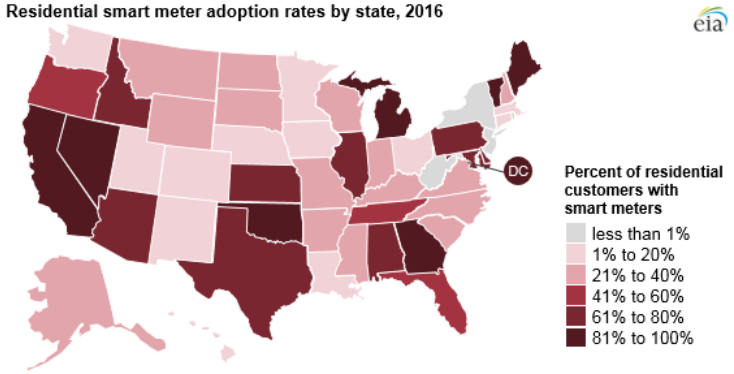
Big Data in Distribution Systems: Volume

- In 2017, the U.S. electric utilities had about 78.9 million AMI installations covering over 50% of 150 million electricity customers.
- The smart meter installation worldwide will surpass 1.1 billion by 2022.
- In 2012, the AMI data collected in the U.S. alone amounted to well above 100 terabytes.
- By 2022, the electric utility industry will be swamped by more than 2 petabytes of meter data alone.

U.S. Smart Meter Installations Projected to Reach 90 Million by 2020



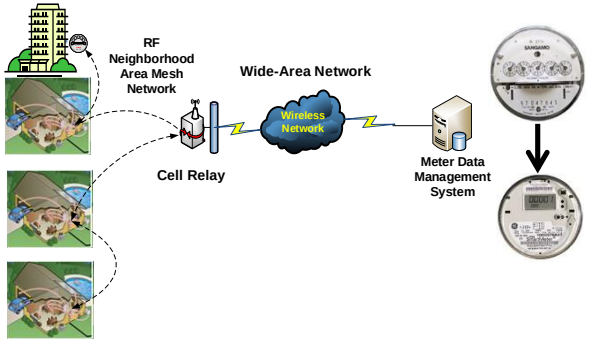
Source: Institute for Electric Innovation



Source: U.S. Energy Information Administration

Big Data in Distribution Systems: Variety

- Advanced Metering Infrastructure
  - Electricity usage (15-minute, hourly)
  - Voltage magnitude
- Weather Station
- Geographical Information System
- Census Data (block group level)
  - Household variables: ownership, appliance, # of rooms
  - Person variables: age, sex, race, income, education
- SCADA Information
- Micro-PMU
  - Time synchronized measurements with phase angles
- Equipment Monitors

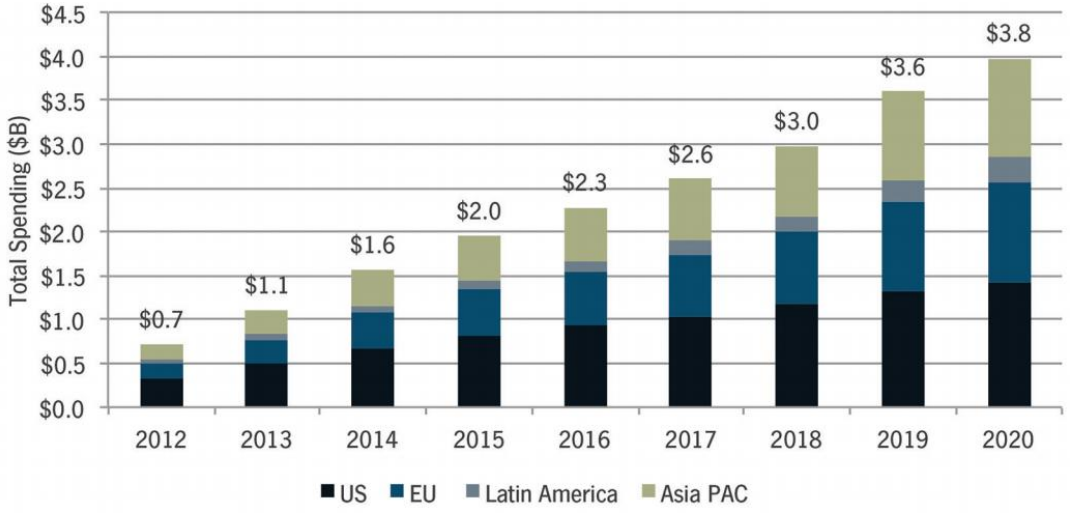


Big Data in Distribution Systems: Velocity

- Sampling Frequency
  - AMI's data recording frequency increases from **once a month** to one reading **every 15 minutes to one hour**.
  - Micro-PMU hundreds (512) of samples per cycle at 50/60 Hz
- Bottleneck in Communication Systems
  - Limited bandwidth for zigbee network
  - Most of the utilities in the US receives smart meter data with ~24 hour delay
- Edge Computing Trend
  - Itron and Landis+Gyr extend edge computing capability of smart meters
  - Increasing data transmission range and computing capabilities of smart meters
  - Centralized → distributed / decentralized

Big Data in Distribution Systems: Value

- The big data collected in the power distribution system had utterly swamped the traditional software tools used for processing them.
- Lack of innovative use cases and applications to unleash the full value of the big data sets in power distribution systems[1].
- Insufficient research on machine learning and big data analytics for power distribution systems.
- Electric utilities around the world will spend over $3.8 billion on data analytics solutions in 2020.



Source: GTM Research

1. Nanpeng Yu, Sunil Shah, Raymond Johnson, Robert Sherick, Mingguo Hong and Kenneth Loparo, “Big Data Analytics in Power Distribution Systems” *IEEE PES ISGT*, Washington DC, Feb. 2015.

Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

Applications of Big Data Analytics and Machine Learning in Power Distribution Systems

Spatio-temporal Forecasting

Electric Load / DERs – Short-Term / Long-Term

Anomaly Detection

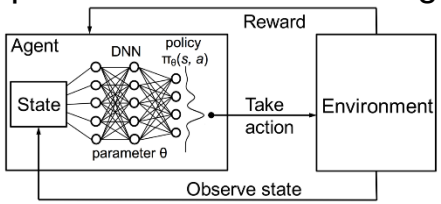
Electricity Theft, Unauthorized Solar Interconnection

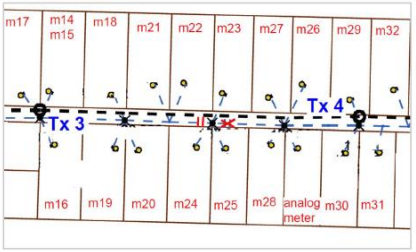
System Monitoring

State Estimation & Visualization

Distribution System Controls

Deep Reinforcement Learning





Network Topology and Parameter Identification

Transformer-to-customer, Phase connectivity, Impedance estimation

Equipment Monitoring

Predictive Maintenance
Online Diagnosis

Customer Behavior Analysis

Customer segmentation, nonintrusive load monitoring, demand response

Publications: Big Data Analytics & Machine Learning in Smart Grid

1. N. Yu, S. Shah, R. Johnson, R. Sherick, Mingguo Hong and Kenneth Loparo, "Big Data Analytics in Power Distribution Systems", IEEE PES Conference on Intelligent Smart Grid Technology, Washington DC, Feb. 2015.

2. Xiaoyang Zhou, Nanpeng Yu, Weixin Yao and Raymond Johnson, “Forecast load impact from demand response resources” *Power and Energy Society General Meeting*, pp. 1-5, Boston, USA, 2016.

3. W. Wang, N. Yu, B. Foggo, and J. Davis, “Phase identification in electric power distribution systems by clustering of smart meter data” *15th IEEE International Conference on Machine Learning and Applications* (ICMLA), pp. 1-7, Anaheim, CA, 2016.

4. Jie Shi and Nanpeng Yu, “Spatio-temporal modeling of electric loads” in *49th North American Power Symposium*, pp.1-6, Morgantown, WV, 2017.

5. W. Wang, N. Yu, and R. Johnson “A model for commercial adoption of photovoltaic systems in California” *Journal of Renewable and Sustainable Energy*, Vol. 9, Issue, 2, pp.1-15, 2017.

6. Yuanqi Gao and Nanpeng Yu, “State estimation for unbalanced electric power distribution systems using AMI data” *The Eighth Conference on Innovative Smart Grid Technologies (ISGT 2017)*, pp. 1-5, Arlington, VA.

7. Wenyu. Wang and Nanpeng Yu, "AMI Data Driven Phase Identification in Smart Grid," *the Second International Conference on Green Communications, Computing and Technologies*, pp. 1-8, Rome, Italy, Sep. 2017.

8. Jinhui Yang, Nanpeng Yu, Weixin Yao, Alec Wong, Larry Juang, and Raymond Johnson, “Evaluate the effectiveness of CVR with robust regression” in *Probabilistic Methods Applied to Power Systems*, pp.1-6, 2018.

9. Brandon Foggo, Nanpeng Yu, “A comprehensive evaluation of supervised machine learning for the phase identification problem”, *the 20th International Conference on Machine Learning and Applications*, pp.1-9, Copenhagen, Denmark, 2018.

10. Ke Wang, Haiwang Zhong, Nanpeng Yu, and Qing Xia, “Nonintrusive load monitoring based on sequence-to-sequence model with attention mechanism”, *Proceedings of the CSEE*, 2018.

11. Farzana Kabir, Brandon Foggo, and Nanpeng Yu, "Data Driven Predictive Maintenance of Distribution Transformers," in *the 8th China International Conference on Electricity Distribution*, pp. 1-5 2018.

12. Wei Wang and Nanpeng Yu, " A Machine Learning Framework for Algorithmic Trading with Virtual Bids in Electricity Markets," to appear in *IEEE Power and Energy Society General Meeting*, 2019.

13. Yuanqi Gao, Brandon Foggo, and Nanpeng Yu, “A physically inspired data-driven model for electricity theft detection with smart meter data” to appear in *IEEE Transactions on Industrial Informatics,* 2019.

14. Wang, Wenyu, and Nanpeng Yu. "Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems." *arXiv preprint arXiv:1902.09686* (2019).

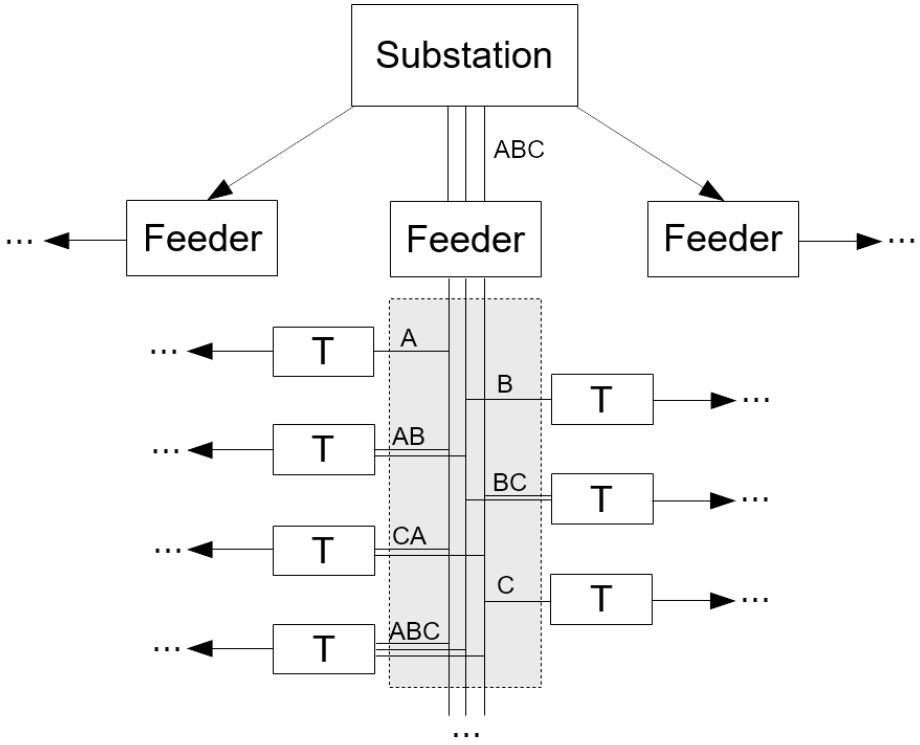
https://intra.ece.ucr.edu/~nyu/

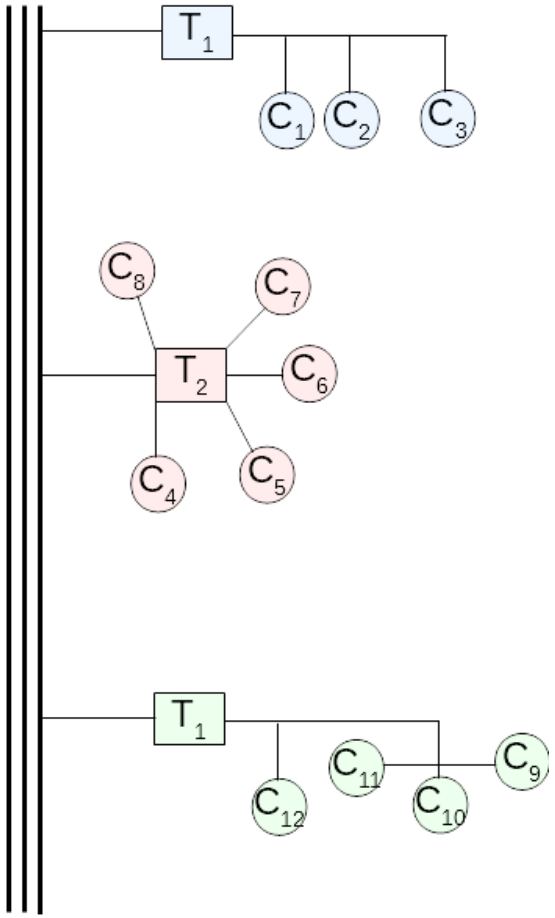
Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

Distribution System Topology Identification

- The distribution system topology identification problem can be broken down into two sub-problems
  - The phase connectivity identification problem
  - The customer to transformer association problem



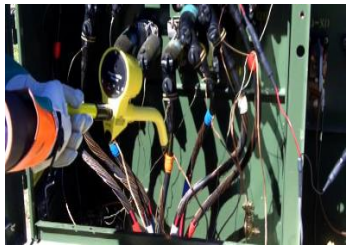


Phase Connectivity Identification

- Problem Definition
  - Identify the phase connectivity of each customer & structure in the power distribution network.
  - Very few electric utility companies have completely accurate phase connectivity information in GIS!
- Why is it important? (Business Value)
  - Phase connectivity is crucial to an array of distribution system analysis & operation tools including
    - 3-phase Power flow
    - Load balancing
    - Distribution network state estimation
    - 3-phase optimal power flow
    - Volt-VAR control
    - Distribution network reconfiguration and restoration

Phase Connectivity Identification

- Primary Data Set
  - Advanced Metering Infrastructure, SCADA, GIS, OMS
  - Training data (field validated phase connectivity)
- Solution Methods
  - Physical approach with Special Sensors
    - Micro-synchrophasors, Phase Meters
    - Drawback: expensive equipment, labor intensive ($2,000 per feeder), 3,000 feeders for a regional electric utility company ($6 million)

Phase Connectivity Identification

- Solution Methods
  - Integer Optimization, Regression and Correlation based Approach
    - 0-1 integer linear programming (IBM)
    - Correlation/Regression based methods (EPRI)
    - Drawback: cannot handle delta connected Secondaries, low tolerance for erroneous or missing data, low accuracy and high computational cost
  - Data-driven phase identification technology
    - Synergistically combine machine learning techniques and physical understanding of electric power distribution networks.
    - Unsupervised and supervised machine learning algorithms
    - High accuracy on all types of distribution circuits. (overhead, underground, phase-to-neutral, phase-to-phase, pilot demonstration on over 100 distribution feeders)

SOUTHERN CALIFORNIA EDISON®
An EDISON INTERNATIONAL® Company

PG&E Pacific Gas and Electric Company®

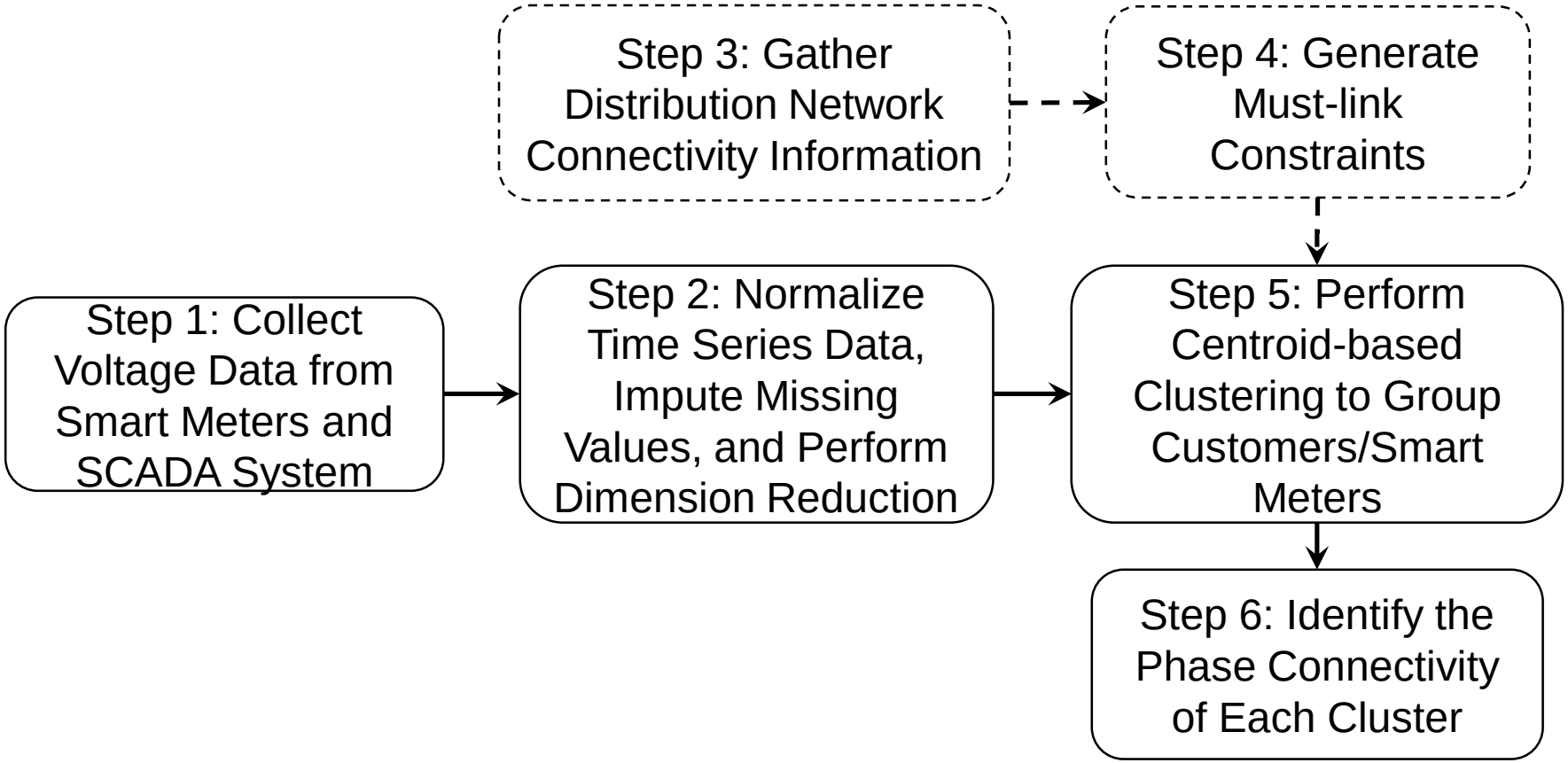
FORTIS BC™

Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

Unsupervised Machine Learning Algorithm[1]

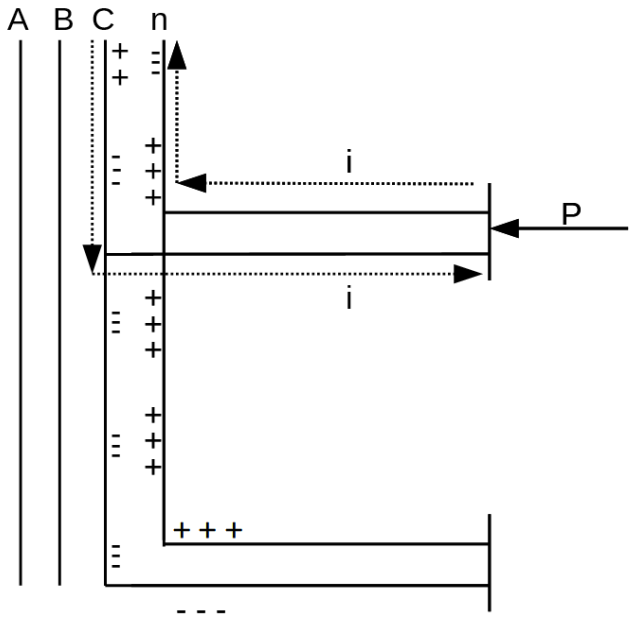
General Framework



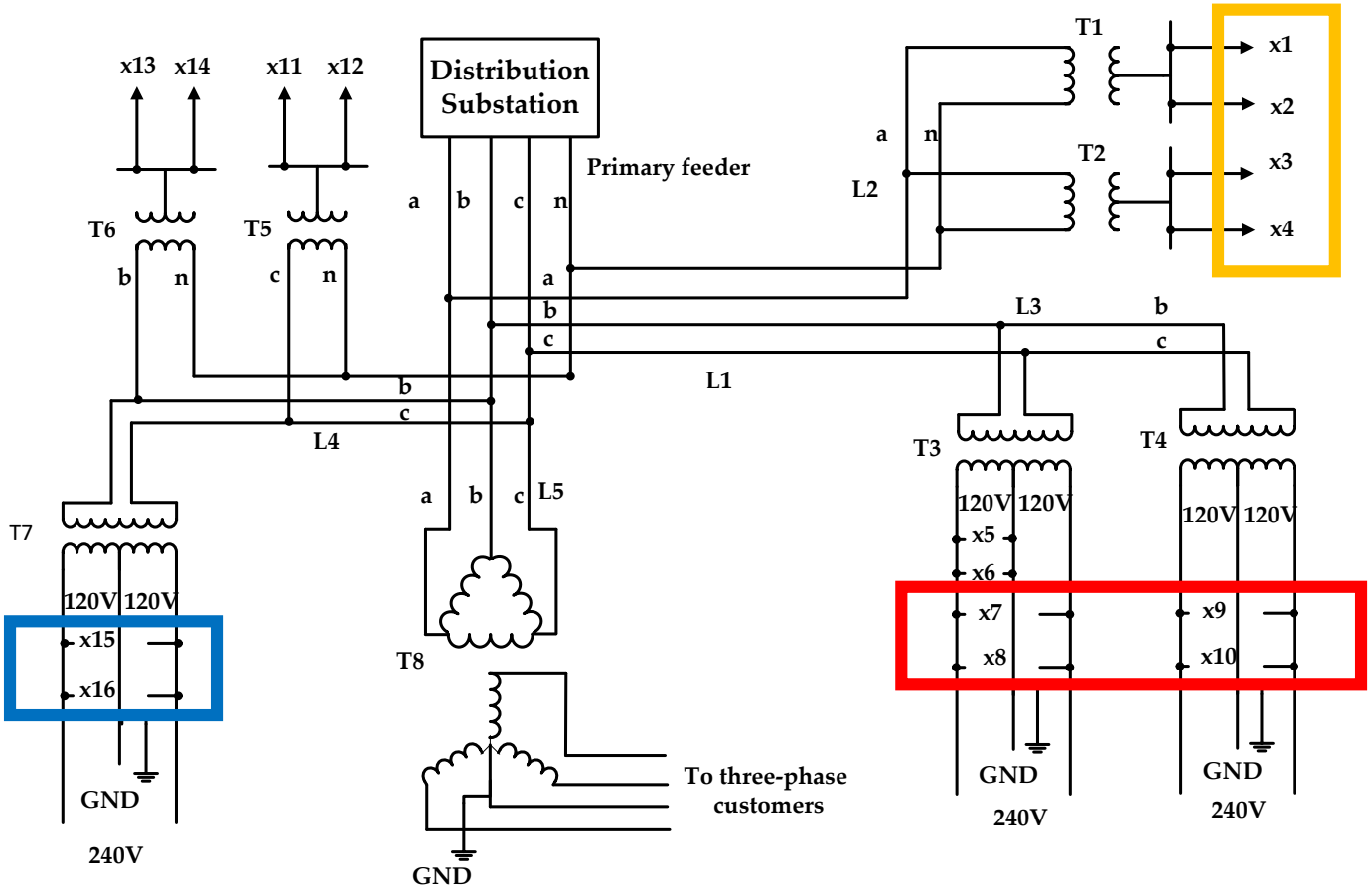
1. W. Wang, N. Yu, B. Foggo, and J. Davis, "Phase identification in electric power distribution systems by clustering of smart meter data" *15th IEEE International Conference on Machine Learning and Applications* (ICMLA), pp. 1-7, Anaheim, CA, 2016.

Why Voltage Data Is Predictive of Phase?

- Voltage data is fairly informative of phase type
  - Consider a power injection at bus k whose phase type is AB.
  - This induces a current along the lines A and B.
  - Any customer also feeding from either of those lines will notice a change.
  - Due to the capacitive and inductive effects of the primary feeder, both lines will also induce a voltage change along the lines C and n.
  - However, the off-diagonal elements of the phase impedance and shunt admittance matrices are much smaller than the diagonal ones.
  - Hence, the power injection at bus k will have much less effect on phase C than phase A and B.



Must-link Constraints



Customers connected to the same secondary must have the same phase connections

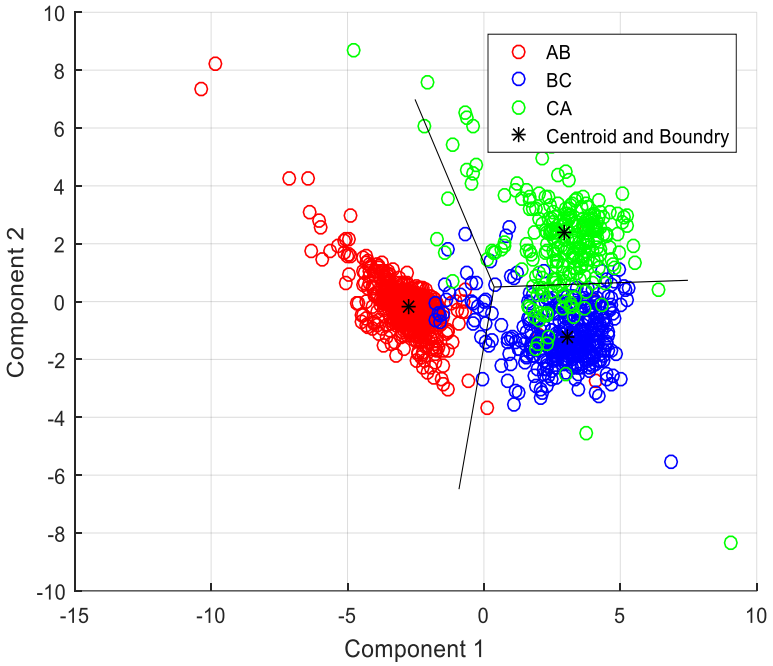
Case Study: Southern California Edison Distribution Circuit

Voltage Level	12.47 kV
Peak load	~5 MW
Number of Customers	~1500
Customer type	95% residential

- Most of the customers served by a three-wire single-phase system through center-tapped transformers (120/240 V).
- Highly unbalanced in terms of phase currents.
- 6 month of smart meter data and SCADA data.
- Engineers gather actual phase connectivity of each building and structure through field validation.

Unsupervised Learning: Unconstrained Clustering

Phase Identification Accuracy: **92.89%**

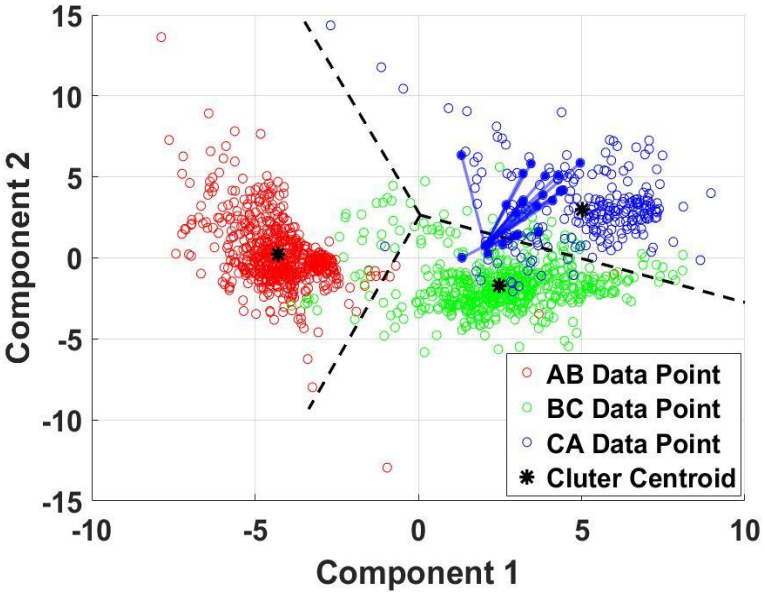


Cluster number	Number of customers	Accuracy (%)	Phase
1	226	94.25	CA
2	647	95.21	AB
3	364	87.91	BC

- The circuit is highly unbalanced and has 3 possible phase connections.
- Even linear dimension reduction technique results in reasonable separation among customers with different phase connections.

Supervised Learning: Constrained Clustering

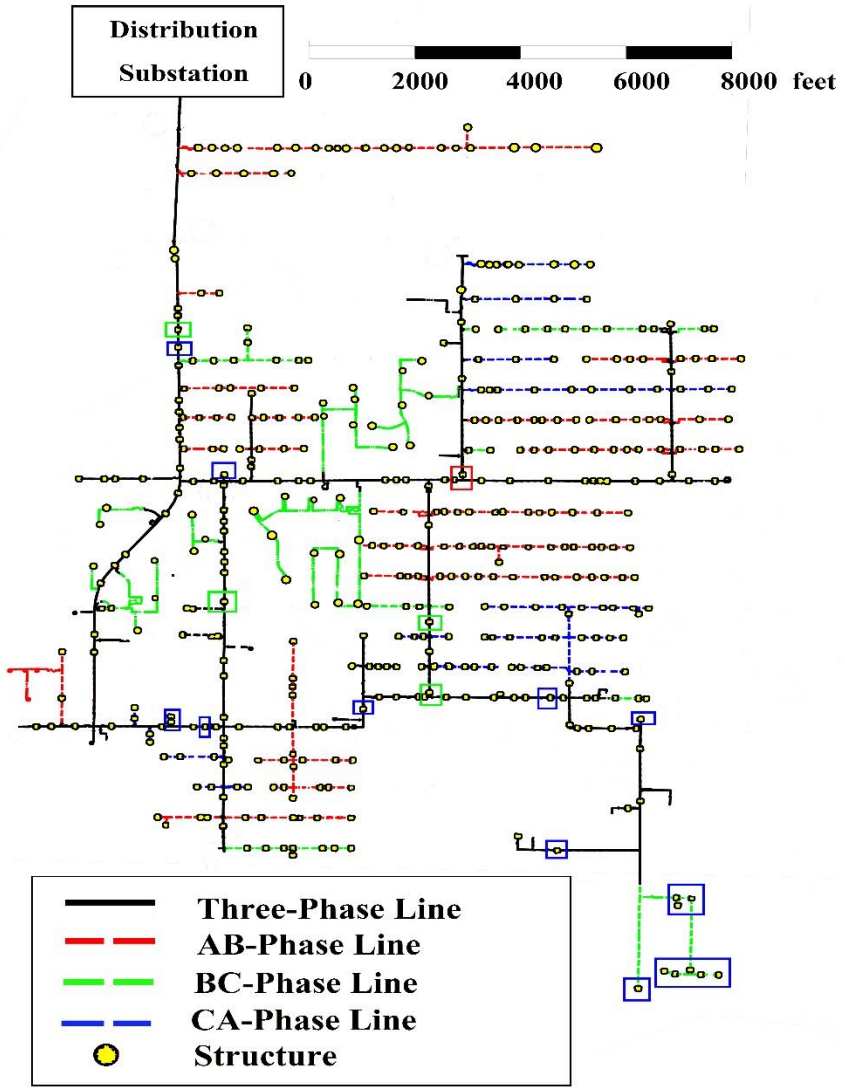
Phase Identification Accuracy: **96.69%**



Cluster number	Number of customers	Accuracy (%)	Phase
1	618	99.84	AB
2	384	91.41	BC
3	235	97.02	CA

- The must-link constraints pulled some of the blue points (customers with phase connections of CA) in the green region back to the blue area.
- The must-link constraints improve the phase identification accuracy.

Visualization of Phase Identification Accuracy



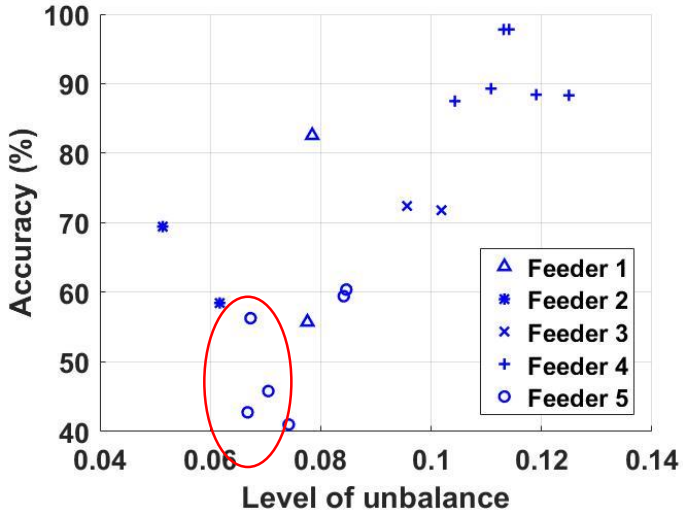
- With GIS inputs, visualization of distribution circuit with phase connection information can be generated automatically
  - Each line is colored according to its actual phase
  - Each structure is represented by a small dot
  - A colored rectangle is overlaid on top of a structure if it is assigned to the wrong cluster.

Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

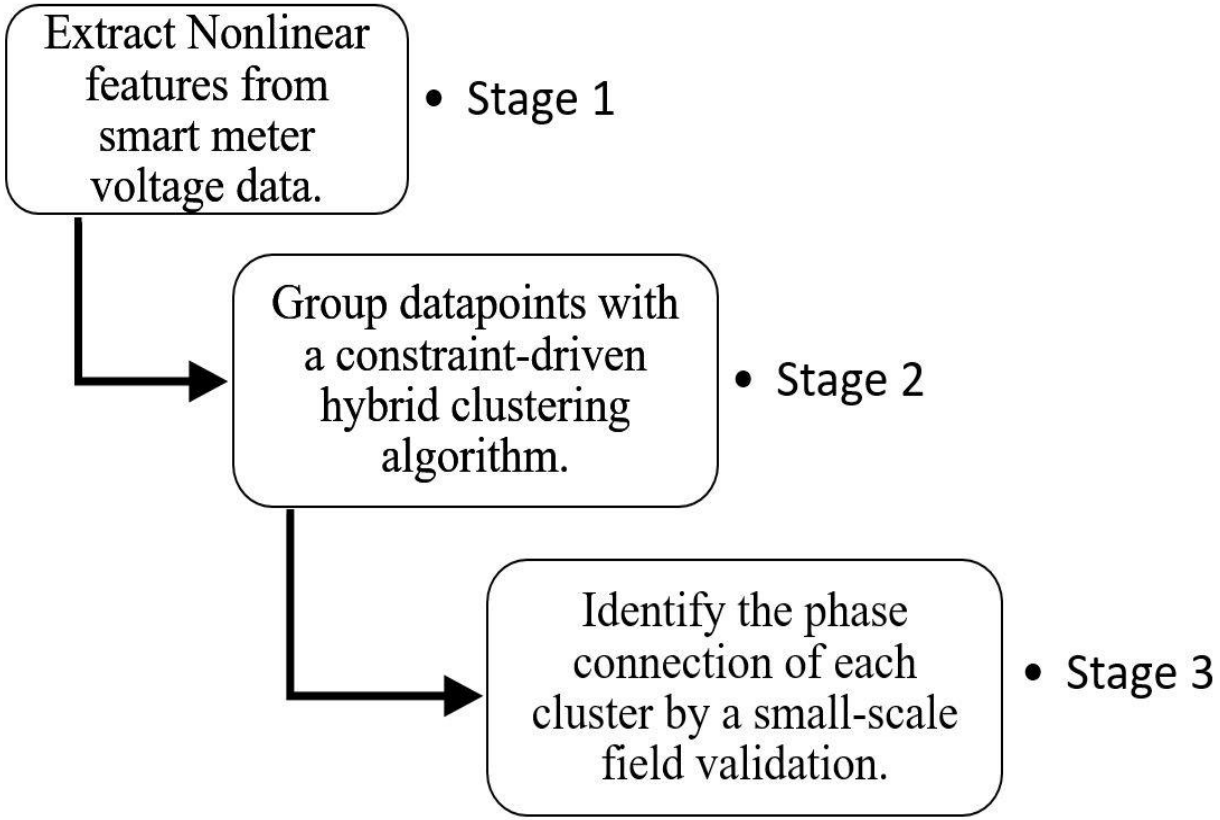
Drawbacks of Constrained K-means Clustering Algorithm (CK-Means)

- First, all of the prior proposed methods assume that the number of phase connections are known.
  - E.g., in the CK-Means algorithm, the number of phase connections/clusters needs to be know as prior knowledge
- Second, the existing methods can not provide accurate phase identification results when there is a mix of phase-to-neutral and phase-to-phase connected smart meters and structures.
  - The phase identification accuracy decreases as the number of possible phase connection increases.
- Third, the existing methods are quite sensitive to the level of unbalance in a distribution feeder.
  - The phase identification accuracy decrease as the level of unbalance decreases.



Nonlinear Dimension Reduction & Density-based Clustering[2]

General Framework



2. W. Wang and N. Yu, "AMI Data Driven Phase Identification in Smart Grid," the Second International Conference on Green Communications, Computing and Technologies, pp. 1-8, Rome, Italy, Sep. 2017.

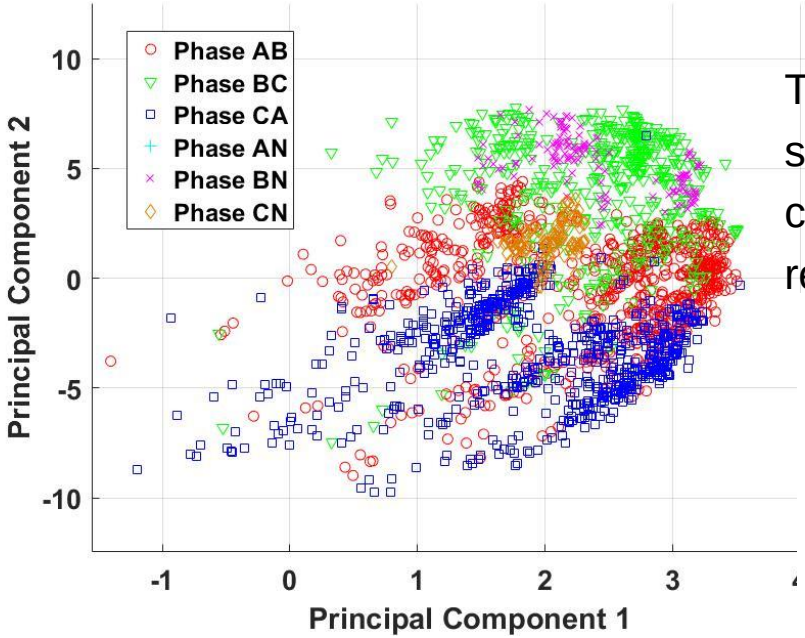
Stage 1 Feature Extraction from Voltage Time Series

- Dimension reduction techniques
  - Linear dimension reduction techniques (E.g., PCA)
    - Drawbacks
      1. Restricted to learning only linear manifolds. High-dimensional data lies on or near a low-dimensional, non-linear manifold.
      2. Difficult for linear mappings to keep the low-dimensional representations of very similar points close together.
    - Explains the lower accuracy of phase identification algorithm using linear features for less unbalanced feeders.
  - Nonlinear dimensionality reduction techniques
    - Sammon mapping, curvilinear components analysis (CCA), Isomap, and t-distributed stochastic neighbor embedding (t-SNE).
    - We adopt t-SNE, because it has been shown to work well with a wide range of data sets and captures both local and global data structures.
    - t-SNE improves upon SNE by

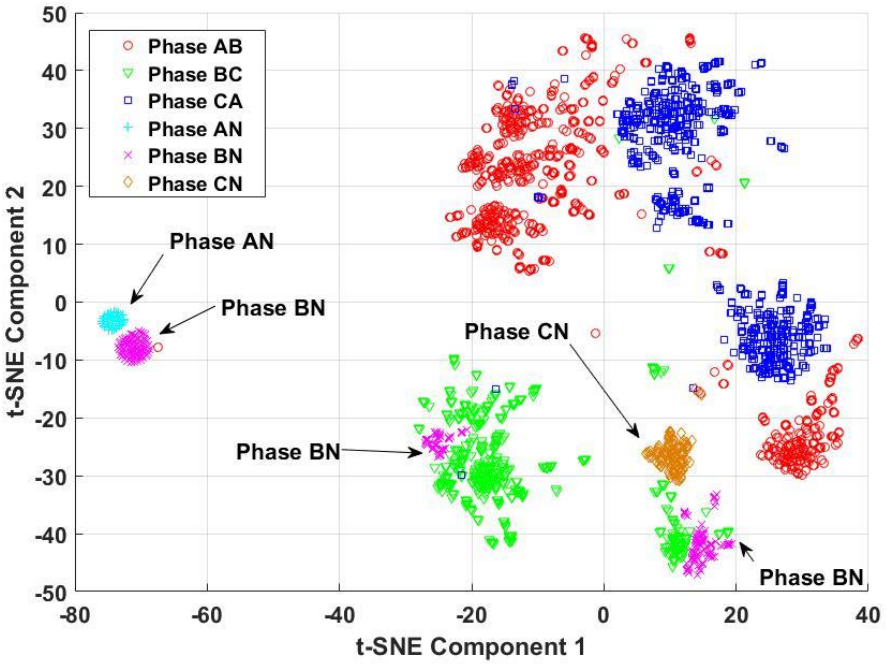
    1. Simplifying the gradient calculation with a symmetrized version of the SNE cost function
    2. Adopting a Student-t distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space

Comparison between PCA & t-SNE

Feeder 5, data set 18 with a low level of unbalance

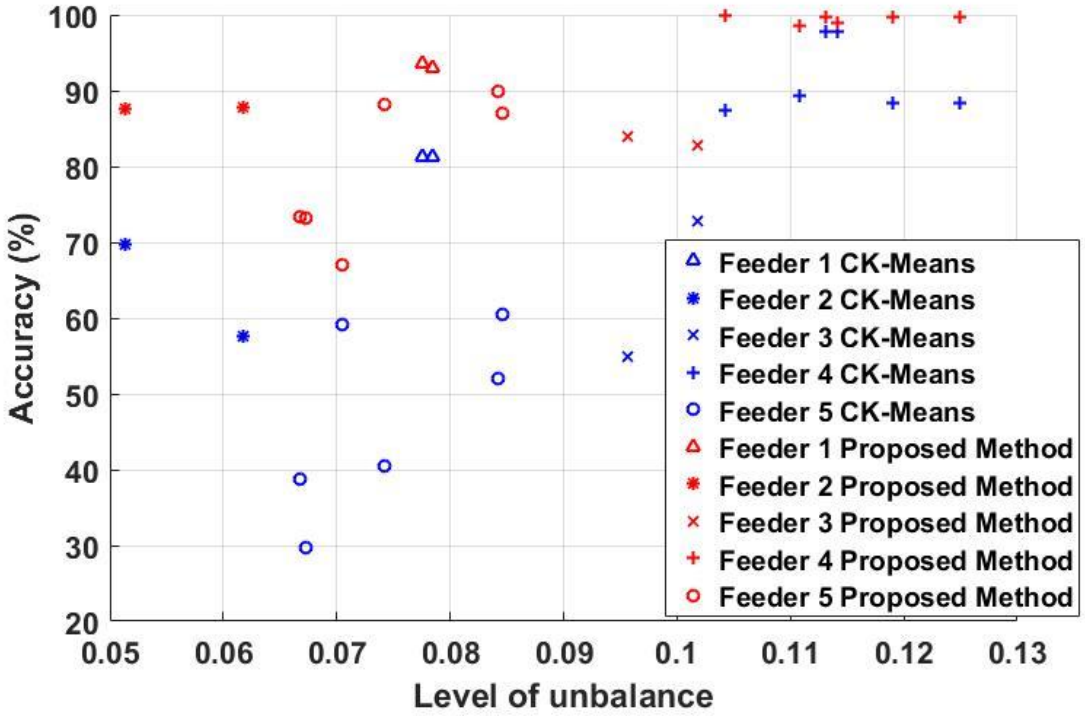


The data points are not well separated according to phase connection with linear dimension reduction.



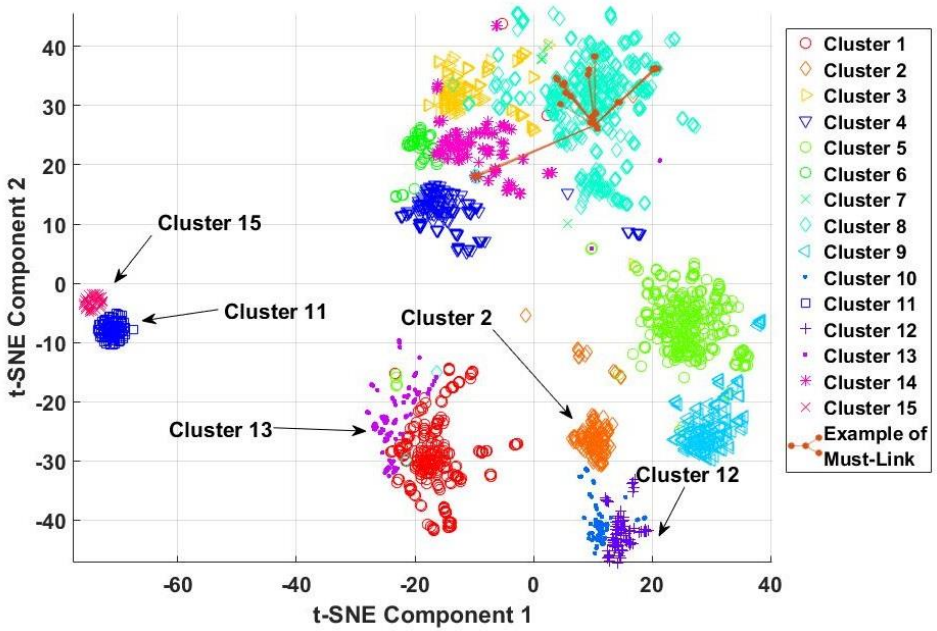
The non-linear dimensionality reduction technique does a much better job in extracting hidden features from the voltage time series during a less unbalanced period for the feeders.

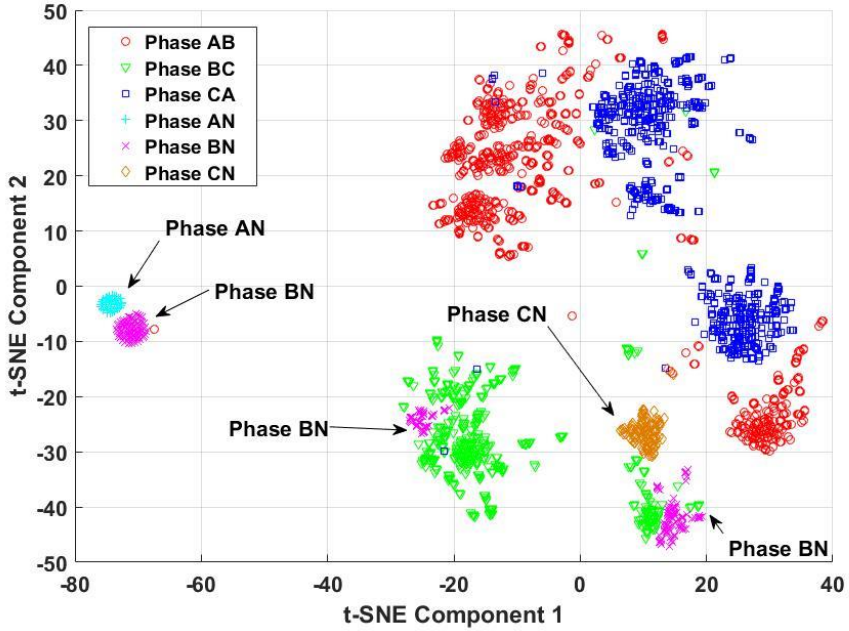
Phase Identification Accuracy with CK-Means and the Proposed Method



- The proposed phase identification algorithm significantly outperforms the CK-Means method with all data sets in terms of accuracy.
- On average, the proposed phase identification algorithm improves the identification accuracy by 19.81% over the CK-Means algorithm.

Clustering Results of the Proposed Method





- Nonconvex clusters are identified.
- The proposed phase identification algorithm not only groups phase-to-phase meters for phase AB, BC, and CA accurately, but also groups single-phase meters with high accuracy

Impact of Data Granularity on Accuracy

Feeder	Data Set	Granularity of Meter Readings		
		1 hour	15-minute	5-minute
1	s1	93.06%	93.93%	93.88%
	s2	93.62%	94.32%	94.40%
2	s3	87.55%	88.86%	92.03%
	s4	87.79%	90.47%	89.93%
3	s5	83.94%	90.02%	91.56%
	s6	82.83%	84.51%	87.16%

- As the granularity of meter readings increases from hourly to every 15 minutes and then 5 minutes, the phase identification accuracy increases.
- The average increase in phase identification accuracy over the 3 distribution circuits is 3.36% when the meter reading granularity increases from hourly to 5 minutes.
- More granular voltage readings allows extraction of features/patterns that may not be present in coarse data sets

Outline

- Why do we focus on electric power distribution systems?
- Big data in power distribution systems
  - Volume, Variety, Velocity, and Value
- Machine learning and big data applications in distribution systems
  - Topology Identification: Phase Connectivity Identification
  - Unsupervised Machine Learning
    - Linear Dimension Reduction & Centroid-based Clustering
    - Nonlinear Dimension Reduction & Density-based Clustering
    - Physically Inspired Maximum Marginal Likelihood Estimation

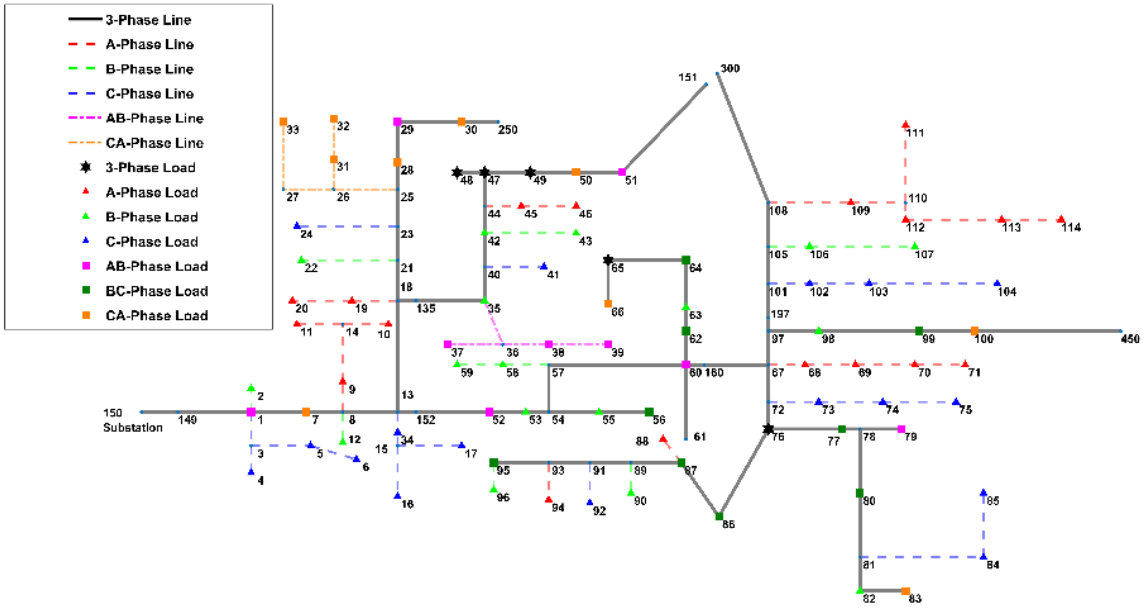
Motivation and Main Idea[3]

- Motivation
  - Existing data-driven approaches lack physical interpretation and theoretical guarantee.
  - Their performance generally deteriorates as the complexity of the network, the number of phase connections, and the level of load balanceness increase.
  - Need a physically inspired data-driven algorithm for phase identification.
- Overall Framework
  - Develop a physical model, which links the phase connections to the voltage magnitudes and power injections via the three-phase power flow manifold.
  - Formulate the phase identification problem as a maximum likelihood estimation (MLE) and a maximum marginal likelihood estimation (MMLE) problem.
  - Prove that the correct phase connection solution achieves the highest log likelihood values for both problems.
  - Develop an efficient solution algorithm for the MMLE problem.

3. Wang, Wenyu, and Nanpeng Yu. "Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems." *arXiv preprint arXiv:1902.09686* (2019).

Problem Setup

- A distribution circuit contains M loads (can connect to the three-phase primary line directly or indirectly through single-phase or two-phase branches).
- The three-phase primary line consists of $N + 1$ nodes.
- Node 0 is the substation/source node.
- Smart meters measure the real and reactive power consumption and the voltage magnitude of each load.



Available Information and Goal

- Available Information
  - For a single-phase load on phase i, we know its power injection and voltage magnitude of phase i
  - For a two-phase delta-connected load between phase i and j, we know its power injection and voltage magnitude across phase i and j.
  - For a three-phase load, we know its total power injection and the voltage magnitude of one of the phases, which needs to be identified.
  - For the source node, we know the voltage measurement (SCADA).
  - The connectivity model of the primary feeder. (GIS)
- Goal
  - Identify which phase(s) each single-phase or two-phase load connects to and which phase's voltage magnitude the three-phase smart meter measures.

Linearized Three-phase Power Flow Model

$$A\begin{bmatrix}\boldsymbol{v}-\overline{\boldsymbol{v}}\\ \boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{v}-\overline{\boldsymbol{v}}\\ \boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{p}\\ \boldsymbol{q}\end{bmatrix}$$

- A_{ij} are $3(N+1)\times 3(N+1)$ matrices derived from the admittance matrix*.
- $\boldsymbol{v}$, $\boldsymbol{\theta}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ are the nodes' voltage magnitude, voltage angle, real and reactive power of the three phases.
- $\bar{v}=\mathbf{1}_{3(N+1)}$ and $\bar{\theta}=[0\times\mathbf{1}_{N+1}^{T},-\frac{2\pi}{3}\times\mathbf{1}_{N+1}^{T},\frac{2\pi}{3}\times\mathbf{1}_{N+1}^{T}]^{T}$ are the flat feasible solution for the underlying nonlinear three-phase power flow.
- Remove the rows & columns of the substation node in A_{mn}, $\boldsymbol{v}$, $\boldsymbol{\theta}$, $\boldsymbol{p}$, and $\boldsymbol{q}$:

$$\check{A}\begin{bmatrix}\check{\boldsymbol{v}}\\ \check{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\widetilde{A_{11}} & \widetilde{A_{12}}\\ \widetilde{A_{21}} & \widetilde{A_{22}}\end{bmatrix}\begin{bmatrix}\check{\boldsymbol{v}}\\ \check{\boldsymbol{\theta}}\end{bmatrix}=\begin{bmatrix}\check{\boldsymbol{p}}\\ \check{\boldsymbol{q}}\end{bmatrix}$$

- Now voltage can be written in terms of real and reactive power injections.

$$\check{\boldsymbol{v}}=(\widetilde{A_{11}}-\widetilde{A_{12}}\widetilde{A_{22}}^{-1}\widetilde{A_{21}})^{-1}\check{\boldsymbol{p}}-(\widetilde{A_{11}}-\widetilde{A_{12}}\widetilde{A_{22}}^{-1}\widetilde{A_{21}})^{-1}\widetilde{A_{12}}\widetilde{A_{22}}^{-1}\check{\boldsymbol{q}}$$

Or in condensed form as

$$\check{\boldsymbol{v}}=K\check{\boldsymbol{p}}-L\check{\boldsymbol{q}}$$

Similarly we have

$$\check{\boldsymbol{\theta}}=\varkappa\check{\boldsymbol{p}}-\mathcal{L}\check{\boldsymbol{q}}$$

*$rank(A)$ is at most $6N$. Need to transform A into a nonsingular form to make the subsequent derivations easier.

Modeling Phase Connections in Three-phase Power Flow

- Decision Variables for Phase Connection
  - x_m^1, x_m^2, and x_m^3 denote phase connections for each load m.
  - $x_m^i = 0$ or 1, and $\sum_i \ x_m^i = 1$, $\forall m$.
  - If load m is single-phase, then x_m^1, x_m^2, and x_m^3 represent AN, BN, and CN connections.
  - If load m is two-phase, then x_m^1, x_m^2, and x_m^3 represent AB, BC, and CA connections.
  - If load m is three-phase, then the measured voltage is between one phase and the neutral, then x_m^1, x_m^2, and x_m^3 represent which of the phases AN, BN, and CN is measured.
  - The phase connection decision variables form a $M \times 3M$ matrix X as

$$X \triangleq diag([x_1^1 \ x_1^2 \ x_1^3], \dots, [x_M^1 \ x_M^2 \ x_M^3])$$

Link Phase Connections to Smart Meter Measurements

- Main Result

$$\hat{\boldsymbol{v}} \approx X\hat{\boldsymbol{v}}^{ref} + X\hat{K}X^T\hat{\boldsymbol{p}} + X\hat{L}X^T\hat{\boldsymbol{q}}$$

  - $\hat{\boldsymbol{v}}$, $\hat{\boldsymbol{p}}$, and $\hat{\boldsymbol{q}}$ denote measured voltage magnitudes, real power and reactive power of each load*. $\hat{\boldsymbol{v}}^{ref} \triangleq [\hat{v}_1^{ref}, \ldots, \hat{v}_M^{ref}]$. $\hat{v}_m^{ref} = \left[v_0^a, v_0^b, v_0^c\right]$ if load m is single-phase or three-phase. $\hat{v}_m^{ref} = \left[v_0^{ab}, v_0^{bc}, v_0^{ca}\right]$ if load m is two-phase.
  - $\hat{K} \triangleq [(U^1K + U^2\boldsymbol{\varkappa})\hat{U}^1 - (U^1L + U^2\mathcal{L})\hat{U}^3]$
  - $\hat{L} \triangleq [(U^1K + U^2\boldsymbol{\varkappa})\hat{U}^2 - (U^1L + U^2\mathcal{L})\hat{U}^1]$
  - U^1, U^2, $\hat{U}^1$, $\hat{U}^2$, and $\hat{U}^3$ are $3N \times 3M$ matrices calculated based on the topology of the three-phase primary feeder.
- The time difference version of the physical model

$$\tilde{\boldsymbol{v}}(t) = X\tilde{\boldsymbol{v}}^{ref}(t) + X\hat{K}X^T\tilde{\boldsymbol{p}}(t) + X\hat{L}X^T\tilde{\boldsymbol{q}}(t) + \boldsymbol{n}(t)$$

  - $\tilde{\boldsymbol{v}}(t) \triangleq \hat{\boldsymbol{v}}(t) - \hat{\boldsymbol{v}}(t-1)$. $\tilde{\boldsymbol{v}}^{ref}(t)$, $\tilde{\boldsymbol{p}}(t)$, and $\tilde{\boldsymbol{q}}(t)$ are defined in a similar way.
  - $\boldsymbol{n}(t)$ is the "noise term" representing the error of the linearized power flow model, the measurement error, and other sources of noise not considered.

* The derivation of measured voltage magnitudes, real power and reactive power from the corresponding variables can be found in the arXiv version of the paper.

Formulate Phase Identification as a Maximum Likelihood Estimation (MLE) Problem

- MLE Problem Formulation
  - Let $\boldsymbol{x} \triangleq [x_1^1, x_1^2, x_1^3, \dots, x_M^1, x_M^2, x_M^3]^T$ be the phase connection decision variable vector.
  - Define $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$ as the theoretical difference voltage measurement $\tilde{\boldsymbol{v}}(t)$ with phase connection $\boldsymbol{x}$.
  - Assume that the noise follows a Gaussian distribution $\boldsymbol{n}(t) \sim \mathcal{N}(\boldsymbol{0}_{M\times 1}, \Sigma_N)$, where Σ_N is an unknown underlying covariance matrix.
  - Assume that $\boldsymbol{n}(t)$ is i.i.d. and independent of $\tilde{\boldsymbol{v}}^{ref}(t), \tilde{\boldsymbol{p}}(t), and\ \tilde{\boldsymbol{q}}(t)$. Given these conditions, $\boldsymbol{n}(t)$ is also independent of $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$.
  - The likelihood of observing $\{\tilde{\boldsymbol{v}}(t)\}_{t=1}^T$, given $\boldsymbol{x}$, $\{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T$ and $\{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T$ is

$$Prob(\{\tilde{\boldsymbol{v}}(t)\}_{t=1}^T \mid \{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T; \boldsymbol{x})$$
$$= \frac{|\Sigma_N|^{-\frac{T}{2}}}{(2\pi)^{\frac{MT}{2}}} \times \exp\{-\frac{1}{2}\sum_{t=1}^{T}[\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_N^{-1} [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]\}$$

  - Taking the negative logarithm of likelihood function, removing the constant, and scaling by $^2/_T$, we get

$$f(\boldsymbol{x}) \triangleq \frac{1}{T}\sum_{t=1}^{T}[\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]^T \Sigma_N^{-1} [\tilde{\boldsymbol{v}}(t) - \tilde{\boldsymbol{v}}(t, \boldsymbol{x})]$$

Theoretical Guarantee

- The correct phase connection $\boldsymbol{x}^*$ maximizes the likelihood function and minimizes the function $f(\boldsymbol{x})$ under two mild assumptions.

  Lemma 1. Let $\boldsymbol{x}^*$ be the correct phase connection. If the following two conditions are satisfied, then as $T \to \infty$, $\boldsymbol{x}^*$ is a global optimizer of $f(\boldsymbol{x})$.

  1. $\boldsymbol{n}(t_k)$ is i.i.d. and independent of $\widetilde{\boldsymbol{v}}^{ref}(t_l), \widetilde{\boldsymbol{p}}(t_l), and\ \widetilde{\boldsymbol{q}}(t_l)$, for $\forall t_k, t_l \in Z^+$.
  2. $\widetilde{\boldsymbol{v}}^{ref}(t_k), \widetilde{\boldsymbol{p}}(t_k), and\ \widetilde{\boldsymbol{q}}(t_k)$ are independent of $\widetilde{\boldsymbol{v}}^{ref}(t_l), \widetilde{\boldsymbol{p}}(t_l), and\ \widetilde{\boldsymbol{q}}(t_l)$, for $\forall t_k, t_l \in Z^+$, $t_k \neq t_l$.

- Directly minimizing $f(\boldsymbol{x})$ is very difficult due to its nonlinearity and nonconvexity. Furthermore, the actual values of Σ_N is unknown.
- Therefore, we will convert the phase identification problem into a maximum marginal likelihood estimation (MMLE) problem.
- We will also prove that the correct phase connection is a also a global optimizer of the MMLE problem.

Phase Identification as a Maximum Marginal Likelihood Estimation (MMLE) Problem

- Let $\tilde{v}_m(t)$ be the mth entry of $\tilde{\boldsymbol{v}}(t)$, $\tilde{v}_m(t, \boldsymbol{x})$ be the mth entry of $\tilde{\boldsymbol{v}}(t, \boldsymbol{x})$, and $n_m\ (t)$ be the mth entry of $\boldsymbol{n}(t)$.
- The marginal likelihood of observing $\{\tilde{v}_m(t)\}_{t=1}^T$, given $\boldsymbol{x}$, $\{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T$ and $\{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T$ is

$$Prob(\ \{\tilde{v}_m(t)\}_{t=1}^T \mid \{\tilde{\boldsymbol{p}}(t)\}_{t=1}^T, \{\tilde{\boldsymbol{q}}(t)\}_{t=1}^T; \boldsymbol{x})$$

$$= \frac{|\Sigma_N(m,m)|^{-\frac{T}{2}}}{(2\pi)^{\frac{T}{2}}} \times \exp\{-\frac{1}{2}\sum_{t=1}^{T} \frac{[\tilde{v}_m(t) - \tilde{v}_m(t, \boldsymbol{x})]^2}{\Sigma_N(m,m)}\}$$

  - Where $\Sigma_N(m,m)$ is the mth diagonal entry of Σ_N. Taking the negative logarithm of the likelihood function, removing the constant terms and scaling by $2\Sigma_N(m,m)/T$, we have

$$f_m(\boldsymbol{x}) \triangleq \frac{1}{T}\sum_{t=1}^{T} [\tilde{v}_m(t) - \tilde{v}_m(t, \boldsymbol{x})]^2$$

- **Lemma 2**. Let $\boldsymbol{x}^*$ be the correct phase connection. If the following two conditions in Lemma 1 hold, then as $T \to \infty$, $\boldsymbol{x}^*$ is a global optimizer of $f_m(\boldsymbol{x})$.

Solution Method

- Directly minimizing $f_m(\boldsymbol{x})$ is still a difficult task.
- We further simplify the optimization problem by first solving three sub-problems $minf_{m,i}(\boldsymbol{x}_{-m}), i\epsilon\{1,2,3\}$.

$$f_{m,i}(\boldsymbol{x}_{-m}) \triangleq f_m(\boldsymbol{x})$$

Subject to $x_m^i = 1$ and $x_m^j = 0$ for $j \neq i$

Where $\boldsymbol{x}_{-m}$ is a $(3M-3)\times 1$ vector containing every element in $\boldsymbol{x}$ except x_m^1, x_m^2, and x_m^3. Then we have

$$minf_m(\boldsymbol{x}) = \min\{minf_{m,1}(\boldsymbol{x}_{-m}),\ minf_{m,2}(\boldsymbol{x}_{-m}), f_{m,3}(\boldsymbol{x}_{-m})\}$$

- Now the sub-problem for MMLE can be formulated as

Find $\quad \boldsymbol{x}_{-m,i}^{\dagger} = \underset{\boldsymbol{x}_{-m}}{\operatorname{argmin}} f_{m,i}(\boldsymbol{x}_{-m})$

Subject to $\ x_k^j = 0$ or $1\ \forall j$ and $k \neq m$

$$\textstyle\sum_j x_k^j = 1\ \forall\ k \neq m$$

- This is a binary least-square problem which can be converted to convex quadratic programming by relaxing the problem by replacing the binary constraints by their convex hull. The sub-problem can be solved in polynomial time.

Summary of Solution Algorithm

Algorithm 1 Phase Identification Algorithm

Input: $\tilde{\boldsymbol{v}}(t)$, $\tilde{\boldsymbol{v}}^{\text{ref}}(t)$, $\tilde{\boldsymbol{p}}(t)$, $\tilde{\boldsymbol{q}}(t)$, $\hat{K}$, and $\hat{L}$, $t = 1, ..., T$.
Output: Estimated phase connections for the M loads.

1: **for** $m = 1$ to M **do**
2: **for** $i = 1$ to 3 **do**
3: Use the input to calculate $v_{m,i}^{\text{tot}}(t)$ and $\boldsymbol{\varphi}_{m,i}^{T}(t)$ and find the solution $\boldsymbol{x}_{-m,i}^{\dagger}$ to the sub-problem in (45).
4: **end for**
5: Use $\boldsymbol{x}_{-m,i}^{\dagger}$, $i \in \{1, 2, 3\}$ to find the $\boldsymbol{x}$ that minimizes $f_m(\boldsymbol{x})$ in (34). Record the solution as $\boldsymbol{x}^{\dagger}{}_m$.
6: **end for**
7: Generate two phase identification results based on M sets of $\boldsymbol{x}^{\dagger}{}_m$ using two approaches: the target-only approach and the voting approach.
8: Calculate $\sum_{m=1}^{M} f_m(\boldsymbol{x})$ based on both the target-only and the voting approach. Select the solution with the lower sum of square error.

- From step 1 to 6, we solve M MMLE problems, each of which contains three binary least-square sub problems.
- Step 3 solves the sub-problems of MMLE.
- Step 5 solves the mth MMLE problem by finding which of the three $\boldsymbol{x}_{-m,i}^{\dagger}$ minimizes $f_{m,i}(\boldsymbol{x}_{-m})$.
- The chosen $\boldsymbol{x}_{-m,i}^{\dagger}$ combined with the corresponding $x_m^i = 1$, $x_m^j = 0$ $(j \neq i)$ forms the $3M \times 1$ solution $x_m^{\dagger}$ of the mth MMLE problem.
- The M sets of $x_m^{\dagger}$ may not be all correct due to the limited number of measurements, and measurement noise.
- In step 7, we design two approaches to integrate M sets of $x_m^{\dagger}$ into two phase identification solutions. The final solution has a lower sum of square error.

1. Target-only Approach. The phase connection of each load m is the corresponding connection shown in the mth solution $x_m^{\dagger}$.
2. Voting Approach. For single-phase and two-phase load m, the phase connection is the corresponding phase connection that receives the most votes in the M sets of $x_m^{\dagger}$.

Numerical Study Setup

- Test Circuits (Modified IEEE distribution feeders)
  - Radial primary: IEEE 37-bus, 123-bus & heavily meshed primary: 342-bus.

Feeder	A	B	C	AB	BC	CA	ABC	Total
37-bus	5	5	6	3	2	2	2	25
123-bus	18	17	17	9	9	10	5	85
342-bus	30	38	31	35	31	33	10	208

Number of Loads per Phase in the IEEE Test Circuits

- Smart Meter Data
  - Length: 90 days of hourly average real power consumption data (2160 data points)
  - Source: a distribution feeder managed by FortisBC.
- Power Flow Simulated with OpenDSS to Generate Theoretical Nodal Voltage
- Measurement noise follows a zero-mean Gaussian distribution with three-sigma deviation matching 0.1% and 0.2% of nominal values. (0.1 and 0.2 accuracy class smart meters established in ANSI.)
- After applying measurement noise, the voltage measurements are rounded to the nearest 1 V for the primary loads and 0.1 V for the secondary loads to make the phase identification task more difficult.

Phase Identification Algorithm Performance

Accuracy of the Proposed Phase Identification Method

Feeder	Meter Class	30 days	60 days	90 days
37-bus (radial)	0.1%	100%	100%	100%
	0.2%	92%	100%	100%
123-bus (radial)	0.1%	96.47%	100%	100%
	0.2%	63.53%	96.47%	100%
342-bus (meshed)	0.1%	96.63%	100%	100%
	0.2%	72.60%	99.52%	100%

- The performance of the proposed MMLE-based algorithm on three IEEE distribution test circuits, two meter accuracy classes, and three time windows are shown here.
- With 90 days of hourly meter measurements, the proposed algorithm achieved 100% accuracy for all three circuits. (Works well for radial and meshed circuits).
- Phase identification accuracy increases as smart meter measurement error decreases and addition smart meter data becomes available.

Comparison with Existing Methods

Phase Identification Accuracy of Different Methods with 90 days of Meter Data

Method	Meter Class	37-Bus Feeder	123-Bus Feeder	342-Bus Feeder
Correlation-based Approach*	0.1%	100%	98.75%	81.82%
	0.2%	100%	97.5%	81.31%
Clustering-based Approach#	0.1%	100%	100%	93.43%
	0.2%	100%	98.75%	91.41%
MMLE-based Algorithm	0.1%	100%	100%	100%
	0.2%	100%	100%	100%

- The proposed MMLE-based algorithm outperforms the correlation and clustering-based approaches.
- The improvement in accuracy increases as the complexity of the distribution feeder increases.

* M. Xu, R. Li, and F. Li, “Phase identification with incomplete data,” *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2777-2785, 2018

W. Wang and N. Yu, "AMI Data Driven Phase Identification in Smart Grid," the Second International Conference on Green Communications, Computing and Technologies, pp. 1-8, Rome, Italy, Sep. 2017.

Conclusion

- Develop a physically inspired data-driven algorithm for the phase identification in power distribution system.
- The phase identification problem is formulated as an MLE and MMLE problem based on the three-phase power flow manifold.
- We prove that the correct phase connection is a global optimizer for both the MLE and the MMLE problems.
- A computationally efficient algorithm is developed to solve the MMLE problem, which involves synthesizing the solutions from the sub-problems.
- Comprehensive simulation results show that our proposed algorithm yields high accuracy and outperforms existing methods.

The Center for Grid Engineering Education - Short Course: Big Data Analytics and Machine Learning in Smart Grid

- Date: May 9th 8:00 am – 5:00 pm
- Location: Hilton St. Louis at The Ballpark
  1 South Broadway, Gateway Ballroom
  St. Louis, Missouri
- PDH's available: 8 hours
- Registration Fee charged by EPRI
  - $800 per person
  - 20% discount for organizations with three or more attendees
  - 25% discount for government employees (non-utility)
  - 25% discount for university professors*
  - 75% discount for graduate students*
  - *University IDs required to qualify for professor or graduate student discounts.

https://intra.ece.ucr.edu/~nyu/Teaching/ML-BD-Smart-Grid_2019.pdf

The Center for Grid Engineering Education - Short Course: Big Data Analytics and Machine Learning in Smart Grid

- EPRI Contacts: Amy Feser, afeser@epri.com
- (865) 218-5909
- Course Topics: Big Data Analytics and Machine Learning
  - Distribution System
    - Topology Identification
    - Theft Detection
    - Predictive Maintenance of Distribution Equipment
    - Estimation of Behind-the-meter Solar Generation
    - Reinforcement Learning based Volt-VAR Control and Network Reconfiguration
  - Electricity Market
    - Algorithmic Trading with Virtual Bids in Electricity Market
  - Transmission System
    - Anomaly Detection with PMU Data
    - Motifs and Signatures Discovery with PMU Data
    - Segmentation of PMU Data

https://intra.ece.ucr.edu/~nyu/Teaching/ML-BD-Smart-Grid_2019.pdf

Thank You

- Contact information
  - Dr. Nanpeng Yu
  - Department of Electrical and Computer Engineering, UC Riverside, United States
  - Phone: 951.827.3688
  - Email: nyu@ece.ucr.edu
  - Website: http://www.ece.ucr.edu/~nyu/