This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

Topology Identification of Distribution Networks Using A Split-EM based Data-Driven Approach

Li Ma, *Member*, *IEEE*, Lingfeng Wang, *Senior Member*, *IEEE*, and Zhaoxi Liu, *Member*, *IEEE*

Abstract—To improve the topology observability in power distribution networks (PDNs), a two-stage topology identification framework is proposed to recognize the mixed topologies in a large set of historical data and predict the real-time topology based on the nodal measurements. A split expectation-maximization (split-EM) method is proposed considering the measurement errors to deal with the topology identification problem on the historical batch data, in which the number of topology categories does not need to be given in advance. Based on the topology identification results of historical data, the number of topology categories is reduced. Then, feasible classifiers are trained using machine learning methods to predict the real-time topology efficiently. An error-correcting mechanism is proposed for the real-time identification involving the credibility analysis and the reidentification based on the Bayesian recursion model. Finally, via a practical example, the effectiveness of the proposed models is verified by efficiently identifying the PDN's topologies in both the historical batch data with mixed topologies and real-time measurements. In addition, the partition-based extension application solution of the topology identification models for large-scale PDNs is proposed without extra measurements to relieve the calculation burden and reduce the identification time notably while maintaining the accuracy as the non-partitioned scheme.

***Index Terms*— Historical topology identification, real-time topology identification, resilience improvement, power distribution system, split-EM method, classifier training.**

I. Introduction

The power distribution network (PDN) is different from the transmission grids where the topologies are regularly measured and verified. Topology information of distribution networks is inaccurate or even unavailable due to uninformed changes that happen from time to time, such as network reconfiguration, repairs, maintenance and load balancing [1]. Although topology sensors (such as Feeder Terminal Unit, FTU) are being utilized in PDNs, they are placed only at special locations due to budget constraints [2]. Topology identification in PDNs is critical due to its important roles in carrying out many tasks, including power flow analysis, real-time contingency analysis, resiliency enhancements against natural disasters or cyber-physical attacks, efficient integration of renewable energy sources (RES), and so on [3-6].

The importance of topology identification to power grids is receiving growing attention, and some related research has been performed in the past few years based on the ongoing deployment of advanced metering infrastructures (AMI), micro-phasor measurement unit (μPMU)-type sensors [7] and GPS timing devices [1] at the buses. The existing research on power grid topology identification problems can be classified into static models [1, 2, 8-10] and time-varying models [3, 11-16], depending on whether the grid topology is fixed or not during the computation. For the static models, reference [1] proposed an error-in-variables model to jointly estimate the line parameter and topology in a maximum-likelihood framework. Reference [2] proposed a mutual information-based topology identification model for the distribution grid with new data from sensor-equipped DER devices. Reference [8] proposed a structure learning algorithm to solve the topology estimation problem in structurally meshed but operationally radial distribution networks. Reference [9] proposed a topology identification algorithm based on the measurements from a few line current sensors, and the problem was modeled as a mixed integer linear program (MILP). Reference [10] used the Markov Random Field algorithm to explore the nodal correlation, and a revised maximum likelihood method was devised to solve the model. Then for the time-varying models, the research is composed of real-time identification with single topology [3, 11-14, 16], and historical identification with multiple topologies [15]. For real-time topology identification, the sparse-recovery methods were utilized in [3] and [11] to solve the smart grid topology identification problems, where the power network was regarded as an interconnected graph and the DC power flow model was considered. A "Learning-to-Infer" method was developed in [12] for identifying the line status of the power network efficiently in real time, and the line outage detector optimization was solved as a discriminative learning model. Reference [13] proposed a model for identifying network changes based on the Bayesian approach, and the model was tested on the 11 kV distribution networks of the U.K. Generic Distribution System (UKGDS). Reference [14] studied the parameters and topology estimation problem in a polyphase distribution network via the least absolute shrinkage and selection operator (LASSO) regression. Reference [16] proposed a distribution network dynamic topology awareness method that only requires the synchronized voltage amplitude measurements of a few nodes in the grid. In [15], the authors improved their previous static state framework for the single topology identification in [1], and a parameter and topology joint identification model was proposed. In addition to the topology identification, references [17, 18] also addressed the phase identification problem using a maximum marginal likelihood estimation and an MILP model, respectively.

In general, the distribution network is structurally meshed but operationally radial. To overcome the drawbacks of the radial systems, prevent service interruptions, and reduce losses, the meshed configuration has been studied [19, 20]. In recent years, the large-scale integration of RES has also brought great changes to the distribution systems, and the conceptions such as microgrid and energy hub have also been implemented [21, 22].

This work was supported in part by the U.S. National Science Foundation under Award ECCS1711617. *(Corresponding author: Lingfeng Wang.)*
The authors are with the Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA (e-mail: optimali@qq.com; l.f.wang@ieee.org; zhaoxil@uwm.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

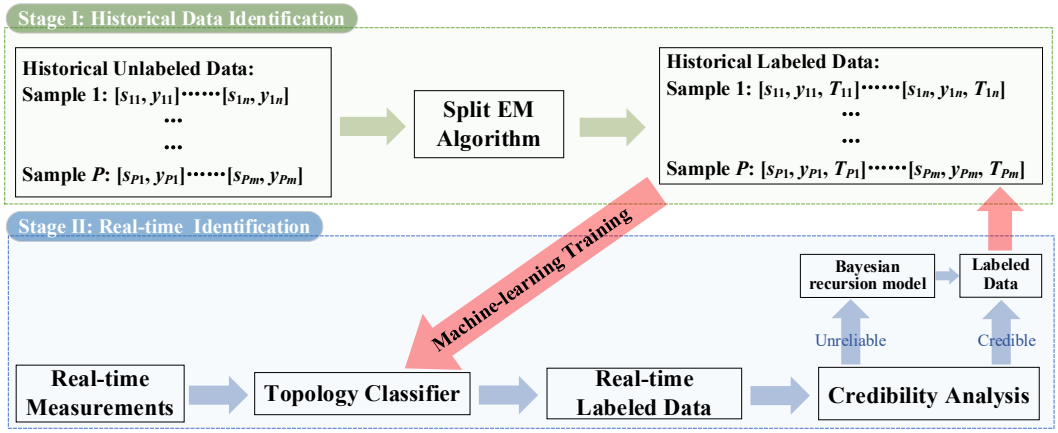

Fig. 1. The two-stage topology identification procedure.

Accordingly, the partially islanding operation mode will also be adopted in the distribution networks under some circumstances. Considering these changes, the topology identification tools for distribution networks should be applicable to diversified network topologies. Several models in the existing literature have considered the meshed configuration [1, 14], but the networks with islands have seldom been considered. In addition, most of the existing studies assumed that the collected data samples correspond to the same topology [23-26]. However, the network topology changes can happen rather frequently in the distribution system operation, as much as once every several hours [27]. These topology changes can be induced by the routine reconfiguration, manual maintenance, etc. [2], and it is not easy to determine when and how the topology changes. Thus, the number of topology categories for a set of historical measurement data is usually unknown in practice, and identifying the mixed topologies in a large set of historical data simultaneously is a challenging task. Moreover, as the historical and real-time topology identifications were realized separately in the existing literature, the historical topology information has not been fully exploited in the real-time topology prediction, and in some cases the labeled data has to be generated artificially. To date, an integrated framework involving both historical and real-time topology identification has not been researched. In this context, this paper proposes a two-stage topology identification framework for PDNs to recognize the mixed topologies in historical batch data and predict the real-time topology based on the available nodal measurements. The measurements include nodal active/reactive power injections, voltage amplitudes and phase angles, where the voltage phase angles are measured by low-cost μPMUs [28], and other measurement are obtained from widely-used smart meters [29]. The main contributions of the paper are listed as follows:

1) A modified EM algorithm named split-EM is proposed for historical batch data topology identification, where the number of topology categories in the historical data is not necessary. For a data sample including multiple records, the proposed model can simultaneously find all the topology categories and identify the topology each record belongs to. It is applicable to different types of topologies, including radial, meshed, and islanded networks; and different system models, such as three-phase balanced systems and unbalanced systems. Further, it is also applicable when the nodal voltage phase angles are not measured.

2) A two-stage topology identification framework is proposed based on the split-EM historical identification, in which the number of the topology categories could be narrowed down. Then the classifiers are trained using machine learning methods and adopted in the real-time topology prediction more efficiently.

3) An extension application solution of the topology identification models for large-scale networks is designed without any extra measurements. By partitioning the network into subsystems, the calculation burden is reduced. The overall topology information can be obtained dynamically and efficiently with the proposed solution.

This rest of the paper is organized as follows. The overall two-stage topology identification framework is introduced in section II. The historical and real-time topology identification models are proposed in section III. In section IV, the application solution of the proposed models for large-scale networks is designed. The case study is presented in section V. Lastly, the conclusions are drawn in section VI.

II. Two-Stage Topology Identification Framework

The two-stage topology identification framework is shown in Fig. 1. In stage I, for historical data identification, the unlabeled records can be divided into several smaller samples (each sample includes a number of records, and one record refers to the nodal measurements at a point in time), and the topology identification operations can be performed with the split-EM method in a parallel way to improve the efficiency. For each record $[s_{ij}, y_{ij}]$, it is the nodal measurements of all nodes in the PDN at a point in time, and s_{ij} includes the voltage amplitudes and phase angles, while vector y_{ij} includes the active and reactive power injections. The topology category T_{ij} is determined after the topology identification procedure is performed, which is added to the original record to generate the labeled data $[s_{ij}, y_{ij}, T_{ij}]$.

The machine-learning training will be carried out to generate several topology classifiers with all the historical labeled data. The inputs of the classifiers are the nodal measurements and the output is the corresponding topology categories. The trained classifiers will be used in stage II for real-time identification. The real-time measurement (similar to one record in the history data) can be labeled using the trained classifiers, and the credibility analysis is performed to prevent the rare occasions that the classifiers cannot correctly label some measurements under new topology parameters (which

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

may never appear in the historical data). Based on the result of the credibility analysis, the credible labeled data will be sent to the historical labeled database to update the machine-learning training of the classifier, while the unreliable data will be reidentified using the Bayesian recursion model, which will take a longer time than the classifier.

III. Historical and Real-time Topology Identification

A. Historical Topology Identification

1) The EM model

When the topology information is unknown, the topology identification problem of the historical data can be regarded as an unsupervised classification problem. The unsupervised classification of historical data can further be considered as a parameter estimation problem with unknown mixture of topology categories. The historical data identification model should integrate topology estimation and category selection in one algorithm, and also be applicable when there is a great variety of topology categories in a large-scale network. The inputs of the model are the nodal historical measurements including the active and reactive power injections, voltage amplitudes and phase angles in the PDN, and the output is the corresponding topology categories for all the measurement records. The parameters in the historical topology identification problem can be represented as $\boldsymbol{\Theta} = \{(T_m, \alpha_m), m \in [1,2,\dots,M]\}$, where M is the total number of the topology categories, and α_m is the proportion of the records with the m-th topology category in the sample, T_m is the topology parameter vector of the m-th category, representing the states of the lines with unknown connectivity. T_m's dimension is the number of lines with unknown states. The element of T_m is binary which equals 1 if the corresponding line is connected, and 0 otherwise. The logarithmic likelihood function of sample $\boldsymbol{X}$ (including N records $[x_1,\dots,x_N]$) under $\boldsymbol{\Theta}$ (i.e., $\mathbf{L}(\boldsymbol{X};\boldsymbol{\Theta})$) can be expressed as:

$$\begin{aligned}\mathbf{L}(\boldsymbol{X};\boldsymbol{\Theta}) &= \log\textstyle\prod_{i=1}^{N} \mathrm{p}(x_i;\boldsymbol{\Theta}) \\ &= \textstyle\sum_{i=1}^{N} \log\big(\mathrm{p}(x_i;\boldsymbol{\Theta})\big) \\ &= \textstyle\sum_{i=1}^{N} \log \sum_{j=1}^{M} \mathrm{p}\big(x_i,T_j;\boldsymbol{\Theta}\big) \\ &= \textstyle\sum_{i=1}^{N} \log \sum_{j=1}^{M} Q_{x_i}^{T_j} \frac{\mathrm{p}(x_i,T_j;\boldsymbol{\Theta})}{Q_{x_i}^{T_j}} \end{aligned} \quad (1)$$

$$\mathbf{L}^*(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q}) = \textstyle\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{x_i}^{T_j} \log \frac{\mathrm{p}(x_i,T_j;\boldsymbol{\Theta})}{Q_{x_i}^{T_j}} \quad (2)$$

$$\mathbf{L}^*(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q}) \le \mathbf{L}(\boldsymbol{X};\boldsymbol{\Theta}) \quad (3)$$

where $\mathbf{L}^*(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q})$ can be regarded as the lower bound of the logarithmic likelihood function, and the inequality in expression (3) is derived from the Jensen-inequality [30]. It has been proved that $\mathbf{L}(\boldsymbol{X};\boldsymbol{\Theta}) = \mathbf{L}^*(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q})$ when

$$Q_{x_i}^{T_j} = \frac{\alpha_j \mathrm{p}(x_i|T_j;\boldsymbol{\Theta})}{\sum_{j=1}^{M} \alpha_j \mathrm{p}(x_i|T_j;\boldsymbol{\Theta})} \quad (4)$$

$\mathrm{p}(x_i;\boldsymbol{\Theta})$ is the probability of x_i within parameter $\boldsymbol{\Theta}$, $\mathrm{p}(x_i,T_j;\boldsymbol{\Theta})$ is the probability of x_i belonging to T_j within parameter $\boldsymbol{\Theta}$, and $\mathrm{p}(x_i|T_j;\boldsymbol{\Theta})$ is the probability of x_i given T_j within parameter $\boldsymbol{\Theta}$. $Q_{x_i}^{T_j}$ is also known as the conditional distribution of the j-th topology for the i-th record in the sample:

$$\textstyle\sum_{j=1}^{M} Q_{x_i}^{T_j} = 1 \quad (5)$$

$$\mathrm{p}(x_i,T_j;\boldsymbol{\Theta}) = \alpha_j \mathrm{p}(x_i|T_j;\boldsymbol{\Theta}) \quad (6)$$

The parameters of mixture models can be estimated using the expectation-maximization (EM) algorithm [31], which is an iterative method to find the maximum likelihood of parameter estimates in the statistical models depending on the unobserved latent variables (which are the states of the lines with unknown connectivity here). The EM iteration alternates between the expectation step (E-step) and maximization step (M-step). The E-step creates a function for the expectation of the log-likelihood evaluated using the current parameter estimation results, and the M-step maximizes the expected log-likelihood function determined in the E-step. The E-step and M-step for estimating the parameters $\Theta = \{(\alpha_m, T_m), m \in [1,M]\}$ are as follows:

- The E-step calculates $\boldsymbol{Q}(t+1)$ (with element $Q_{x_i}^{T_j}(t+1)$ in the matrix) based on the current estimate of the parameters $\boldsymbol{\Theta}(t)$. The conditional distribution $Q_{x_i}^{T_j}(t+1)$ can also be regarded as the posterior probability of the i-th record belonging to the j-th topology category.

$$Q_{x_i}^{T_j}(t+1) = \frac{\alpha_j(t)\mathrm{p}\big(x_i|T_j(t);\boldsymbol{\Theta}(t)\big)}{\sum_{m=1}^{M}\alpha_m(t)\mathrm{p}\big(x_i|T_m(t);\boldsymbol{\Theta}(t)\big)} \quad (7)$$

- The M-step maximizes $\mathbf{L}^*(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q}(t+1))$ to update the estimate of the parameter $\boldsymbol{\Theta}(\boldsymbol{t+1})$:

$$\alpha_m(t+1) = \textstyle\sum_{i=1}^{N} Q_{x_i}^{T_m}(t+1)/\boldsymbol{N} \quad (8)$$

$$\begin{aligned}\boldsymbol{\Theta}(\boldsymbol{t+1}) &= \operatorname{argmax} \mathbf{L}^*\big(\boldsymbol{X},\boldsymbol{\Theta};\boldsymbol{Q}(t+1)\big) \\ &= \operatorname{argmax}\Big\{\textstyle\sum_{i=1}^{N}\sum_{j=1}^{M} Q_{x_i}^{T_j}(t+1)\log\frac{\mathrm{p}(x_i,T_j;\boldsymbol{\Theta})}{Q_{x_i}^{T_j}(t+1)}\Big\}\end{aligned} \quad (9)$$

The variables in the optimization problem (9) of the M-step are $\{(T_m)\}, m \in [1,2,\dots,M]$, and the evaluation of $\mathrm{p}(x_i,T_j;\boldsymbol{\Theta})$ is closely related to the power flow calculation, which will be introduced in detail in section III-B. Then the model in (9) is a nonlinear integer programming problem, which is also a non-convex problem.

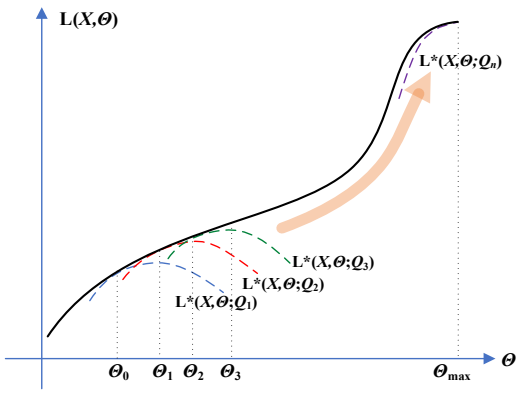

Fig. 2. Schematic diagram of the EM algorithm in topology identification.

The convergence analysis for the EM algorithm can be referred to [32]. A common issue associated with the EM algorithm is the local optimum problem [33]. In other words, if the logarithmic likelihood function has multiple peaks, the EM process is easy to fall into a local optimal solution. In topology identification studies, it seems unlikely that two topologies have the same logarithmic likelihood values for certain records, which means multiple peaks are almost impossible in the topology identification problem (this statement will also be verified in the case study section). Then, it is assumed that $\mathbf{L}(\boldsymbol{X};\boldsymbol{\Theta})$ will not get the same value between all the possible

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

topology parameters $\boldsymbol{\Theta}$ in this paper. Regarding $\boldsymbol{\Theta}$ as the variable, the logarithmic likelihood function can be sorted in an ascending order, and the realization of the EM process can be depicted in Fig. 2.

As shown in Fig. 2, the EM algorithm starts from a randomly given parameter vector $\boldsymbol{\Theta_0}$, then Q_1 can be determined using expression (7). Based on Q_1, we can get function $\mathbf{L}^*(\boldsymbol{X}, \boldsymbol{\Theta}; Q_1)$, and $\boldsymbol{\Theta_1}$ can be solved by optimizing $\mathbf{L}^*(\boldsymbol{X}, \boldsymbol{\Theta}; Q_1)$. This process is repeated until $\boldsymbol{\Theta_{max}}$ is found.

2) The split-EM model

The EM-based topology identification model introduced above is suitable for the case with known number of topology categories. As mentioned in section I, the number of topology categories may not be available in practice when a group of historical data is given. To deal with this problem, we propose a modified EM algorithm named split-EM in this paper. The procedure of the proposed split-EM method is shown in Fig. 3.

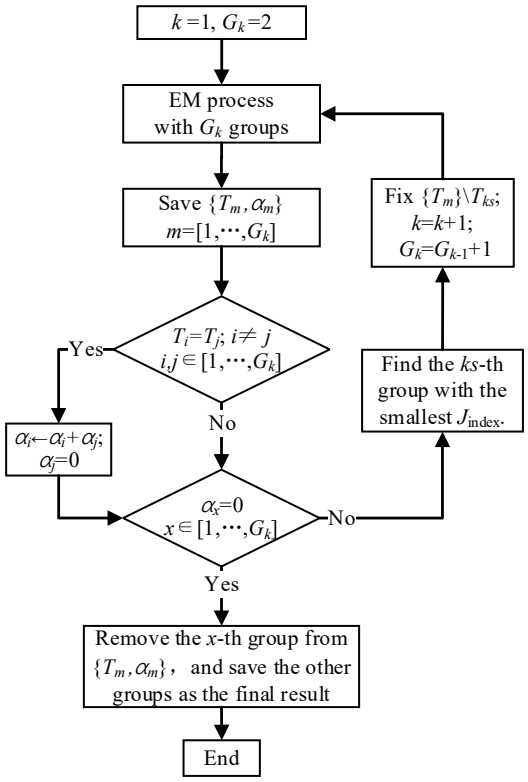

Fig. 3. The realization procedure of the split-EM method.

In the procedure, k is the round number in the split-EM process, and G_k is the number of the topology categories in the k-th round. The whole procedure will start from $k = 1, G_k = 2$. The aforementioned EM process is performed in the beginning of each round, and the estimated parameters $\{(T_m, \alpha_m)\}, m \in [1, \dots, G_k]$ are obtained. Then we judge whether two categories hold the same topology parameters, and if there exist any two categories i and j satisfying $T_i = T_j$, α_i is set to be $\alpha_i + \alpha_j$, and α_j is set to be 0. If no two categories hold the same topology parameters, we proceed to decide if there exists any category x with $\alpha_x = 0$. If there exists a category with $\alpha_x = 0$, the split-EM process is ended, and the final parameter results are $\{(T_m, \alpha_m)\}, m \in [1, \dots, G_k]$ with the x-th category's parameters removed. If there doesn't exist any category x with $\alpha_x = 0$, the process goes to the next round. A specified ks-th category is chosen to split in the next round based on a judgement index as expressed in expression (10), when the ks-th category holds the smallest J_{index} among all the topology categories in the current round. The topologies' parameters except the ks-th category are retained in the new round, which means $\{T_m\}\backslash T_{ks}, m \in [1, \dots, G_k]$ is fixed in the new round of the EM process, while $\{\alpha_m\}, m \in [1, \dots, G_{k+1}]$ still need to be estimated. The split-EM is continued until the termination condition is satisfied. The judgement index of the j-th group is expressed as follows to reflect the credibility of T_j.

$$\boldsymbol{J_{\text{index\_}j}} = \frac{\sum_{i=1}^{N} Q_{x_i}^{T_j} \log(\mathrm{p}(x_i|T_j;\boldsymbol{\Theta}))}{\sum_{i=1}^{N} Q_{x_i}^{T_j}} \tag{10}$$

Remark: Comparison between the split-EM and EM algorithms

The computational complexity of the EM algorithm is $O(NM)$ for every iteration [34], where N is the number of the records and M is the number of the topology categories. For the split-EM process, 2 topology categories need to be identified in each round as other topology categories are determined according to the results of the previous round. Then the computational complexity for one iteration of all the rounds can be expressed as $O(2NM)$ (the number of rounds is equal to the total number of topology categories M). In the traditional EM algorithm, if we try from 2 topology categories to (M+1) topology categories, the computational complexity for one iteration of all the rounds can be expressed as $O(2N+3N+\cdots+(M+1)N) = O(NM(M+3)/2)$, which is greater than or equal to $O(2NM)$ in the split-EM method when $M \geqslant 2$. The gap between the split-EM and traditional EM will be more obvious when M is larger.

B. Probability Density Calculation

This section mainly focuses on how to determine the probability of a data record belonging to a specified topology category, which is an important element in the split-EM process.

1) Three-phase balanced system

For the three-phase balanced power system, we usually use one single phase to represent the overall system. Then the AC power flow equations with the line states are expressed as follows [1]:

$$p_i = \sum_{j=1}^{m_e} g_j |l_{ji}| \left(v_i^2 - v_{u_{j1}} v_{u_{j2}} \cos\left(l_{ji} \left(\theta_{u_{j1}} - \theta_{u_{j2}} \right) \right) \right) - b_j |l_{ji}| v_{u_{j1}} v_{u_{j2}} \sin\left(l_{ji} \left(\theta_{u_{j1}} - \theta_{u_{j2}} \right) \right) \tag{11}$$

$$q_i = \sum_{j=1}^{m_e} b_j |l_{ji}| \left(v_{u_{j1}} v_{u_{j2}} \cos\left(l_{ji} \left(\theta_{u_{j1}} - \theta_{u_{j2}} \right) \right) - v_i^2 \right) - g_j |l_{ji}| v_{u_{j1}} v_{u_{j2}} \sin(l_{ji}(\theta_{u_{j1}} - \theta_{u_{j2}})) \tag{12}$$

where p_i and q_i are the active and reactive power injections at the i-th node, g_i and b_i are the conductance and susceptance on the j-th line, v_i and θ_i are the voltage amplitude and phase angle at the i-th node, m_e is the total number of lines. l_{ji} is the element in the incidence matrix $\mathbf{L}$ of PDN, $l_{ji} \in \{1, -1, 0\}$ represents the j-th line leaves from, enters, or separates from the i-th node, respectively. u_{j1}, u_{j2} are the elements in the incidence matrix $\boldsymbol{u}$ of PDN,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

where u_{j1} and u_{j2} represent "from" and "to" node numbers of the *j*-th line. The power flow equations for all the nodes can also be expressed as the following:

$$[\boldsymbol{p}, \boldsymbol{q}] = \boldsymbol{h}(\boldsymbol{v}, \boldsymbol{\theta}) \tag{13}$$

where the power flow function $\boldsymbol{h}$ corresponds to a specified topology.

In expression (2), $\mathrm{p}(x_i|T_j;\boldsymbol{\Theta})$ is a multi-dimensional density model corresponding to the *j*-th topology category. T_j represents the topology parameters in the *j*-th category (the line parameters such as length and impedance are known, only the connection states will be considered here). We use $\mathrm{p}(x_i|T_j)$ to replace $\mathrm{p}(x_i|T_j;\boldsymbol{\Theta})$ for simplicity in the following section. $\mathrm{p}(x_i|T_j)$ is the same as $\mathrm{p}([s_i, y_i]|T_j)$, and can also be expressed as $\mathrm{p}([s_i, y_i, s_i', y_i']|T_j)$, where $[s_i, y_i]$ denotes one measurement record, and $[s_i', y_i']$ are the corresponding real values. $\mathrm{p}([s_i, y_i, s_i', y_i']|T_j)$ can be calculated based on the specified error distributions' probability density function [1, 13], when errors are obtained using $[s_i, y_i]$ and $[s_i', y_i']$. However, it is not easy to obtain $[s_i', y_i']$ accurately in practice. Here, a probability density calculation method only relying on nodal measurements will be used. It is supposed that the measurement errors ε_s and ε_y follow the Gauss distributions as: $\varepsilon_s \sim \mathrm{N}(0, {\sigma_s}^2)$, $\varepsilon_y \sim \mathrm{N}(0, {\sigma_y}^2)$. Then the real values and T_j satisfy the following expressions:

$$y_i' = h_j(s_i') \tag{14}$$

$$s_i' = s_i - \varepsilon_s,\ y_i' = y_i - \varepsilon_y \tag{15}$$

Using the first order Taylor expansion [35], we have:

$$y_i - \varepsilon_y = h_j(s_i - \varepsilon_s) \approx h_j(s_i) - \varepsilon_s h_j'(s_i) \tag{16}$$

$$E_j = y_i - h_j(s_i) \approx -\varepsilon_s h_j'(s_i) + \varepsilon_y \tag{17}$$

According to the above expressions, $y_i - h_j(s_i) \sim N(0, \Sigma_j)$, where $\Sigma_j = \left(h_j'(S)\sigma_s\right)^2 + {\sigma_y}^2$. Then $\mathrm{p}([s_i, y_i, s_i', y_i']|T_j)$ can be replaced by $\mathrm{p}(y_i - h_j(s_i)|T_j)$, and the latter only contains measurement values and can be represented as:

$$\begin{aligned}\mathrm{p}(x_i|T_j) &= \mathrm{p}(E_j|T_j) = \mathrm{p}(y_i - h_j(s_i)|T_j) \\ &= \frac{1}{\sqrt{(2\pi)^n \cdot |\Sigma_j|}} \exp\left(-\frac{1}{2} {E_j}^{\mathrm{T}} \left(\Sigma_j\right)^{-1} E_j\right)\end{aligned} \tag{18}$$

where $n = 2 \times n_e$ (n_e is the total number of nodes), and the standard deviation σ in the covariance matrix Σ_j can be determined based on the relative error of the measurement (error%). For the measured value with a given mean μ, $\mu \pm 3 \cdot \sigma$ can cover more than 99.7% area of the Gaussian curve. For any measured value Φ, σ can be calculated as follows [13]:

$$\sigma = \frac{\Phi \times \mathrm{error}\%}{3 \times 100} \tag{19}$$

2) The Model for Three-phase Unbalanced Systems

For the three-phase unbalanced power systems, the AC power flow equations in [36] are reformulated with the line states as follows:

$$\begin{aligned}p_i^\alpha = \sum_{j=1}^{n_e} \sum_{\beta=\mathrm{a,b,c}} B_{ij} g_{ij}^{\alpha\beta} \Big(&v_i^\alpha v_i^\beta \cos\left(\theta_i^\alpha - \theta_i^\beta\right) - \\ &v_i^\alpha v_j^\beta \cos\left(\theta_i^\alpha - \theta_j^\beta\right)\Big) - B_{ij} b_{ij}^{\alpha\beta} v_i^\alpha v_j^\beta \sin\left(\theta_i^\alpha - \theta_j^\beta\right)\end{aligned} \tag{20}$$

$$\begin{aligned}q_i^\alpha = \sum_{j=1}^{n_e} \sum_{\beta=\mathrm{a,b,c}} B_{ij} b_{ij}^{\alpha\beta} \Big(&v_i^\alpha v_j^\beta \cos\left(\theta_i^\alpha - \theta_j^\beta\right) - \\ &v_i^\alpha v_i^\beta \cos\left(\theta_i^\alpha - \theta_i^\beta\right)\Big) - B_{ij} g_{ij}^{\alpha\beta} v_i^\alpha v_j^\beta \sin\left(\theta_i^\alpha - \theta_j^\beta\right)\end{aligned} \tag{21}$$

where B_{ij} denotes the state of the line from the *i*-th node to the *j*-th node, and

$$\begin{cases} B_{ij} = 1, & \text{connected} \\ B_{ij} = 0, & \text{disconnected} \end{cases}; \tag{22}$$

α and β are phase indexes; p_i^α and q_i^α are the active and reactive power injections of phase α at the *i*-th node; v_i^α and θ_i^α are the voltage amplitude and phase angle of phase α at the *i*-th node; $g_{ij}^{\alpha\beta}$ and $b_{ij}^{\alpha\beta}$ are the conductance and susceptance between phase α and β on the line from the *i*-th node to the *j*-th node. Similar to the model for the single-phase case, the power flow equations for all the nodes in the network can also be expressed as follows:

$$[\boldsymbol{p}^u, \boldsymbol{q}^u] = \boldsymbol{h}^{\mathrm{u}}(\boldsymbol{v}^u, \boldsymbol{\theta}^u) \tag{23}$$

where $\boldsymbol{p}^u = \{p_i^\alpha\}$, $\alpha = \mathrm{a,b,c}$, $i \in [1, \ldots, n_e]$; the definitions for $\boldsymbol{q}^u$, $\boldsymbol{v}^u$, and $\boldsymbol{\theta}^u$ are similar to that of $\boldsymbol{p}^u$; the power flow function $\boldsymbol{h}^{\mathrm{u}}$ corresponds to a specified topology in a three-phase unbalanced system. $\mathrm{p}^u(x_i|T_j;\boldsymbol{\Theta})$ is used to denote the probability of x_i given T_j within parameter $\boldsymbol{\Theta}$ in a three-phase unbalanced system, and the calculation process can be realized as that for $\mathrm{p}(x_i|T_j;\boldsymbol{\Theta})$ in balanced systems, i.e., expressions (14)~(18).

C. Real-Time Topology Identification

Based on the split-EM identification model, a large amount of historical data can be labeled. Then the classifiers for the real-time topology identification can be generated based on the labeled historical data using a series of machine learning methods. For each real-time measurement, its topology category will be labeled using the trained classifiers. The longer the collection time of the historical data is, the better the performance of trained classifiers will be, since more complete topology scenarios can be used in the training process.

1) Credibility analysis of the labels

The identification result of the classifier should be further analyzed as the classifier is not applicable to the records with topologies that did not exist in the historical data samples and a classifier may also have error by itself. The credibility analysis can be performed based on the logarithm of the probability (expression (18)) to determine if the label is credible, because the logarithm probability of a record belonging to the correct topology category is generally much larger than that belonging to a wrong one (more details will be presented in the case study section). Based on the result of the credibility analysis, the unreliable labeled data will be reidentified using the Bayesian recursion model.

2) The Bayesian recursion based reidentification for unreliable labels

The Bayesian recursion model can be utilized to reidentify the real-time measurement with unreliable label [13]. All the possible topology categories should be given in advance in the Bayesian recursion model. For a specified record, the iteration process can be expressed as follows:

$$\mathrm{p}\left(T_j|\boldsymbol{\varepsilon}\right)^k = \frac{\mathrm{p}\left(E_j|T_j\right)\mathrm{p}\left(T_j|\boldsymbol{\varepsilon}\right)^{k-1}}{\sum_{m=1}^{M} \mathrm{p}(E_m|T_m)\mathrm{p}\left(T_m|\boldsymbol{\varepsilon}\right)^{k-1}} \tag{24}$$

where k is the iteration number, $\boldsymbol{\varepsilon} = [E_1, E_2, \ldots, E_M]$ is the error vector and can be calculated using expression (17), and the original probability $\mathrm{p}\left(T_j|\boldsymbol{\varepsilon}\right)^1, j \in [1,2, \ldots, M]$ are all

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

preset to 1/*M*. After enough iterations are performed, the estimation will converge to one topology category, then this category will be chosen as the final topology category.

The labels of the real-time measurements passed the credibility analysis and error correction will be added to the historical labeled data, which will be utilized to update the classifiers training at a specific frequency to make the classifiers more accurate.

IV. Extension Application in Large-scale Network

A. The Overall Application Process

As mentioned in section III-A, the calculation complexity of the split-EM method is closely related to the number of topology categories involved in the historical data. Then, for a large-scale network, an extension application solution can be employed through partitioning the network into several subsystems to reduce the calculation burden. The flowchart of the overall process is shown in Fig. 4.

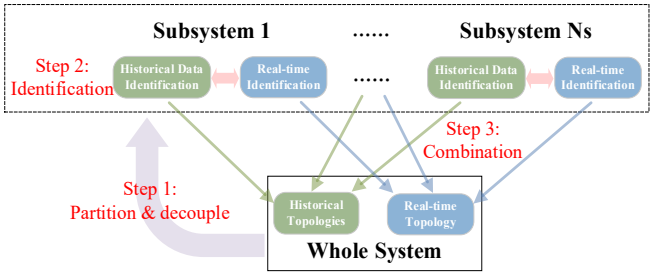

Fig. 4. The overall process of the application in large-scale system.

Step1: The whole PDN is divided into *N*s subsystems through the optimization model for subsystem formation (presented in subsection IV-B). The subsystem formulation can be decoupled from each other by exchanging the nodal voltage amplitudes and phase angles at the nodes of the connecting lines between subsystems, as the power flow equations for each subsystem can be expressed in the following:

$$\begin{aligned}[\boldsymbol{p_1}, \boldsymbol{q_1}] &= \boldsymbol{h_1}(\boldsymbol{v_1}, \boldsymbol{\theta_1}, \boldsymbol{v_{c\_1}}, \boldsymbol{\theta_{c\_1}}) \\ &\vdots \\ [\boldsymbol{p_i}, \boldsymbol{q_i}] &= \boldsymbol{h_i}(\boldsymbol{v_i}, \boldsymbol{\theta_i}, \boldsymbol{v_{c\_i}}, \boldsymbol{\theta_{c\_i}}) \\ &\vdots \\ [\boldsymbol{p_{N_s}}, \boldsymbol{q_{N_s}}] &= \boldsymbol{h_{N_s}}(\boldsymbol{v_{N_s}}, \boldsymbol{\theta_{N_s}}, \boldsymbol{v_{c\_N_s}}, \boldsymbol{\theta_{c\_N_s}})\end{aligned} \tag{25}$$

where $\boldsymbol{h_i}$ is the power flow function for the *i*-th subsystem; $\boldsymbol{p_i}, \boldsymbol{q_i}, \boldsymbol{v_i}, \boldsymbol{\theta_i}$ are the active power vector, reactive power vector, voltage amplitude vector, and voltage phase angle vector at the nodes within the *i*-th subsystem, respectively; $\boldsymbol{v_{c\_i}}, \boldsymbol{\theta_{c\_i}}$ are the voltage amplitude vector, and voltage phase angle vector at the nodes on the other side of the connecting lines between the *i*-th subsystem and other subsystems.

Step2: The topology identification process, including data processing, performing split-EM algorithm for historical data, training the classifers, and predicting the real-time topologies, is performed in each subsystem. The identification process is similar to that in Fig. 1.

Step3: The historical and realtime topology identification results in each subsystem are transmitted to the distribution system operator (DSO) directly and combined together based on the time stamps. For the real-time identification, only the subsystem undergoing the topology change needs to update its real-time toplogy information, making the overall performance of the framework more efficient.

The advantage of this application solution in terms of the calculation complexity will be analyzed here. For a data sample consisting of N records, assuming that the number of the topology categories within the *i*-th subsystem is M_i, then the largest number of the topology categories of the whole distribution network is $\prod_{i=1}^{N_s} M_i$. If the topology identification is performed in the whole distribution network, the calculation complexity in each iteration can be up to $O(2N\prod_{i=1}^{N_s} M_i)$. While the total calculation complexity in each iteration can be expressed as $O(2N\sum_{i=1}^{N_s} M_i)$ if the topology identification is performed in each subsystem. Based on the topology identification results of the subsystems and the acquisition time of the measurement, the topology information for the whole distribution network can be obtained by combining the results of the subsystems together. In most cases, $O(2N\sum_{i=1}^{N_s} M_i)$ is much less than $O(2N\prod_{i=1}^{N_s} M_i)$, which means that the partitioning scheme adopted here can drastically reduce the calculation burden.

B. The Optimization of Subsystem Formation

For a simple system, we may evenly divide it into several subsystems artificially considering the estimated identification time in each subsystem for step1 in Fig. 4. However, it may not be easy to divide a complex system with a mass of nodes and loops. In this case, the formation of the subsystems are important for the overall application process. Too many subsystems will lead to the increase of information interaction between subsystems, and over large subsystems caused by uneven partition or too few subsystems will limit the reduction in the identification time for the whole area. Therefore, an optimization of the subsystem formation is nessesary to further improve the performance of the whole application process in Fig. 4. The optimization for the subsystem formation is modeled as follows:

$$\begin{aligned}\min \quad & \sum_{i=1}^{m_e} C_i \\ \text{s.t.} \quad & N_j + N_{c\_j} \le N_{\text{limit}} \\ & N_{s\_j} \le N_{s\_\text{limit}}\end{aligned} \tag{26}$$

where the objective in the optimization is to minimize the total number of connecting lines (also reflecting minimizing the number of subsystems), and C_i is the connecting line indicator. $C_i = 1$ means the *i*-the line is a connecting line, $C_i = 0$ means not. The interconnected subsystems can be formed supposing that all connecting lines are removed, and two constraints are considered in each subsystem: the number of nodes in the *j*-th subsystem ($N_j + N_{c\_j}$) is less than N_{limit}, and the number of unknown states in the *j*-th subsystem (N_{s_j}) is less than N_{s_limit}. To be noticed, the number of nodes on the other sides of the connecting lines (N_{c_j}) for the *j*-th subsystem is also considered for the calculation of the *j*-th subsystem, and the lines with unknown states are not used as the connecting lines here. N_{limit} and N_{s_limit} can be determined according to the empirical topology identification time with different system scales and unknown state numbers (more details can be found in Section V). The optimization model can be solved based on graph related algorithms and intelligent optimization algorithms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

TABLE I THE RESULTS OF SPLIT-EM PROCESS FOR IEEE 33-BUS SYSTEM

Round	Scenario 1: with all nodal voltage phase angles measured				Scenario 2: With all nodal voltage phase angles unmeasured and set as 0			
	T^*	α	J_{index}	Logarithmic likelihood	T^*	α	J_{index}	Logarithmic likelihood
k=1	[1110110000; 1111100000]	[0.325; 0.675]	[**-529.8**; -215.9]	974.8	[1110110000; 1111100000]	[0.325; 0.675]	[**-315.7**; -155.1]	$-2.14*10^4$
k=2	[1110110000; 1111100000; 0111100100]	[0.30; 0.65; 0.05]	[40.7; **13.9**; 39.5]	2242.2	[1110110000; 1111100000; 0111100100]	[0.30; 0.65; 0.05]	[**−264.1**; −171.2; −41.7]	$-1.93*10^4$
k=3	[1110110000; 1111100000; 0111100100; 1011101000]	[0.30; 0.50; 0.05; 0.15]	[40.7; **35.2**; 39.5; 40.0]	3665.0	[1110110000; 1111100000; 0111100100; 1011101000]	[0.30; 0.50; 0.05; 0.15]	[**−264.1**; −152.2; −41.7; −146.7]	$-1.81*10^4$
k=4	[1110110000; 1111100000; 0111100100; 1011101000; 1110111011]	[0.30; 0.50; 0.05; 0.15; **0**]	——	3665.0	[1110110000; 1111100000; 0111100100; 1011101000; 1111100000]	[0.30; 0.25; 0.05; 0.15; 0.25]	——	$-1.81*10^4$

V. CASE STUDIES

A. Historical Topology Identification in IEEE 33-bus System

In the actual operation of PDNs, the basic topology information can be obtained through the GIS (Geographic Information System), and the states of unmonitored switches need to be identified. The verification of the topology identification mainly focuses on the identification of the states of unmonitored switches in this section. The proposed topology identification model is applied in the IEEE 33-bus system as shown in Fig. 5. In this paper, it is assumed the following lines' connection states are unknown (i.e., the lines with unmonitored switches): line 11-12, line 14-15, line 15-16, line 2-19, line 28-29, line 8-21, line 9-15, line 12-22, line 18-33, line 25-29 (the numbers are the end-nodes' serial numbers of the lines), and the lines with unknown connectivity are shown as dotted lines and numbered from 1 to 10 as shown in Fig. 5. There are 2^{10}=1024 possible topologies in this test system.

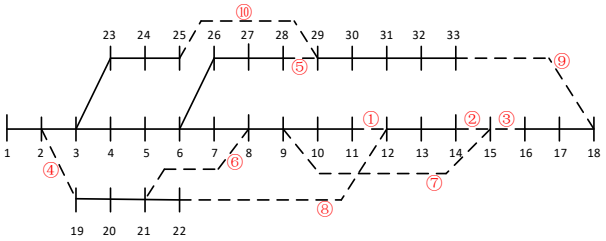

Fig. 5. The structure of IEEE 33-bus test system.

The voltage of bus 1 in the PDN is 12.66 kV. Except for the first bus, the nodal active/reactive power injections in the PDN are randomly generated within predefined ranges, and the average values for the active and reactive power are 120 kW and 80 kVar respectively. The nodal voltage amplitudes and phase angles (except for the first bus) and the first bus's power injections in the PDN are generated using the Matpower toolbox in MATLAB [37, 38]. Measurement errors are randomly generated and added to the above generated data. According to the American National Standard Institute (ANSI) C12.20 Standard [39], the standard deviations of errors (in terms of p, q, v, θ) are all set to be 0.1% of the measurements (we use 0.1% in the text below for simplicity), and the error of each measurement is generated based on its error standard deviation and limited by three times of the standard deviation (i.e., $\varepsilon \le \pm 3 \cdot \sigma$). In this case, error% is 0.3%, which satisfies the accuracy standard in [39]. The error settings are also applicable to other parts in the case study except for the sensitivity analysis of the measurement errors. The MATLAB software package is employed to solve the split-EM model and generate classifiers based on the labeled historical data.

1) The application in different topology types

- The application in radial networks

Here the topology identification is performed for a data sample including 100 records. There are 4 topology categories (all of them are radial networks) in this data sample, i.e., T1, T2, T3, T4, corresponding to records 1~50, 51~80, 81~95, and 96~100, respectively. The ralated information of the sample are listed in Table II.

TABLE II TESTING DATA FOR THE CASE CONSISTING OF RADIAL NETWORKS

Topology category	The sequential states of the switches*	Record numbers	Record proportion
T1	[1111100000]	1~50	50%
T2	[1110110000]	51~80	30%
T3	[1011101000]	81~95	15%
T4	[0111100100]	96~100	5%

*** For the state of the switch, "1" means the switch is closed, while "0" indicates it is open.**

The OPTI toolbox and NOMAD solver [40, 41] are used to solve the non-convex optimization in the split-EM solving process. The initial states of switches are all set to 1, and this setting is also used in other cases in the case study section. The results of the split-EM process for the above mentioned case are shown in Table I. The scenarios with and without nodal voltage phase angles are both verified (denoted as Scenario 1 and Scenario 2, respectively). As shown in Table I, the topology identification in both scenarios experiences 4 rounds in the split-EM processes, and the split-EM processes can efficiently identify all the topologies in the historical records, even for category T4 with only 5 records. In each round, the topology category with smallest J_{index} is chosen to be split in the next round, while the other topology categories remain the same. As there is a zero element in $\boldsymbol{\alpha}$ within Scenario 1 and two repeated topology categories in T* within Scenario 2 when k=4, the split-EM processes stop according to the realization procedure in Fig. 3. The logarithmic likelihood values increase with the split-EM processes in both scenarios, as more correct

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

topologies are recognized. Although the logarithmic likelihood values in Scenario 2 are much smaller than those in Scenario 1, the topology categories and which category each record belongs to can also be identified correctly.

- The application in the network with meshes

In this section, a data sample including the meshed topology is verified. Each record in the data sample is generated using the same process as described previously, but different topologies are used when calculating power flows using the Matpower toolbox in MATLAB. The data sample generation in the network with islands in the following section is also similar. There are 2 topology categories in this data sample, i.e., T1 (1111100000) and T2 (1111110001, with mesh), corresponding to records 1~50 and 51~80 respectively, and the identification results are listed in Table III. The split-EM process is similar with that in Table I), and the identification results are all correct in this case.

TABLE III THE RESULTS IN SPLIT-EM PROCESS FOR THE CASE WITH MESH

Round	T	α	J_{index}	Logarithmic likelihood
k=1	[1111100000; 1111110001]	[0.625; 0.375]	[34.4; **29.1**]	2538.3
k=2	[1111100000; 0100000010; 1111110001]	[0.625; **0**; 0.375]	——	2538.3

- The application in the network with islands

In this section, a data sample including topology with island is verified. T1 (1111100000) and T2 (0111000010, with island) correspond to records 1~50 and 51~80 respectively, and the identification results are listed in Table IV. The topology identification results are also all correct in this case.

TABLE IV THE RESULTS IN SPLIT-EM PROCESS FOR THE CASE WITH ISLAND

Round	T	α	J_{index}	Logarithmic likelihood
k=1	[0111000010; 1111100000]	[0.375; 0.625]	[33.4; **32.4**]	2566.8
k=2	[0111000010; 0111111100; 1111100000]	[0.375; **0**; 0.625]	——	2566.8

2) Sensitivity analysis of line parameters and measurement errors

In this section, we test 400 records corresponding to 4 topology categories ([1111100000], [1011101000], [0111100100], [1111000001], and 100 records for each topology) to analyze the sensitivities of the line parameters and measurement errors by adjusting the amplitudes of the parameter variations and the standard deviations of the errors. As the accuracy of the split-EM method mainly depends on whether the logarithmic probability of a record within the correct topology is the largest, S_{index} in expression (27) is designed and utilized to reflect the sensitivity of the line parameters and measurement errors.

$$S_{\text{index}} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{N}\sum_{i=1}^{N} S_{ik}\right) \tag{27}$$

$$S_{ik} = \begin{cases} 1, & \mathrm{p}(x_{ik}(T_{\mathrm{c}})|T_{\mathrm{c}}) = \max\limits_{j\in[1,\ldots,M]} \mathrm{p}(x_{ik}(T_{\mathrm{c}})|T_j) \\ 0, & \mathrm{p}(x_{ik}(T_{\mathrm{c}})|T_{\mathrm{c}}) \neq \max\limits_{j\in[1,\ldots,M]} \mathrm{p}(x_{ik}(T_{\mathrm{c}})|T_j) \end{cases} \tag{28}$$

where N is the number of the records, K is the number of parameter error sets or measurement error sets; $x_{ik}(T_{\mathrm{c}})$ is the record corresponding to topology T_{c}, while T_j corresponds to the j-th topology category among all the topology categories.

In this study, 20 cases of error standard deviations and 23 cases of parameter variation ranges are considered, and 100 parameter error sets or measurement error sets are randomly generated for each case of the error standard deviation or parameter variation range. The sensitivity analyzing results are shown in Fig. 6. According to Fig. 6, S_{index} is more sensitive to the standard deviations of the measurement errors compared with the parameter errors of the lines in the PDN. It is also found that no two topology categories have the same probability for a specified record in the calculation process. It means there is only one maximum value of the logarithmic probability among all the topology categories. When the standard deviation of the measurement errors is within 0.1% [39], S_{index} corresponding to each standard deviation case is nearly equal to 1, indicating the results of the split-EM method are authentic with the practical measurement errors.

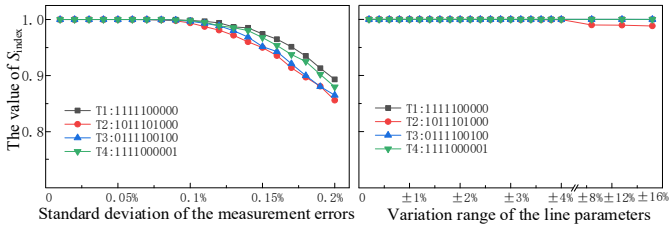

Fig. 6. Sensitivity analysis results of standard deviation and line parameters.

B. Real-time Identification in IEEE 33-bus System

1) The real-time identification based on the classifiers

It is assumed that 24 topology categories are involved in the historical data in this case. 1000 records of each topology category (24000 in total) are used to be trained, and 1000 records are used in the testing process. Among the 1000 testing records, the topology of 20 records is not included in the historical data (the 20 records belong to one topology, [1101100010]). The Classification Learner and Neural Net Pattern Recognition toolboxes in MATLAB are adopted here to generate the trained classifiers. Among all the classification learner models, the Fine Tree, Linear SVM, Quadratic SVM, and Neural Network (using the Neural Net Pattern Recognition toolbox), etc., perform much better than the others in the toolbox, and the training/testing accuracy and prediction time of these models for a single record are listed in Table V.

TABLE V THE COMPARISON BETWEEN CLASSIFIERS IN THE REAL-TIME PREDICTION

Trained classifier	Training accuracy	Testing accuracy	Prediction time (sec)
Fine Tree	99.10%	97.20%	0.03
Linear SVM	99.99%	97.80%	0.26
Quadratic SVM	99.99%	97.90%	0.32
Cubic SVM	99.90%	97.80%	0.35
Medium Gaussian SVM	98.80%	97.50%	0.31
Coarse Gaussian SVM	99.50%	97.20%	0.33
Bagged Trees	98.30%	97.30%	0.10
Subspace Discriminant	99.80%	97.60%	0.09
Neural Network	99.96%	98.00%	0.01

2) Credibility analysis and error corrections

In this section, taking Quadratic SVM as an example, the prediction results will be analyzed and corrected. The logarithmic probability of each record with the topology category identified by the Quadratic SVM is shown in Fig. 7

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

(a). As the topology category for the records from the 381st to the 400th doesn't exist in the historical data, their logarithmic probabilities are much smaller than other records. Thus, we can obtain a threshold value of the logarithmic probability based on a mass of data considering the correct topology and wrong topologies. The threshold value can be used to determine whether an unreliable topology is assigned to a data record. Here we set the threshold as -50, and the records from the 381st to the 400th are regarded as unreliable ones and are reidentified using the Bayesian recursion model, which involves all possible topology categories (the number is 2^{10}=1024). The Bayesian method costs 6.09 sec for each record, and the topology identification results are all correct for these 20 data records, as shown in Fig. 7 (b). After the error correction process, the accuracy has been improved from 97.9% to 99.9%, which is accurate enough for industrial applications.

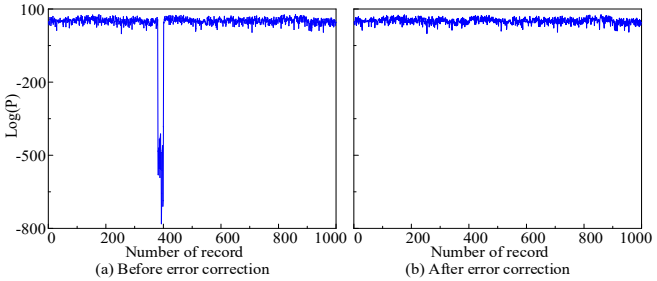

Fig. 7. The logarithmic probabilities of the testing records

3) Sensitivity analysis of the number of topology categories

The sensitivity of the number of topology categories in the training process is analyzed in this section. In Table VI, 100 records for each topology are used in the training process, and 1000 records are used in the testing process.

TABLE VI COMPARISON BETWEEN DIFFERENT NUMBERS OF TOPOLOGIES

Number of topologies in training	Training time (sec)	Training accuracy	Testing accuracy	Prediction time (sec)
20	15.8	99.20%	99.40%	0.11
40	96.8	99.30%	99.00%	0.28
60	292.5	96.00%	97.70%	0.71
80	652.4	93.10%	94.40%	1.25
100	1269.9	93.20%	95.00%	2.02
120	2275.4	93.60%	94.60%	2.91

Note: the results are based on Quadratic SVM.

As shown in Table VI, with the number of the topology categories in training increasing, the training/prediction time tends to increase, while the training/testing accuracy decreases. Then the historical identification process not only provides useful topology information for the DSO to support the optimal operation and planning of the distribution network, but also reduces the number of the topology categories, which could shorten the prediction time and improve the prediction accuracy in the real-time topology identification. It is reasonably assumed that most topology categories have appeared in historical data for the distribution networks that are in operation for a long time. The topologies that were never occurred can be identified through the credibility analysis and error correction process, which is also helpful for the proposed models to be deployed in distribution networks which have been in operation for a short time.

C. *Application in a Three-phase Unbalanced System*

The proposed topology identification model for three-phase unbalanced networks is applied in a pratical distribution system as shown in Fig. 8. The states of the switches/lines directly connected to the transformer stations are monitored in practice and need not to be identifed. The lines with unknown connection states are marked using circled numbers in Fig. 8. As the topology identification focuses on the medium-voltage networks, the measurement points are set at the incoming lines of distribution transformers, which can be regarded as aggregated load points in the PDN model. The nodal currents, voltages and power-factor angles of each phase are measured and used in this test case, and active and reactive power injections are calculated before the topology identification. It can be noticed that the loads are unbalanced between the three phases at each node, and the nodal voltage phase angles of each phase are not necessary in this actual case.

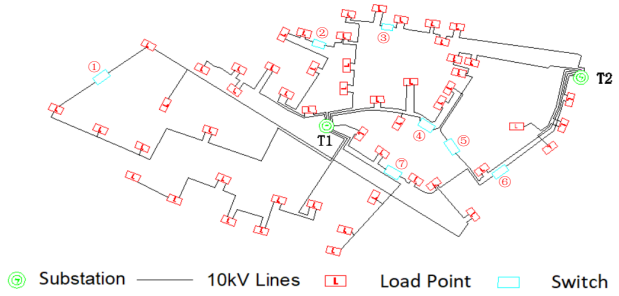

Fig. 8. The structure of an actual three-phase unbalanced system.

1) The historical topology identification

The topology identification is performed on a data sample including 30 records. There are 2 topology categories (both of them are radial networks) in this data sample, i.e., T1 (0000001), T2 (1000000), corresponding to records 1~15 and 16~30, respectively. The identification results are listed in Table VII, and the identification results are all correct.

TABLE VII THE RESULTS IN SPLIT-EM PROCESS FOR THE ACTUAL CASE

Round	T	α	J_{index}	Logarithmic likelihood
k=1	[0000001; 1000000]	[0.5; 0.5]	[-462.3; -475.3]	$-2.07*10^4$
k=2	[0000001; 1000000; 0000001]	[0.5; **0.5**; 0]	——	$-2.07*10^4$

2) The real-time topology identification

TABLE VIII THE COMPARISON BETWEEN CLASSIFIERS IN THE REAL-TIME PREDICTION FOR THREE-PHASE UNBALANCED SYSTEM

Trained classifier	Training accuracy	Testing accuracy	Prediction time (sec)
Fine Tree	97.50%	97.50%	0.02
Boosted Tree	97.90%	96.25%	0.03
Quadratic SVM	99.40%	98.33%	0.04
Cubic SVM	99.20%	99.17%	0.03
Fine KNN	96.80%	97.50%	0.02
Weighted KNN	98.00%	97.50%	0.02

It is assumed that 6 topology categories are involved in the historical data in this case. 200 records of each topology category (1200 in total) are used to be trained, and 240 records are used in the testing process. The Classification Learner toolbox in MATLAB are adopted to generate the trained classifiers. Among all the classification learner models, the models with better performances are listed in Table VIII, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

well as the training/testing accuracy and prediction time (per record) of these models.

D. Application of the Large-scale Network

The proposed topology identification method is also tested on the modified 135-bus test system [42, 43] as shown in Fig. 9. It is assumed that the historical identification is conducted in the end of the day in this case, and the data are collected every 15 minutes, which means there are about one hundred data records for a day. Considering different settings of number of nodes and number of unknown states in the system, Table IX shows time consumption of the empirical historical topology identification with a data sample involving 100 records and two topologies (assuming that the cases with one or two topologies are the most likely scenarios within a data sample of a day). It can be observed that the historical topology identification time is affected by both the number of nodes and the number of unknown states in the system. More nodes and more unknown states will both increase the identification time.

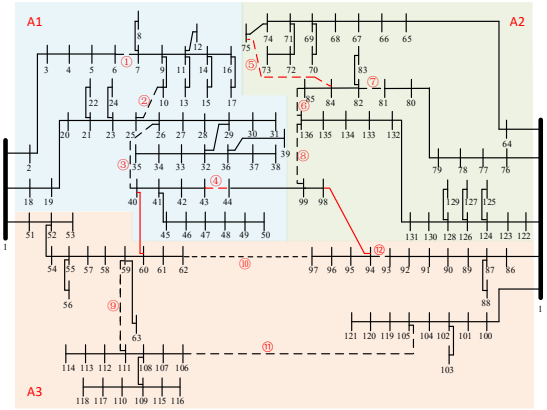

Fig. 9. The 135-bus test system.

TABLE IX TOPOLOGY IDENTIFICATION TIME WITHIN DIFFERENT SYSTEM SCALE AND UNKNOWN STATES (SECONDS)

Number of nodes	Number of unknown states					
	2	4	6	8	10	…
30	3	32	144	213	307	
40	5	80	290	463	472	
50	14	150	824	1619	1764	…
60	20	**179**	1255	1956	1847	
70	34	253	1469	2373	2816	
…	…					

For the test system in Fig. 9, it is assumed that the historical identification is required to be finished within 5 minutes. Therefore, N_{limit} and $N_{\text{s\_limit}}$ are set as 60 and 4 according to Table IX, and the subsystems are highlighted by different background colors in Fig. 9. The proposed topology identification model can be used directly in each subsystem based on the power flow equations in (25). Taking subsystem A1 (highlighted in blue color) as an example, nodes 2~50 are within subsystem A1, and nodes 1, 60, and 99 are the nodes on the other side of the connecting lines of subsystem A1. The nodal active/reactive power injections at nodes 2~50, and the nodal voltage amplitudes/phase angles at nodes [1~50, 60, 99] are measured. In this way, the subsystems are decoupled from each other, and the proposed split-EM process and real-time prediction can be adopted in each subsystem. The split-EM process and real-time prediction are tested for the whole area and the subsystems. The results are presented and compared as follows.

1) The historical topology identification

The testing data for the 135-bus system are shown in Table X. There are 12 lines in the entire system whose connection states are unknown, and each subsystem holds 4 of them. It is assumed that each subsystem has 2 topology categories and the whole area has 5 topology categories within the testing data, which includes 100 records. The states of the switches of each topology category, and the topology category of each record are presented in Table X and Fig. 10.

TABLE X TESTING DATA FOR THE LARGE-SCALE NETWORK

Area range	Serial numbers of the switches	The states of the switches in each topology category
A1	①②③④	T_{11}:[1000]; T_{12}:[0100]
A2	⑤⑥⑦⑧	T_{21}:[0010]; T_{22}:[1000]
A3	⑨⑩⑪⑫	T_{31}:[1001]; T_{32}:[0110]
Whole area	①②③④⑤⑥ ⑦⑧⑨⑩⑪⑫	T_1:[T_{11},T_{21},T_{31}]; T_2:[T_{11},T_{21},T_{32}]; T_3:[T_{11},T_{22},T_{32}]; T_4:[T_{12},T_{22},T_{31}]; T_5:[T_{12},T_{22},T_{32}]

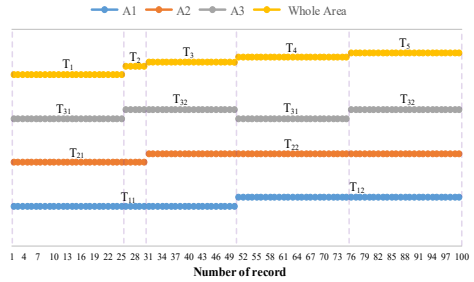

Fig. 10. The distribution of the testing data in the large-scale network.

The split-EM procedure in Fig. 3 is utilized in the whole area and all the subsystems, respectively. The corresponding split-EM processes are shown in Fig. 11 and Fig. 12. Although the results are correct in both scenarios, the split-EM process for the entire system needs 5 rounds and the average calculation time of each round is about 2000 seconds, while it only takes 2 rounds in each subsystem and the total calculation time of each subsystem is just about 150 seconds on average.

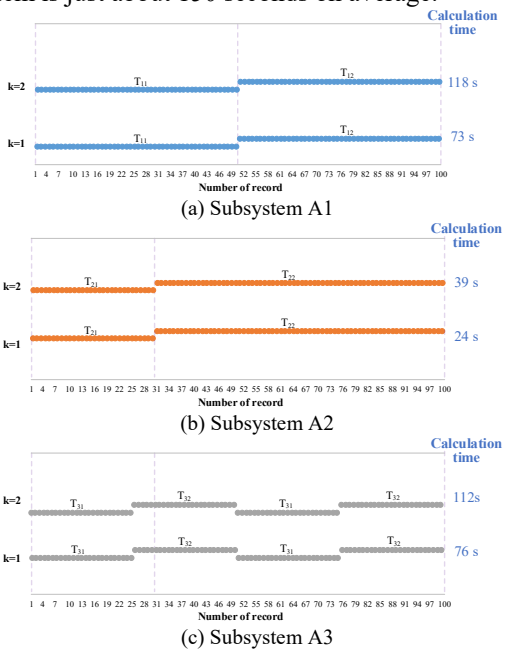

Fig. 11. The results of split-EM process for each subsystem.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPWRS.2021.3119649, IEEE Transactions on Power Systems

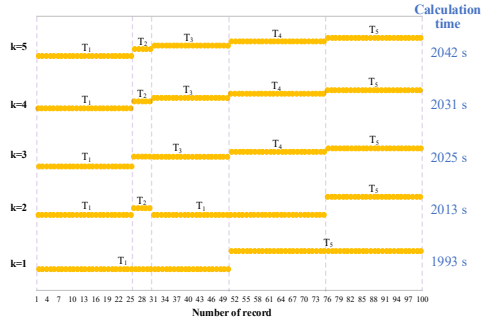

Fig. 12. The results of split-EM process for the whole area.

2) The real-time topology identification

In this section, the real-time identification processes of the whole area and the subsystems will be compared. We consider 4 topology categories in each subsystem and 12 topology categories in the whole area involved in the historical data in this testing. 12000 records are used in the training process for the whole area and each subsystem, and 1200 records are used in the testing process. The comparison results are shown in Table XI.

TABLE XI THE COMPARISON OF REAL-TIME RESULTS OF 135-BUS SYSTEM

Area range	Training time (sec)	Training Accuracy	Testing Accuracy	Prediction time (sec)
Whole area	97.3	99.50%	98.00%	0.10
A1	23.9	99.60%	98.20%	0.01
A2	18.8	99.20%	98.90%	0.01
A3	13.1	99.80%	99.40%	0.01

As shown in Table XI, the training/prediction time of the whole area is longer than those of the subsystems. The training accuracy difference between each subsystem and the whole area is within ± 0.5%, and the testing accuracy of subsystems are a bit higher than that of the whole area.

In conclusion, the performance of both the historical and real-time identification in the subsystems is as good as that in the whole area in terms of accuracy, and better in terms of identification time. After the topology identification of each subsystem is done, the topology information of the whole area can be obtained by combining the results of the subsystems based on the acquisition time of the measurements.

3) Subsystem optimization in the 874-bus system

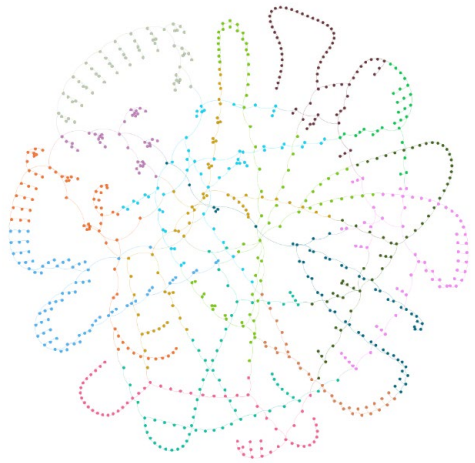

Fig. 13. Subsystem optimization result of the 874-bus system.

Asides from the 135-bus system, an 874-bus system (as shown in Fig. 13) is used to validate the model in subsection V-B. GA and graph related toolbox [44, 45] are utilized in the Matlab environment to realize the optimization of the subsystems of the 874-bus system, and N_{limit} and $N_{\text{s\_limit}}$ are set as 80 and 5. The optimized subsystem formation result is shown in Fig. 13, and 15 subsystems are formed according to the optimization. The optimized subsystems are marked with different colors, and the edge with two terminal nodes in different colors are the connecting lines between the subsystems. The topology identification tasks in each subsystem are performed similarly to those in the 135-bus system, and will not be presented due to limitation of space.

VI. CONCLUSION

To improve the topology observability in the PDN, a two-stage topology identification framework is designed to recognize the mixed topologies in the historical batch data and predict the real-time topology. A split expectation-maximization (EM) method is proposed to deal with the topology identification problem of a large set of historical data in which the number of topology categories is unknown. The calculation complexities are compared between the split-EM and traditional EM methods, proving that the proposed split-EM method consumes less calculation resources in most cases. To predict the real-time topology efficiently, the topology classifiers are trained based on the labeled historical records through machine learning methods. An error-correcting mechanism consisting of the credibility analysis and reidentification process based on the Bayesian recursion model is also proposed for the real-time identification to improve its performance. The effectiveness of the models is verified in a test system. The proposed split-EM can identify the topology categories correctly in the cases for radial network and the networks with meshes and islands, and it also works well when the nodal phase angles are not measured. The split-EM model is also extended to adapt to the three-phase unbalanced systems. For the real-time topology identification, several highly-accurate classifiers are generated using the Classification Learner and Neural Net Pattern Recognition toolboxes in MATLAB, and their prediction time for a single case is all less than 0.4 sec. Taking the classifier based on the Quadratic SVM model as an example, the error-correcting mechanism is verified in terms of improving the real-time prediction accuracy from 97.9% to 99.9%. The sensitivity of the topology categories is also analyzed, and the results show that reducing the number of topology categories in the historical identification will benefit real-time identification in terms of reducing prediction time and improving prediction accuracy. In addition, the application solution of the topology identification models in large-scale PDNs is proposed without extra measurements, which can update the overall topology information dynamically and efficiently with less calculation burden. The application in the 135-bus system has verified the advantages of the proposed extension framework both in historical and real-time identification. In the future, the proposed models in the paper can be further expanded to address other related issues, such as the real-time topology identification combined with fault detection and localization.

REFERENCES

[1] J. Yu, Y. Weng, and R. Rajagopal, "PaToPa: A data-driven parameter and topology joint estimation framework in distribution grids," *IEEE Transactions on Power Systems,* vol. 33, no. 4, pp. 4335-4347, 2018.

[2] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems,* vol. 32, no. 4, pp. 2682-2694, 2016.

[3] M. Babakmehr, M. G. Simoes, M. B. Wakin, A. Al Durra, and F. Harirchi, "Smart-grid topology identification using sparse recovery," *IEEE Transactions on Industry Applications,* vol. 52, no. 5, pp. 4375-4384, 2016.

[4] S. Mousavizadeh, M.-R. Haghifam, and M.-H. Shariatkhah, "A linear two-stage method for resiliency analysis in distribution systems considering renewable energy and demand response resources," *Applied Energy,* vol. 211, pp. 443-460, 2018.

[5] S. Chanda, A. K. Srivastava, M. U. Mohanpurkar, and R. Hovsapian, "Quantifying power distribution system resiliency using code-based metric," *IEEE Transactions on Industry Applications,* vol. 54, no. 4, pp. 3676-3686, 2018.

[6] M. Thomson, and D. G. Infield, "Network power-flow analysis for a high penetration of distributed generation," *IEEE Transactions on Power Systems,* vol. 22, no. 3, pp. 1157-1162, 2007.

[7] J. H. Eto, E. Stewart, T. Smith, M. Buckner, H. Kirkham, F. Tuffner, and D. Schoenwald, "Scoping study on research and development priorities for distribution-system phasor measurement units," Lawrence Berkeley National Laboratory, Berkeley, CA, USA, Technical Report LBNL-1003915, Dec. 2015.

[8] D. Deka, S. Backhaus, and M. Chertkov, "Structure learning in power distribution networks," *IEEE Transactions on Control of Network Systems,* vol. 5, no. 3, pp. 1061-1074, 2018.

[9] M. Farajollahi, A. Shahsavari, and H. Mohsenian-Rad, "Topology identification in distribution systems using line current sensors: An milp approach," *IEEE Transactions on Smart Grid,* vol. 11, no. 2, pp. 1159-1170, 2019.

[10] J. Zhao, L. Li, Z. Xu, X. Wang, H. Wang, and X. Shao, "Full-scale Distribution System Topology Identification Using Markov Random Field," *IEEE Transactions on Smart Grid,* vol. 11, no. 6, pp. 4714-4726, 2020.

[11] M. Babakmehr, M. G. Simões, M. B. Wakin, and F. Harirchi, "Compressive sensing-based topology identification for smart grids," *IEEE Transactions on Industrial Informatics,* vol. 12, no. 2, pp. 532-543, 2016.

[12] Y. Zhao, J. Chen, and H. V. Poor, "A learning-to-infer method for real-time power grid multi-line outage identification," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 555-564, 2019.

[13] R. Singh, E. Manitsas, B. C. Pal, and G. Strbac, "A recursive Bayesian approach for identification of network configuration changes in distribution system state estimation," *IEEE Transactions on Power Systems,* vol. 25, no. 3, pp. 1329-1336, 2010.

[14] O. Ardakanian, Y. Yuan, V. Wong, R. Dobbe, S. Low, A. von Meier, and C. J. Tomlin, "On identification of distribution grids," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 950-960, 2019.

[15] J. Yu, Y. Weng, and R. Rajagopal, "PaToPaEM: A Data-Driven Parameter and Topology Joint Estimation Framework for Time-Varying System in Distribution Grids," *IEEE Transactions on Power Systems,* vol. 34, no. 3, pp. 1682-1692, 2018.

[16] N. Zhou, L. Luo, G. Sheng, and X. Jiang, "Power Distribution Network Dynamic Topology Awareness and Localization Based on Subspace Perturbation Model," *IEEE Transactions on Power Systems,* vol. 35, no. 2, pp. 1479-1488, 2019.

[17] W. Wang, and N. Yu, "Maximum Marginal Likelihood Estimation of Phase Connections in Power Distribution Systems," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 3906-3917, 2020.

[18] S. Liu, X. Cui, Z. Lin, Z. Lian, Z. Lin, F. Wen, Y. Ding, Q. Wang, L. Yang, and R. Jin, "Practical Method for Mitigating Three-Phase Unbalance Based on Data-Driven User Phase Identification," *IEEE Transactions on Power Systems,* vol. 35, no. 2, pp. 1653-1656, 2020.

[19] T. H. Chen, W. T. Huang, J. C. Gu, G. C. Pu, Y. F. Hsu, and T. Y. Guo, "Feasibility study of upgrading primary feeders from radial and open-loop to normally closed-loop arrangement," *IEEE Transactions on Power Systems,* vol. 19, no. 3, pp. 1308-1316, 2004.

[20] J. C. Kim, S. M. Cho, and H. S. Shin, "Advanced power distribution system configuration for smart grid," *IEEE Transactions on Smart Grid,* vol. 4, no. 1, pp. 353-358, 2013.

[21] R. H. Lasseter, and P. Paigi, "Microgrid: A conceptual solution," in IEEE 35th Annual Power Electronics Specialists Conference, 2004, pp. 4285-4290.

[22] M. Geidl, G. Koeppel, P. Favre-Perrod, and B. Klockl, "Energy hubs for the future," *Power & Energy Magazine IEEE,* vol. 5, no. 1, pp. 24-30, 2007.

[23] T. Li, L. Werner, and S. H. Low, "Learning Graphs from Linear Measurements: Fundamental Trade-offs and Applications," *IEEE Transactions on Signal and Information Processing over Networks,* vol. 6, pp. 163-178, 2020.

[24] Z. Tian, W. Wu, and B. Zhang, "A mixed integer quadratic programming model for topology identification in distribution network," *IEEE Transactions on Power Systems,* vol. 31, no. 1, pp. 823-824, 2015.

[25] T. Erseghe, S. Tomasin, and A. Vigato, "Topology estimation for smart micro grids via powerline communications," *IEEE Transactions on Signal Processing,* vol. 61, no. 13, pp. 3368-3377, 2013.

[26] J. Zhang, Y. Wang, Y. Weng, and N. Zhang, "Topology identification and line parameter estimation for non-PMU distribution network: A numerical method," *IEEE Transactions on Smart Grid,* vol. 11, no. 5, pp. 4440-4453, 2020.

[27] R. A. Jabr, "Minimum loss operation of distribution networks with photovoltaic generation," *IET Renewable Power Generation,* vol. 8, no. 1, pp. 33-44, 2014.

[28] D. Schofield, F. Gonzalez-Longatt, and D. Bogdanov, "Design and implementation of a low-cost phasor measurement unit: a comprehensive review," in Seventh Balkan Conference on Lighting (BalkanLight), 2018, pp. 1-6.

[29] E. Y. Song, G. J. FitzPatrick, and K. B. Lee, "Smart sensors and standard-based interoperability in smart grids," *IEEE sensors journal,* vol. 17, no. 23, pp. 7723-7730, 2017.

[30] D. Chandler, "Introduction to modern statistical mechanics," *Introduction to Modern Statistical Mechanics, by David Chandler, pp. 288. Foreword by David Chandler. Oxford University Press, Sep 1987. ISBN-10: 0195042778. ISBN-13: 9780195042771*, pp. 288, 1987.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological),* vol. 39, no. 1, pp. 1-22, 1977.

[32] C. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics,* vol. 11, no. 1, pp. 95-103, 1983.

[33] N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton, "SMEM algorithm for mixture models," *Neural computation,* vol. 12, no. 9, pp. 2109-2128, 2000.

[34] L. Parra, and H. H. Barrett, "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET," *IEEE transactions on medical imaging,* vol. 17, no. 2, pp. 228-235, 1998.

[35] V. Quintana, and T. Van Cutsem, "Power system network parameter estimation," *Optimal Control Applications and Methods,* vol. 9, no. 3, pp. 303-323, 1988.

[36] Y. Wang, N. Zhang, H. Li, J. Yang, and C. Kang, "Linear three-phase power flow for unbalanced active distribution networks with PV nodes," *CSEE Journal of Power and Energy Systems,* vol. 3, no. 3, pp. 321-324, 2017.

[37] R. D. Zimmerman, C. E. Murillo-Sánchez, and D. Gan, "Matpower," *PSERC.[Online]. Software Available at: http://www. pserc. cornell. edu/matpower*, 1997.

[38] D. M. Etter, D. C. Kuncicky, and D. W. Hull, *Introduction to MATLAB*: Prentice Hall, 2002.

[39] ANSI C12.20-2015-Electricity Meters-0.1, 0.2, and 0.5 Accuracy Classes [Online]. Available: https://blog.ansi.org/2017/05/ansi-c1220-2015-electricity-meters-accuracy-classes/

[40] OPTI Toolbox [Online]. Available: https://inverseproblem.co.nz/OPTI/

[41] S. Le Digabel, *NOMAD: Nonlinear optimization with the MADS algorithm*: Groupe d'études et de recherche en analyse des décisions, 2010.

[42] J. R. Mantovani, F. Casari, and R. A. Romero, "Reconfiguração de sistemas de distribuição radiais utilizando o critério de queda de tensão," *Controle and Automacao*, pp. 150-159, 2000.

[43] Data for 135-bus System [Online]. Available: https://people.ece.ubc.ca/hameda/download_files/node_135.m

[44] Genetic Algorithm [Online]. Available: https://www.mathworks.com/help/gads/genetic-algorithm.html

[45] Graph and Network Algorithms [Online]. Available: https://www.mathworks.com/help/matlab/graph-and-network-algorithms.html