

Partial Correlation Analysis Based Identification of Distribution Network Topology

Yanli Liu, *Member, IEEE, Member, CSEE*, and Peng Wang, *Member, IEEE*

Abstract—Accurately identifying distribution network topology, which tends to be a mesh configuration with increasing penetration rate of distributed energy resources (DERs), is critical for reliable operation of a smart distribution network. Multicollinearity among node voltages makes existing topology identification methods unstable and inaccurate. Considering partial correlation analysis can reveal the intrinsic correlation of two variables by eliminating the influence of other variables, this paper develops a novel data-driven method based on partial correlation analysis to identify distribution network topology (radial, mesh, or including DERs) using only historical voltage amplitude data. First, maximum spanning tree of network is generated through Prim algorithm. Then, the loops of network are identified by taking tree neighbors as controlling variables in partial correlation analysis. Finally, a new topology verification mechanism based on partial correlation analysis is developed to correct wrong connections caused by multicollinearity. Test results on IEEE 33-node system, IEEE 123-node system and practical distribution network demonstrate that our method outperforms common data-driven methods, and can robustly identify both radial and mesh distribution network with DERs.

Index Terms—Data-driven, linear correlation, partial correlation, smart meter, topology identification.

I. INTRODUCTION

ACCURATE information of topology structure is crucial for distribution network operation, since its knowledge is the prerequisite for most advanced system analyses such as state estimation, optimal power flow, security assessment and fault location [1], [2]. Compared with transmission grids, a distribution network has lower measurement redundancy but more changeable topology [3], [4]. To achieve zero carbon emissions goal, a high proportion of distributed energy resources (DERs) will be integrated into distribution networks. With amounts of DERs, distribution networks may frequently perform topology reconstruction to deal with power balance and overvoltage problems [5]. Besides, to reduce grid losses

and improve reliability and efficiency, operators may operate the distribution network in a closed-ring layout [6], [7]. These issues have posed a huge challenge to topology identification in distribution networks. Therefore, an efficient topology identification method for smart distribution network is urgently needed more than ever before.

Several methods have been developed to identify distribution network topology in the past decade. Classical state estimation methods evaluate all possible topologies and minimize residuals error given measurements of the actual state of the grid [8], [9]. All switch locations are assumed available, then the right combination of switches is screened out in [10], [11]. These methods require prior knowledge of topology or circuit breakers, which may be inaccurate or even unavailable in distribution networks. Besides, additional equipment needs to be installed to monitor topology changes [12], [13], which involves expensive investment. A mixed-integer linear program (MILP) model-based split method is proposed to recognize the mixed topologies [14], which requires nodal active and reactive power injections. Furthermore, some recent strategies leverage machine learning methods to identify topology. [15] introduce two learning algorithms based on graphical models to detect topology estimation changes. [16] utilizes convolutional neural network (CNN) to identify the topology online. However, current machine learning methods usually require prior knowledge or a sufficient training dataset, which may be unavailable in practice.

Widespread deployment of smart meters provides a new opportunity for distribution network topology identification. As more useful measurement data becomes available, data-driven approaches can analyze mathematically specific and intrinsic relationships existing in measurement data to identify topology. Compared with the above methods, data-driven methods are more effective, economically feasible, and do not require prior knowledge. Existing methods can be classified two main groups.

The first group identifies the topology by analyzing correlation level among node voltages. [17], [18] extract the most relevant connections in voltage correlation matrix to generate a maximum spanning tree of radial network. [19] calculates mutual information among nodes, and then uses Kruskal algorithm to obtain radial topology. In [20], Markov random field is used to describe a distribution network and branches are screened out by adding penalty items.

The second group focuses on a multiple linear regression method due to its mathematical simplicity. [21], [22] formulate topology identification problems as a linear regression

Manuscript received November 10, 2021; revised December 14, 2021; accepted December 28, 2021. Date of online publication January 25, 2023; date of current version March 28, 2023. This work was supported by the National Key R&D Program of China (2020YFB0905900), and science and technology project of SGCC (State Grid Corporation of China) (SGTJDK00DWJS 2100223). Authors would like to thank ISSDA and Vienna University of Technology for providing households load data. Authors also would like to thank the researchers who share the representative distribution networks.

Y. L. Liu (corresponding author, email: yanlilui@tju.edu.cn), and P. Wang are with the School of electrical and information engineering, Tianjin University, Tianjin 300072, China.

DOI: 10.17775/CSEEJPES.2021.08360

problem, and set L1 regular penalty term in lasso regression to filter out neighboring nodes. Radial structure of topology is obtained by regression based on the principle of power conservation in [23]. Line parameter estimation results are used to modify basic topology obtained by regression [24]. Based on decoupled linear power flow (DLPF) equations, a total least squares regression method is proposed to estimate admittance matrix and topology of radial network [25].

However, multicollinearity exists among node voltage. For the first group method, it makes the correlation matrix very dense, so the first group only can be applicable to radial topology. For the second group method, multicollinearity makes regression matrix ill-conditioned, which has great impact on linear regression's performance and will further affect accuracy of topology identification method. Besides, many methods require multiple types of data such as active power, reactive power, voltage amplitude, phase angle, and data loss and asynchronous problems may further reduce accuracy of topology identification.

To solve the above issues, partial correlation analysis is introduced to identify topology in this paper. Partial correlation analysis can analyze the intrinsic correlation between two random variables through removing the influence of controlling variables [26], [27]. Mathematically, partial correlation analysis quantifies the intrinsic correlation level among node voltage reflected by original node voltage equation. Based on this, through gradually obtaining precise controlling variables (i.e., tree neighbors and loop neighbors) topology is identified with high accuracy.

Specifically, the maximum spanning tree of topology is firstly generated through Prim algorithm. Then tree neighbors are regarded as controlling variables in partial correlation analysis to screen out loop neighbors. Finally, a topology verification mechanism based on partial correlation analysis is designed to address errors caused by multicollinearity. The proposed method is tested on IEEE 33-node, 123-node system, and practical distribution networks with different types of topologies and results demonstrate the accuracy and efficiency of our method. Compared with existing approaches, the contributions of this paper are as follows:

1) Partial correlation analysis is first introduced into topology identification, which quantifies the intrinsic correlation among node voltages and weakens the adverse effects of multicollinearity.

2) Compared with current voltage-correlation-based approaches, the proposed method can identify a mesh distribution network without any prior knowledge.

3) The proposed method, based on partial correlation analysis, is easy to implement and has high accuracy and robustness for both radial and mesh topologies.

The rest of the paper is organized as follows: Section II is problem formulation. Section III uses partial correlation analysis to quantify the inherent linear correlation among node voltages. Section IV introduces specific details of the method. Case studies and conclusions are given in Section V and VI, respectively.

II. PROBLEM FORMULATION

In this section, certain concepts of graph theory pertaining to this work are introduced. Then the formulation of topology identification problem to be solved in this paper is given.

1) Graphs and Sub-graph

A m -nodes topology can be represented by graphs G containing node set $v = \{1, 2, \dots, m\}$ and edge set $\varepsilon = \{\varepsilon_{ij}, i, j \in v\}$. A graph is said to be directed if its edges are directed from one node to another, otherwise the graph is undirected. If $G_s = (v_s, \varepsilon_s)$ with set of nodes $v_s \subseteq v$ and set of edges $\varepsilon_s \subseteq \varepsilon$ such that G_s contains all edges with both end nodes in v_s , G_s is a sub-graph of G . A graph is said to be connected if any two nodes have a path, otherwise the graph is disconnected.

2) Trees and Loops

If G_s contains all the nodes of G , no loops and connections, then G_s is a tree, which is denoted by matrix T in this paper. There is one and only one path between any two nodes in a tree, and the number of edges of a tree is $m - 1$. If a tree is the generated subgraph of a undirected graph G , then the tree is called the spanning tree of G , and any connected graph G has at least one spanning tree. The weight of spanning tree is the sum of the weight of all edges

$$W(T) = \sum_{\varepsilon_{ij} \in T} W(\varepsilon_{ij}) \quad (1)$$

where $W(\cdot)$ represents weight function. The maximum spanning tree is the tree with the largest weight.

If the number of edges in a subgraph is greater than $m - 1$, the subgraph must contain loops. In this paper, matrix L is used to represent loops.

3) Adjacency Matrix

For a m -node distribution system, the adjacency matrix $A = (a_{ij})$ is a m -order Boolean matrix (i.e., 0, 1 matrix). Elements in the adjacency matrix A are defined as:

$$a_{ij} = \begin{cases} 1, & j \in N_i \\ 0, & j \notin N_i \end{cases} \quad (2)$$

where N_i represents the set of all nodes connected to node i .

4) Problem Formulation

The problem addressed in this paper is finding tree T and loops L , then generating and correcting adjacency matrix A of distribution network given all node voltage magnitude measurements except slack node.

III. NODE CORRELATION ANALYSIS VIA PARTIAL CORRELATION ANALYSIS

This section first introduces a partial correlation coefficient. Then the second subsection analyzes linear correlation among node voltages based on original node voltage equation. Finally, partial correlation analysis method is used to quantify the inherent correlation level among node voltages.

A. Partial Correlation Coefficient

For n time series $\{x_1(t)\}, \{x_2(t)\}, \dots, \{x_n(t)\}$, where $t = 1, 2, \dots, N$, N is the number of sampling data, let $X = \{x_1(t), x_2(t), \dots, x_n(t)\}$, consider any two time series $\{x_i(t)\}$ and $\{x_j(t)\}$ in X , the partial correlation between them can be briefly introduced as follows:

1) First let controlling variables $Z = X \setminus \{x_i(t), x_j(t)\}$ be the set of all variables except $\{x_i(t)\}$ and $\{x_j(t)\}$ in set X . Solving the linear regression problem amounts to finding regression coefficient vectors W_i^* and W_j^* such that

$$\begin{cases} W_i^* = \arg \min_w \sum_{t=1}^N (x_i(t) - \langle w, Z \rangle)^2 \\ W_j^* = \arg \min_w \sum_{t=1}^N (x_j(t) - \langle w, Z \rangle)^2 \end{cases} \quad (3)$$

where w is the regression coefficient, $\langle w, Z \rangle$ is the scalar product between w and Z .

2) Then calculate the residual

$$\begin{cases} e_i = x_i(t) - \langle W_i^*, Z \rangle \\ e_j = x_j(t) - \langle W_j^*, Z \rangle \end{cases} \quad (4)$$

3) Finally, the partial correlation coefficient between $\{x_i(t)\}$ and $\{x_j(t)\}$ given controlling variables Z , written ρ_{ij} , is the correlation between the residuals e_i and e_j

$$\rho_{ij} = \frac{e_i^T e_j}{\|e_i\| \cdot \|e_j\|} \quad (5)$$

In order to further explain the partial correlation coefficient, Fig. 1 shows its geometric interpretation. Z is the controlling variable or additional variable, and the residual vector lies on the hyperplane S_Z perpendicular to Z . The desired partial correlation is then the cosine of the angle φ between the projections e_i and e_j of $x_i(t)$ and $x_j(t)$, respectively, onto the hyperplane perpendicular to Z . That is, through performing correlation analysis on the hyperplane S_Z , the partial correlation coefficient removes the influence of the controlling variables Z [28].

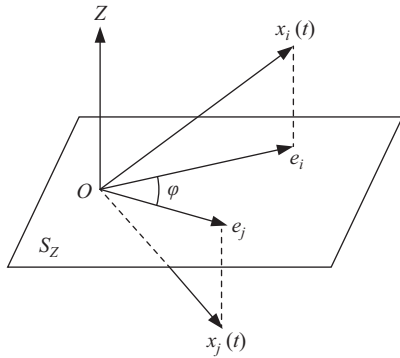


Fig. 1. Geometric interpretation of partial correlation coefficient.

Accordingly, when given voltage measurement data of n nodes, we can generate a $n \times n$ partial correlation coefficient matrix

$$\rho = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{bmatrix}_{n \times n} \quad (6)$$

Actually, Pearson correlation coefficient is the 0th-order partial correlation coefficient (i.e., with $|Z| = 0$), but in a distribution network, any two node voltages' Pearson correlation coefficient is usually close to 1 [20], serious multicollinearity exists so the intrinsic correlation among node voltages is hidden.

B. Intrinsic Correlation Among Node Voltages

For a m node distribution network, we have the following node voltage equation:

$$\dot{I}_i = \dot{V}_i y_{ii} - \sum_{k \in N_i} \dot{V}_k y_{ik} \quad (7)$$

where $\dot{I}_i$ is the current injection of node i , $\dot{V}_i$ is the voltage of node i , $\dot{V}_k (k \in N_i)$ are the voltages on neighboring node of node i , and y_{ii} , y_{ik} are the self admittance and mutual admittance.

Rewrite the equation above as follows

$$\dot{V}_i = \sum_{k \in N_i} \frac{y_{ik}}{y_{ii}} \dot{V}_k + \frac{1}{y_{ii}} \dot{I}_i \quad (8)$$

y_{ik} , $\dot{V}_i$ and $\dot{I}_i$ are all complex numbers, which can be expressed in polar coordinates, i.e. $y_{ik} = |y_{ik}| \exp(j\phi_{ik})$, $\dot{V}_i = V_i \exp(j\theta_i)$ and $\dot{I}_i = I_i \exp(j\beta_i)$, Then (8) becomes:

$$\begin{aligned} V_i \exp(j\theta_i) &= \sum_{k \in N_i} \frac{|y_{ik}|}{|y_{ii}|} V_k \exp(j(\phi_{ik} - \phi_{ii} + \theta_k)) \\ &+ \frac{I_i}{|y_{ii}|} \exp(j(\beta_i - \phi_{ii})) \end{aligned} \quad (9)$$

Changing the above equation, we have

$$\begin{aligned} V_i &= \sum_{k \in N_i} \frac{|y_{ik}|}{|y_{ii}|} V_k \exp(j(\phi_{ik} - \phi_{ii} + \theta_k - \theta_i)) \\ &+ \frac{I_i}{|y_{ii}|} \exp(j(\beta_i - \phi_{ii} - \theta_i)) \end{aligned} \quad (10)$$

Distribution network usually assumes the voltage deviation between adjacent nodes is small, i.e., $\theta_k - \theta_i \approx 0$, the following equation is approximations to (10).

$$\begin{aligned} V_i &\approx \sum_{k \in N_i} \frac{|y_{ik}|}{|y_{ii}|} V_k \exp(j(\phi_{ik} - \phi_{ii})) \\ &+ \frac{I_i}{|y_{ii}|} \exp(j(\beta_i - \phi_{ii} - \theta_i)) \end{aligned} \quad (11)$$

The real part is equal, so we have

$$V_i = \sum_{k \in N_i} \frac{|y_{ik}|}{|y_{ii}|} V_k \cos(\phi_{ik} - \phi_{ii}) + \frac{I_i}{|y_{ii}|} \cos(\beta_i - \phi_{ii} - \theta_i) \quad (12)$$

Then

$$V_i = \sum_{k \in N_i} \lambda_{ik} V_k + \varphi_{vi} I_i \quad (13)$$

where $\lambda_{ik} = \frac{|y_{ik}|}{|y_{ii}|} \cos(\phi_{ik} - \phi_{ii})$, $\varphi_{vi} = \frac{1}{|y_{ii}|} \cos(\beta_i - \phi_{ii} - \theta_i)$, both of them can be regarded as real numbers. $\varphi_{vi} I_i$ is usually much smaller than $\sum_{k \in N_i} \lambda_{ik} V_k$ [22], hence, $\varphi_{vi} I_i$

can be approximated as a noise term in (13), Letting $\varepsilon = \varphi_{vi}I_i$, (13) can be approximate as

$$V_i \approx \sum_{k \in N_i} \lambda_{ij} V_k + \varepsilon \quad (14)$$

Therefore, V_i , V_k ($k \in N_i$) are approximately linearly related. When V_k ($k \in N_i$) is known, V_i and V_j ($j \notin \{i, N_i\}$) are approximately independent of each other. In next subsection, partial correlation analysis quantifies this intrinsic correlation.

C. Analyze Node Voltage Correlation Using Partial Correlation Analysis

In a m node distribution network, the partial correlation level between V_i and V_j ($j \notin \{i, N_i\}$) is analyzed. First, the neighboring node voltages are taken as controlling variables

$$Z = \{V_k(k \in N_i)\} \quad (15)$$

Then, calculate the partial correlation coefficient $\rho_{V_i V_j(j \notin \{i, N_i\})}$ between V_i and V_j ($j \notin \{i, N_i\}$) given Z .

V_i and V_j ($j \notin \{i, N_i\}$) are approximately conditionally independent given Z , so $\rho_{V_i V_j(j \notin \{i, N_i\})}$ will be close to 0. For example, for node 3 in IEEE 33-node system, $N_3 = \{2, 4\}$. Fig. 2 shows the partial correlation coefficients between V_3 and V_i given $Z = \{V_k(k \in N_3)\}$. As shown by the red curve, $\rho_{V_3 V_j(j \notin \{3, N_3\})}$ are close to 0 when ignoring the influence

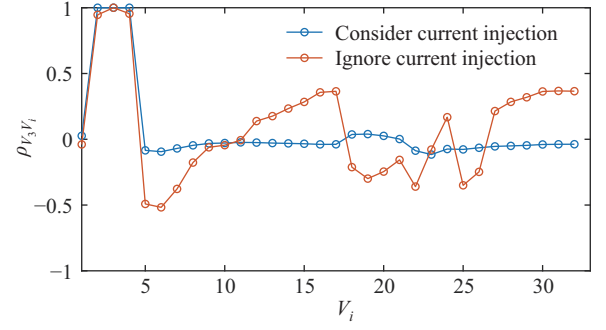


Fig. 2. The partial correlation coefficients between V_3 and V_i .

of I_3 . Let $Z = \{V_k(k \in N_3), I_3\}$, i.e., considering the influence of I_3 , more accurate results are shown by the blue curve, $\rho_{V_3 V_2} \approx \rho_{V_3 V_4} \approx 1$, $\rho_{V_3 V_j(j \notin \{3, N_3\})} \approx 0$. The above observations verify the rationality of above approximation.

Partial correlation analysis accurately quantifies the essential correlation level among node voltages, which can bring more accurate topology identification results. The correlation level among nodes calculated by four common methods are shown in Fig. 3, where color in the square represents the correlation level among nodes. The closer the square color is to yellow, the higher the correlation between the two nodes, as well as higher likelihood the two nodes are connected. As shown in Fig. 3(a)–(c), Pearson correlation matrix and mutual informa-

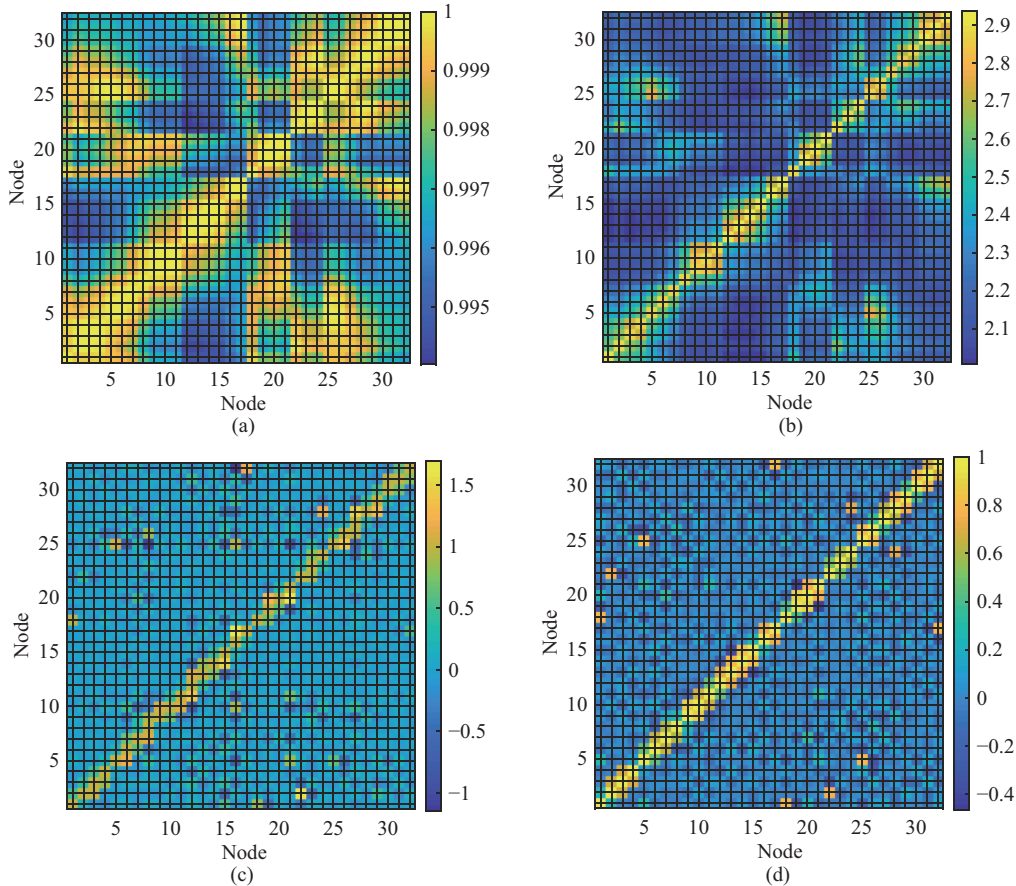


Fig. 3. Correlation matrix heat map of IEEE 33-node system for 4 methods. (a) Pearson correlation. (b) Mutual information. (c) Regression. (d) Partial correlation.

tion matrix are very dense, regression coefficient fluctuates drastically and is not accurate, and none of them can reveal the intrinsic correlation level and connection among nodes. Comparatively, as shown in Fig. 3(d), partial correlation matrix is a sparse matrix, where the partial correlation coefficient between neighboring nodes is close to 1 and others are close to 0. It vividly reveals connections between nodes in the topology.

The next section will gradually obtain accurate controlling variables (i.e., the set of neighboring nodes), and then utilize partial correlation analysis to generate and correct the topology.

IV. TOPOLOGY IDENTIFICATION METHOD BASED ON PARTIAL CORRELATION ANALYSIS

Main framework of the proposed topology identification method is shown in Fig. 4. In step 1, Prim algorithm searches correlation matrix of node voltages to generate the maximum spanning tree T , obtaining branch neighbors. In step 2, loop neighbors are screened out by taking branch neighbors as controlling variables based on partial correlation analysis. In step 3, topology verification is used to further correct wrong connections.

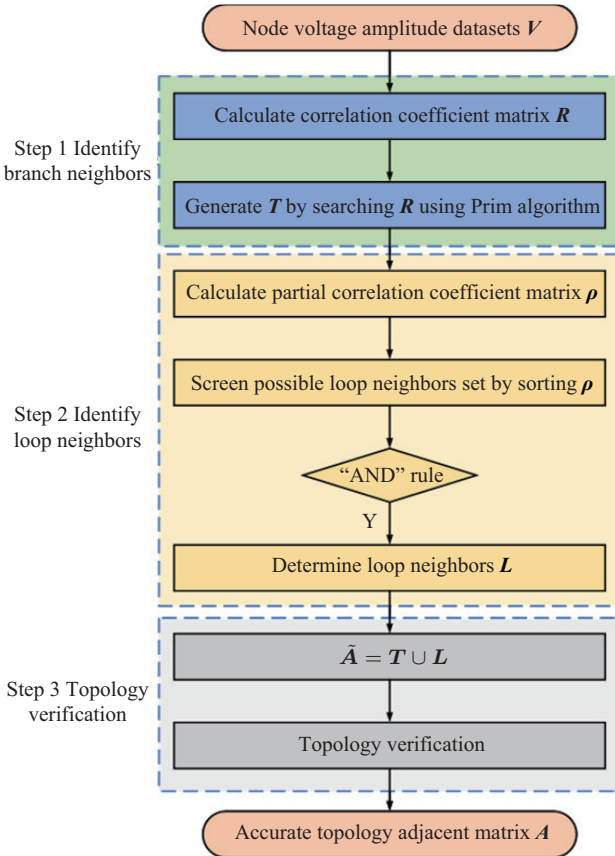


Fig. 4. Main framework of the proposed method.

A. Step 1: Identify Branch Neighbors

In a distribution network, neighboring nodes have the shortest electrical distance and therefore have a higher correlation

level [20]. Prim algorithm [29] is utilized to generate the maximum spanning tree of network. By constantly finding the most relevant connection in correlation matrix R , the algorithm can generate the optimal tree structure of topology [30]. Steps for implementing Prim algorithm are as follows:

Step 1: Initialize the maximum spanning tree with a starting node.

Step 2: Find all the edges that connect the tree to new nodes, find the edge with the maximum weight and add it to the tree.

Step 3: Keep repeating step 2 until get the maximum spanning tree T .

Based on the maximum spanning tree, branch neighbors of each node can be obtained, denoted by $N_T(i)$. The algorithm flow for identifying branch neighbors is shown in Algorithm 1.

Algorithm 1: Identify branch neighbors

Input: $V_i, i = 2, 3, \dots, m$

- 1 Compute correlation coefficient matrix R based on V_i .
- 2 Generate the maximum spanning tree T using Prim algorithm.
- 3 Obtain $N_T(i)$ based on T .

Output: T

B. Step 2: Identify Loop Neighbors

For each node i , a two-step framework is adopted to find its loop neighbors: the first step screens out a possible loop neighbors set. The second one takes branch neighbors as controlling variables in partial correlation analysis to exclude disconnection nodes in possible loop neighbors set and identify loop neighbors.

1) Screen Possible Loop Neighbors

Node i has a higher correlation level with its neighbor nodes, by selecting the top λ nodes with the largest correlation level to node i , the possible loop neighbors set $N_\lambda(i)$ can be screened out. Notice that any common methods such as Pearson correlation, mutual information and partial correlation is feasible. This paper uses partial correlation matrix ρ to obtain possible loop neighbors set $N_\lambda(i)$. The value of λ will be discussed in Section IV-D.

2) Determine Loop Neighbors

Voltages on non-neighboring nodes are approximately conditionally independent when neighboring nodes are given, so the partial correlation coefficient between them is close to 0. For each node i , branch neighbors set $N_T(i)$ and possible loop neighbors set $N_\lambda(i)$ are available. When $N_T(i)$ are treated as controlling variables, most nodes in $N_\lambda(i)$ that are not connected to node i can easily be excluded.

To be more specific, for each node i , its branch neighboring node voltages are treated as controlling variables

$$Z_i = \{V_k(k \in N_T(i))\} \quad (16)$$

Then, for each node j in loop neighbors set $N_\lambda(i)$, calculate the partial correlation coefficient ρ_{ij} between V_i and V_j given Z_i .

Similarly, take the branch neighboring node voltages of node j as controlling variables

$$Z_j = \{V_k(k \in N_T(j))\} \quad (17)$$

Then, calculate the partial correlation coefficient ρ_{ji} between V_i and V_j given Z_j .

Neighbor nodes should meet “AND” rule, which can find each node’s neighbor under the premise of ensuring accuracy [31]. Specifically, if node i and j are connected, node i is a neighbor of node j , and node j must also be a neighbor of node i . So node i and j must satisfy “AND” rule

$$\min(\rho_{ij}, \rho_{ji}) > \gamma_1 \quad (18)$$

When (18) holds, node j is a loop neighbor of node i . Traverse all the nodes, all the loops in the network can be identified, and store them in matrix $\mathbf{L}$. Threshold γ_1 will be discussed in Section IV-D. The algorithm flow for identifying loop neighbors is shown in Algorithm 2.

Algorithm 2: Identify loop neighbors

Input: $\mathbf{T}, \gamma_1$

- 1 Compute partial correlation coefficient matrix ρ based on V_i .
- 2 Screen possible loop neighbors set $N_\lambda(i)$ for bus i based on ρ .
- 3 **for** $j \in N_\lambda(i)$ **do**
- 4 $Z_i = N_T(i), Z_j = N_T(j)$
- 5 Calculate ρ_{ij} between V_i and V_j given Z_i .
- 6 Calculate ρ_{ji} between V_i and V_j given Z_j .
- 7 **if** $\min(\rho_{ij}, \rho_{ji}) > \gamma_1$ **then**
- 8 $\mathbf{L} \leftarrow (i, j)$
- 9 **end**
- 10 **end**

Output: $\mathbf{L}$

C. Topology Verification

Multicollinearity may cause $\mathbf{T}$ and $\mathbf{L}$ to contain wrong edges. The task of topology verification is to address these errors.

When two nodes are connected, the partial correlation coefficient between them is approximately 1. When the two nodes are disconnected, the partial correlation coefficient between them is close to 0 when taking neighboring nodes as controlling variables. Based on the above principles, the topology verification mechanism is designed to verify and correct the topology.

First, Combining $\mathbf{T}$ and $\mathbf{L}$, the preliminary adjacency matrix $\tilde{\mathbf{A}}$ is generated, let $N_{\tilde{\mathbf{A}}}(i)$ represent the neighboring nodes set of node i in $\tilde{\mathbf{A}}$, which contains both branch and loop neighbors of node i .

For each edge (i, j) , take the neighboring node voltages of node i and j in matrix $\tilde{\mathbf{A}}$ as controlling variables, respectively

$$\tilde{Z}_i = \{V_k(k \in N_{\tilde{\mathbf{A}}}(i))\} \quad (19)$$

$$\tilde{Z}_j = \{V_k(k \in N_{\tilde{\mathbf{A}}}(j))\} \quad (20)$$

Then calculate the partial correlation coefficient $\tilde{\rho}_{ij}$ between V_i and V_j given $\tilde{Z}_i$, and calculate the partial correlation coefficient $\tilde{\rho}_{ji}$ between V_i and V_j given $\tilde{Z}_j$.

If the partial correlation coefficient between two considered nodes is approximately 0, $\tilde{\rho}_{ij}$ and $\tilde{\rho}_{ji}$ will not satisfy “AND” rule. So, the two nodes are not connected or the line is wrong and (i, j) will not appear in the final adjacency matrix $\mathbf{A}$. In this way, we can screen out the real neighbors.

The algorithm flow for topology verification is shown as Algorithm 3.

Algorithm 3: Topology verification

Input: $\mathbf{T}, \mathbf{L}$

- 1 $\tilde{\mathbf{A}} = \mathbf{T} \cup \mathbf{L}$
 - 2 **for** $(i, j) \in \tilde{\mathbf{A}}$ **do**
 - 3 $\tilde{Z}_i = \{V_k(k \in N_{\tilde{\mathbf{A}}}(i))\}, \tilde{Z}_j = \{V_k(k \in N_{\tilde{\mathbf{A}}}(j))\}$
 - 4 Calculate $\tilde{\rho}_{ij}$ between V_i and V_j given $\tilde{Z}_i$.
 - 5 Calculate $\tilde{\rho}_{ji}$ between V_i and V_j given $\tilde{Z}_j$.
 - 6 **if** $\min(\tilde{\rho}_{ij}, \tilde{\rho}_{ji}) > \gamma_1$ **then**
 - 7 $\mathbf{A} \leftarrow (i, j)$
 - 8 **end**
 - 9 **end**
- Output:** $\mathbf{A}$
-

D. Discussion

1) Suitable Value of λ

λ decides whether loop neighbors will be filtered into possible loop neighbors set. The distribution network is a regular complex network. For the distribution network this paper analyzed, the average number of neighbors of each node is 2, and the maximum is 4. So λ can be set as follows for a m nodes distribution network

$$\lambda = \text{ceil}(\sqrt{m}) \quad (21)$$

where $\text{ceil}(\cdot)$ means round up function. For IEEE 33-node system, $\lambda = 6$, and for IEEE 123-node system, $\lambda = 12$. It is much larger than the maximum number of neighboring nodes and usually ensures that loop neighbors will be not missed.

2) Suitable Value of γ_1

γ_1 is a threshold used to determine whether two nodes are connected. If γ_1 is too large, real neighbors will be missed. If γ_1 is too small, it will lose the ability to screen. In this paper, γ_1 is set as follows

$$\gamma_1 = \min(\tilde{T}) \quad (22)$$

where $\tilde{T}$ represents the set of partial correlation coefficient between neighboring nodes in $\mathbf{T}$. In practice, γ_1 fluctuates in the range of 0.01 to 0.3.

V. CASE STUDY

In this section, the proposed method will be tested on IEEE 33-node system [32], IEEE 123-node system [33], and China’s typical rural 49-node system [34]. Section V-A introduces the platform, data source, feeders and evaluation index. Section V-B first shows how the topology identification

framework works, then tests the proposed method on different types of feeders. In Section V-C, comparison results of partial correlation analysis methods, mutual information, Pearson, and regression methods are shown. In the lastSection V-E, how noise levels, load levels, and data resolutions affect method performance are discussed.

A. Platform, Data Source, Feeders and Evaluation Index

1) Simulation Platform

All tests are performed on a laptop with i5-8250U and 8.0 GB RAM, the CPU frequency is 1.6GHZ, and the simulation software is MATLAB R2018a.

2) Data Source

In order to simulate behavior of the power system in a more realistic pattern, the real load curve from CER [35] and “ADRES-Concept” Project of Vienna University of Technology [36] are used as the active power profile in the following simulations. For CER data set, the 48-day active power consumption of 4000 Irish households are used, and the sampling frequency is once every 0.5 h. The “ADRES-Concept” Project load curve contains the 14-day active and reactive load curves of 30 Upper-Austria households, with a sampling interval of 1 s. For the load curve of each node of distribution network, 10 to 300 household loads are allocated for each node depending on the node load capacity. Then node voltage V_i is obtained thorough power flow calculation.

3) Case Description

Case 1: IEEE 33-node system is a 12.66 kV distribution network with 32 transmission lines and four tie lines, as shown in Fig. 5.

Case 2: IEEE 123-node system is a 4.16 kV distribution network with three tie lines. Topology is shown in Fig. 6.

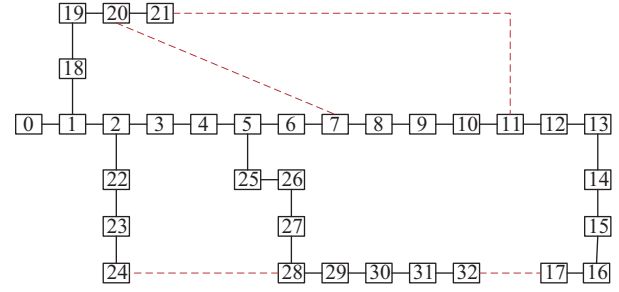


Fig. 5. IEEE 33-node system.

Case 3: China's typical rural 49-node system is a 12.66 kV radial distribution network, as shown in Fig. 7. It originates from the actual distribution network in Jiangsu and Zhejiang, China, and fully takes into account the current status of local DERs penetration. Load capacity is 2980 kW and the capacity of DERs is 760 kW.

4) Topology Connection Evaluation Indicator

Error defined in (23) is used to evaluate effectiveness of the proposed method:

$$\text{Error} = \frac{1}{|\tilde{\varepsilon}|} \left(\underbrace{\sum_{\varepsilon_{ij} \in \tilde{\varepsilon}} \eta(\varepsilon_{ij} \notin \varepsilon)}_{\text{false estimation}} + \underbrace{\sum_{\varepsilon_{ij} \in \varepsilon} \eta(\varepsilon_{ij} \notin \tilde{\varepsilon})}_{\text{missing}} \right) \times 100\% \quad (23)$$

where $\tilde{\varepsilon}$ is the edge set in the output result, ε denotes the true edge set of distribution network, $|\varepsilon|$ represents the size of ε , and $\eta(\cdot)$ is the indicator function. First and second terms represent the number of falsely estimated branches and the

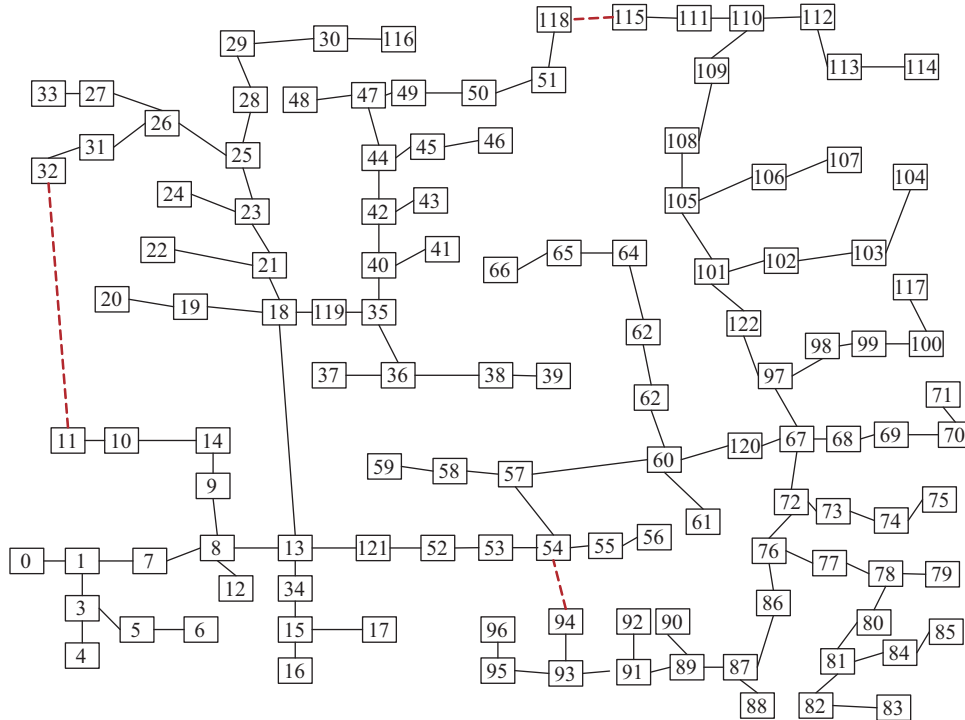


Fig. 6. IEEE123-node system.

number of missing branches, respectively. The value of each error is the average after 10 simulations.

B. Basic Results

1) Output of Each Step of the Framework

IEEE 33-node system is taken as an example to show output results of each step of the framework, as shown in Fig. 8.

Figure 8(a) represents the maximum spanning tree T , which is obtained by searching the Pearson correlation matrix

through Prim algorithm in Step 1. Fig. 8(b) represents the possible loop neighbor set $N_{\lambda}(i)$ of each node in Step 2; they are the set of maximum λ values in each column of matrix ρ . Based on “AND” rule, screened out loop neighbors are stored in matrix L . Combining T and L , the adjacency matrix $\tilde{A}$ is generated, as shown in Fig. 8(c). Topology verification can correct wrong connections in $\tilde{A}$, e.g., (1,3), (4,25), (8,21). The final accurate adjacency matrix A is shown in Fig. 8(d). Results verify the rationality of the proposed method based on

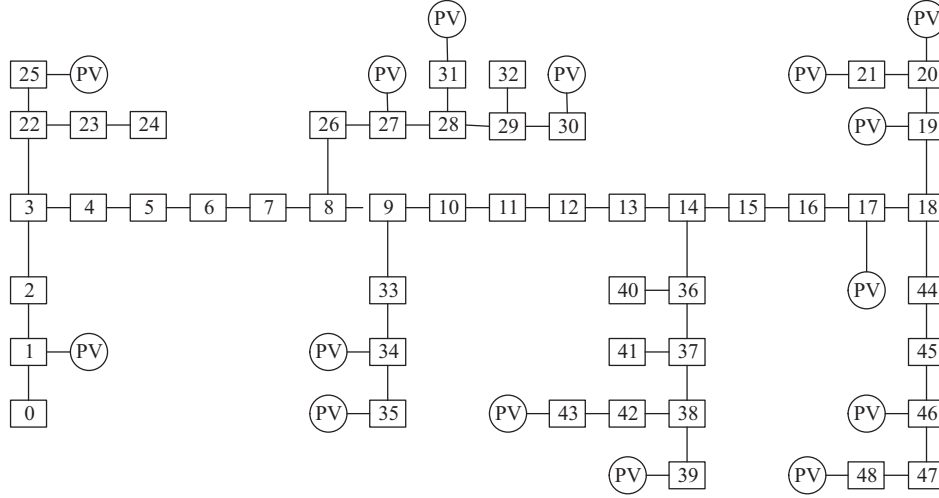


Fig. 7. 49-node system.

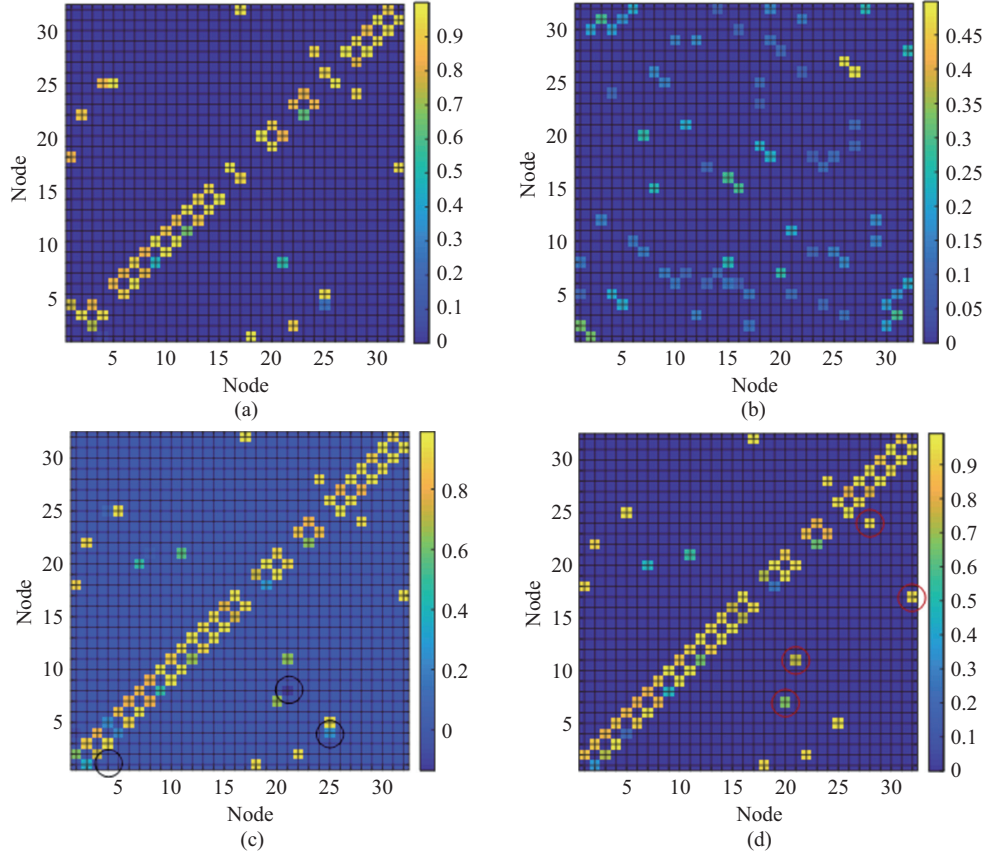


Fig. 8. Output of each step of the framework. Black circle: wrong branches. Red circle: tie lines. (a) Maximum spanning tree T . (b) set of possible loop neighbors. (c) Adjacency matrix $\tilde{A}$. (d) Adjacency matrix A .

partial correlation analysis.

Without prior knowledge, previous work based on voltage correlation analysis only can identify the maximum spanning tree T . For the proposed method, the last three steps are added to identify loops and correct wrong branches caused by multicollinearity, and line status can be visually displayed for operators. Therefore, the proposed method can extend applicability of previous work and improve accuracy of topology identification.

2) Test on Radial Networks Without DERs

Two radial topologies are generated by disconnecting the tie lines in Case 1 and Case 2. Adjacency matrices generated by the proposed method are shown in Fig. 9(a) and 9(b), respectively. Error of two cases are shown in Table I, error in Case 1 is 1.56%, and error in Case 2 is 1.07%. Results show the proposed method has high accuracy for radial topology without DERs.

3) Test on Radial Networks with Different Penetration of DERs

Case 1 and case 3 are taken as the test case and typical distributed power generation data in [34] and layout of DERs in [37] are used. The output adjacency matrix of case 3 is shown in Fig. 9(c). Performance of the proposed method under different penetration of DERs is shown in Table II. In both cases, error under different penetration rates are less than 4%, wrong branches are less than 2. It verifies the effectiveness of the proposed method in distribution networks with DERs.

TABLE I
TEST RESULTS ON RADIAL NETWORK WITHOUT DERs

System	Total branch	Error
33-node (radial)	32	1.56%
123-node (radial)	122	1.07%

TABLE II
TEST RESULTS ON NETWORK WITH DIFFERENT PENETRATION OF DERs

System	Total branch	Penetration rate	Error
33-node	36	15%	2.78%
		30%	1.67%
		45%	2.78%
49-node	48	15%	2.50%
		30%	3.27%
		45%	3.40%

TABLE III
TEST RESULTS ON MESH NETWORK

System	Total branch	Error
33-node with 4 loops	36	2.50%
123-node with 3 loops	125	1.84%

4) Test on Mesh Networks

Closing the tie lines in Case 1 and Case 2, two mesh networks are constructed, i.e., 33-node system with 4 loops and 123-node system with 3 loops. Fig. 9(d) shows the generated 123-node mesh network adjacency matrix. Table III lists the error rate, error in Case 1 is 2.50%, and error in Case 2 is 1.84%, which proves the effectiveness of the proposed

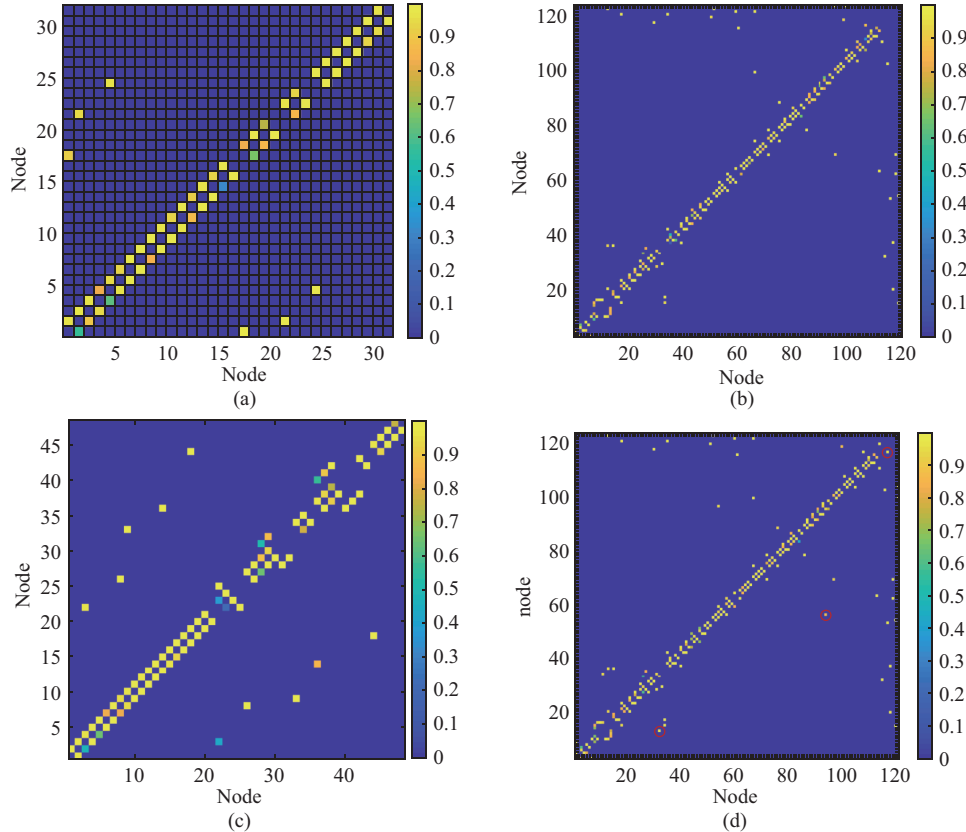


Fig. 9. The adjacency matrix identified by the proposed method in different networks. (a) Radial 33-node system. (b) Radial 123-node system. (c) Radial 49-node system. (d) Mesh 123-node system. Red circle: tie lines

method in a mesh distribution network.

5) Time Consumption

Average calculational time for the proposed method for each case are summarized in Table IV. It takes less than 0.4 s to identify IEEE 33-node system and about 6 s to identify IEEE 123-node system. Results show the proposed method can identify completely unknown topologies in a short time and is applicable to existing distribution networks.

TABLE IV
CALCULATIONAL TIME IN DIFFERENT CASES

System	Time consumption(s)
33-node	0.37
49-node	0.80
123-node	6.15

C. Comparison with Other Methods

Partial correlation analysis method is compared with the other three methods including the method based on the covariance matrix [17], mutual information [19], and linear regression method [24].

For radial topology, the proposed method is compared with covariance and mutual information methods in three cases; the results are shown in Fig. 10. With the same data set, the method based on mutual information performs poorly, the method based on covariance performs relatively good, the proposed method based on partial correlation analysis performs best in three cases. This comparison intuitively shows the advantages of the proposed method in identifying radial topology.

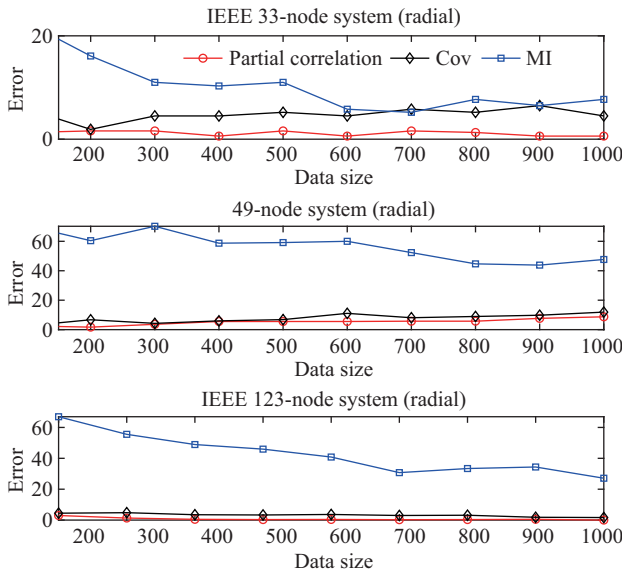


Fig. 10. Comparison of three methods in radial network.

For a mesh network, the proposed method is compared with linear regression method, i.e., Step 1 in [24], which requires active and reactive power data. Test results on IEEE 33-node system and IEEE 123-node system are shown in Fig. 11, when data volume reaches 200, identification accuracy of two methods tends to be stable and the proposed method performs better in both cases.

D. Sensitivity Analysis

1) Sensitivity to Data Accuracy

Performance of data-driven methods relies heavily on measurement data of smart meters. Modern medium-voltage distribution networks are usually equipped with 0.2 s level measuring devices. In the American ANSI C12.20 standard, the measurement error of smart meters is required to be less than 0.5% [38].

When voltage data is added 0.2% and 0.5% gaussian noise respectively, performance of the proposed method is shown in Fig. 12. It can be found the proposed method has sufficient identification accuracy under existing sensor noise level.

2) Sensitivity to Load Level

At different times of one day, the load level of distribution network will change. To understand our method's sensitivity to load levels, accuracy of the method at the specific moment

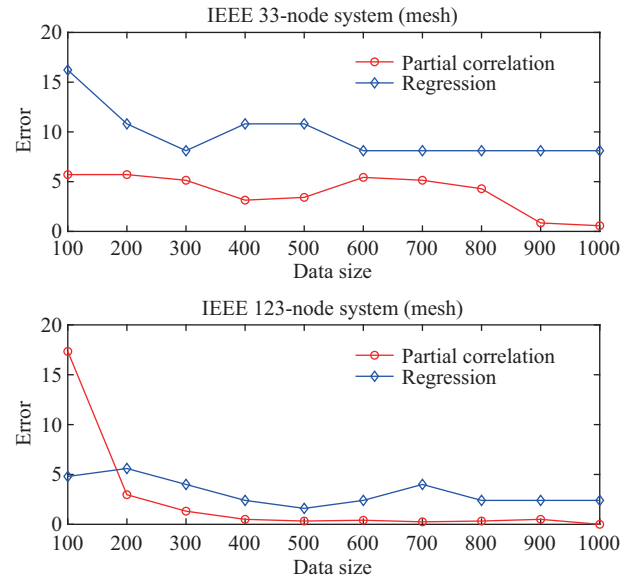


Fig. 11. Comparison of two methods in mesh network.

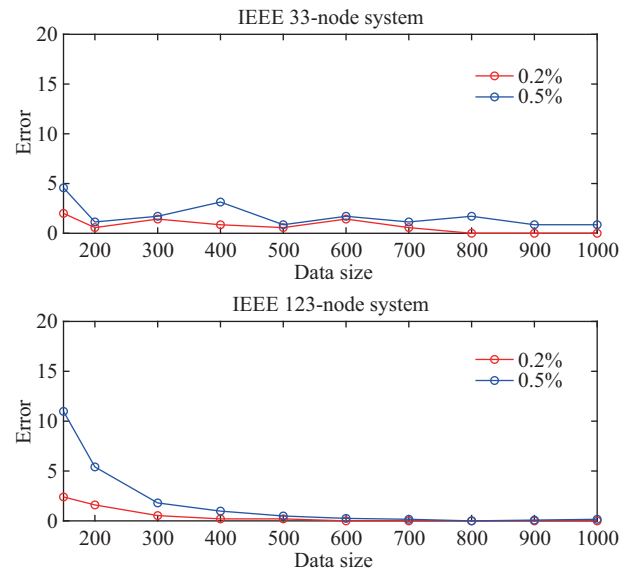


Fig. 12. Error under different noise levels.

is tested on Case 1 with “ADRES-Concept” dataset. The method is tested every 10 minutes with 200 data each time. Performance is shown in Fig. 13. For IEEE 33-node radial system, the average error rate is 0.61%. For mesh IEEE 33-nodes system, the average error rate is 4.74%. The number of wrong branches is close to 1, accuracy is adequate for operator to monitor the system.

3) Sensitivity to Data Resolution

The proposed method is tested on IEEE 123-node system with sampling interval 1 s, 5 min and 30 min, respectively as Fig. 14 shows. It can be obviously seen the accuracy of the proposed method begins to exceed 95% when data size exceeds 200 and has sufficient robustness for different resolutions.

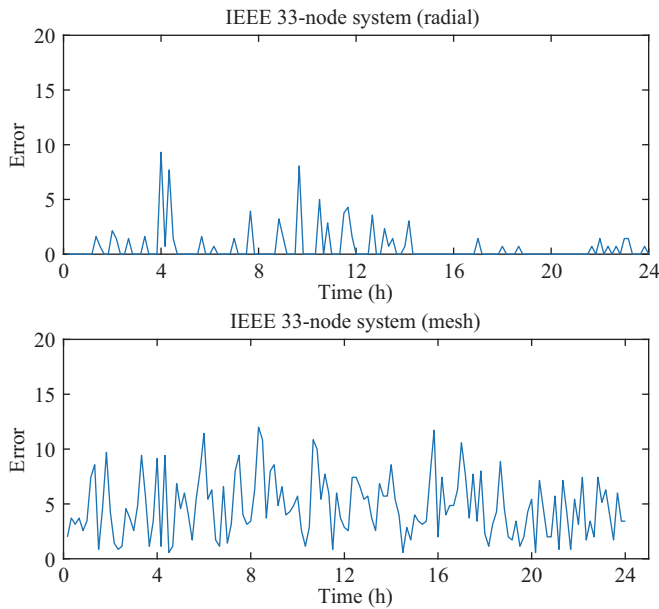


Fig. 13. Error under different load levels in one day.

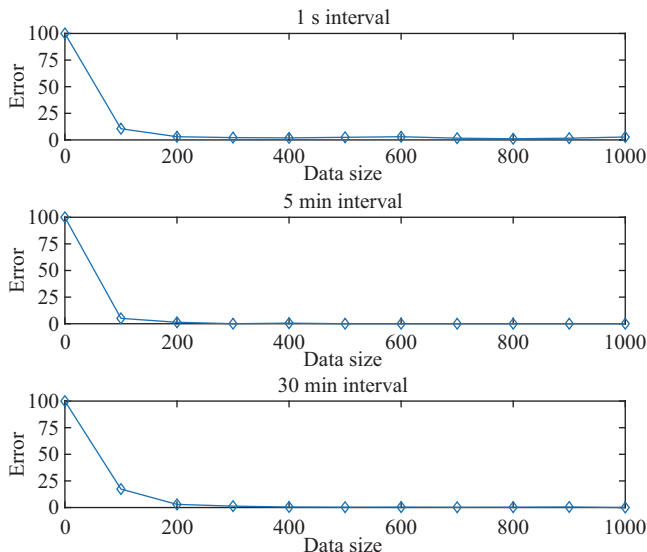


Fig. 14. Error under different data resolutions in IEEE 123-node system.

VI. CONCLUSION

A data-driven method based on partial correlation analysis is proposed in this paper to identify a distribution network topology. By gradually obtaining accurate controlling variables (i.e., the set of neighboring nodes), the method generates and corrects the topology based on partial correlation analysis. First, a branch neighbors set is obtained through Prim algorithm. Then, by eliminating the influence of branch neighbors, the loop neighbors set is identified through partial correlation analysis. Finally, a topology verification mechanism based on partial correlation analysis is designed to correct errors caused by multicollinearity. Case studies on IEEE 33-node system, 123-node system and practical distribution network indicate the proposed method can identify different complex networks (radial, mesh, or including DERs) with high accuracy. Compared with existing methods, the proposed method performs better and can robustly identify topology with limited voltage amplitude measurement data while considering smart meter data quality (i.e., noise, load level, resolution).

REFERENCES

- [1] J. Huang, V. Gupta, and Y. F. Huang, “Electric grid state estimators for distribution systems with microgrids,” in *Proceedings of the 46th Annual Conference on Information Sciences and Systems*, 2012, pp. 1–6.
- [2] W. H. Kersting, *Distribution System Modeling and Analysis*, 2nd ed., Boca Raton, FL, USA: CRC Press, 2007.
- [3] D. Das, “A fuzzy multiobjective approach for network reconfiguration of distribution systems,” *IEEE Transactions on Power Delivery*, vol. 21, no. 1, pp. 202–209, Jan. 2006.
- [4] X. M. Dong, C. Q. Kang, H. Sun, and N. Zhang, “Analysis of power transfer limit considering thermal balance of overhead conductor,” *IET Generation, Transmission & Distribution*, vol. 9, no. 14, pp. 2007–2013, Nov. 2015.
- [5] K. Zare and S. Nojavan, *Operation of Distributed Energy Resources in Smart Distribution Networks*, London: Academic Press, 2018.
- [6] Y. S. Jeon, D. Y. Shin, Y. U. Park, K. S. Kim, J. Y. Yu, and D. M. Kim, “Underground low voltage loop system of KEPCO,” in *Proceedings of the CIRED Workshop 2016*, 2016, pp. 1–4.
- [7] R. J. W. de Groot, J. Morren, and J. G. Slootweg, “Reliable and efficient operation of closed-ring distribution grids supported by distribution automation,” *Sustainable Energy, Grids and Networks*, vol. 15, pp. 53–62, Sep. 2018.
- [8] Y. J. Gao, Z. L. Zhang, W. C. Wu, and H. F. Liang, “A method for the topology identification of distribution system,” in *Proceedings of 2013 IEEE Power & Energy Society General Meeting*, 2013, pp. 1–5.
- [9] V. Zamani and M. Baran, “Topology processing in distribution systems by branch current based state estimation,” in *Proceedings of 2015 North American Power Symposium*, 2015, pp. 1–5.
- [10] Y. Sharon, A. M. Annaswamy, A. L. Motto, and A. Chakraborty, “Topology identification in distribution network with limited measurements,” in *Proceedings of 2012 IEEE PES Innovative Smart Grid Technologies*, 2012, pp. 1–6.
- [11] 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, Sep. 2018.
- [12] C. S. Chen, T. T. Ku, and C. H. Lin, “Design of phase identification system to support three-phase loading balance of distribution feeders,” in *Proceedings of 2011 IEEE Industrial and Commercial Power Systems Technical Conference*, 2011, pp. 1–8.
- [13] Z. Y. Shen, M. Jaksic, P. Mattavelli, D. Boroyevich, J. Verhulst, and M. Belkhat, “Three-phase AC system impedance measurement unit (IMU) using chirp signal injection,” in *Proceedings of 2013 Twenty-Eighth Annual IEEE Applied Power Electronics Conference and Exposition*, 2013, pp. 2666–2673.
- [14] L. Ma, L. Wu, N. Liu and W. Pei, “A two-step approach for multi-topology identification and parameter estimation of power distribution networks,” in *CSEE Journal of Power and Energy Systems*, doi: 10.17775/CSEEJPES.2021.08180.

- [15] D. Deka, S. Talukdar, M. Chertkov, and M. V. Salapaka, "Graphical models in meshed distribution grids: topology estimation, change detection & limitations," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4299–4310, Sep. 2020.
- [16] Y. F. Li, S. Y. Wang, L. Li, and P. Dehghanian, "Artificial intelligence for real-time topology identification in power distribution systems," in *Proceedings of the 52nd North American Power Symposium (NAPS)*, 2021, pp. 1–6, doi: 10.1109/NAPS50074.2021.9449727.
- [17] S. Bolognani, N. Bof, D. Michelotti, R. Muraro, and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *Proceedings of the 52nd IEEE Conference on Decision and Control*, 2013, pp. 1659–1664.
- [18] W. P. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1964–1971, Jul. 2015, doi: 10.1109/TSG.2015.2421304.
- [19] Y. Weng, Y. Z. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2682–2694, Jul. 2017, doi: 10.1109/TPWRS.2016.2628876.
- [20] J. Zhao, L. Li, Z. Xu, X. Y. Wang, H. B. Wang, and X. J. Shao, "Full-scale distribution system topology identification using markov random field," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4714–4726, Nov. 2020, doi: 10.1109/TSG.2020.2995164.
- [21] Y. Z. Liao, Y. Weng, and R. Rajagopal, "Urban distribution grid topology reconstruction via Lasso," in *Proceedings of 2016 IEEE Power and Energy Society General Meeting*, 2016, pp. 1–5.
- [22] Y. Z. Liao, Y. Weng, G. Y. Liu, and R. Rajagopal, "Urban MV and LV distribution grid topology estimation via group lasso," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 12–27, Jan. 2019, doi: 10.1109/TPWRS.2018.2868877.
- [23] S. J. Pappu, N. Bhatt, R. Pasumarthy, and A. Rajeswaran, "Identifying topology of low voltage distribution networks based on smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5113–5122, Sep. 2018, doi: 10.1109/TSG.2017.2680542.
- [24] J. W. 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, Sep. 2020, doi: 10.1109/TSG.2020.2979368.
- [25] J. W. Zhang, P. Wang, and N. Zhang, "Distribution network admittance matrix estimation with linear regression," *IEEE Transactions on Power Systems*, vol. 36, no. 5, pp. 4896–4899, Sep. 2021, doi: 10.1109/TPWRS.2021.3090250.
- [26] R. A. Fisher, "The distribution of the partial correlation coefficient," *Metron*, vol. 3, no. 3–4, pp. 329–332, 1924.
- [27] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, Dec. 2004, doi: 10.1111/j.1467–842X.2004.00360.x.
- [28] C. Yan, L. S. Zhai, H. X. Zhang, H. M. Wang, and N. D. Jin, "Cross-correlation analysis of interfacial wave and droplet entrainment in horizontal liquid-liquid two-phase flows," *Chemical Engineering Journal*, vol. 320, pp. 416–426, Jul. 2017, doi: 10.1016/j.cej.2017.03.044.
- [29] R. C. Prim, "Shortest connection networks and some generalizations," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, Nov. 1957.
- [30] J. B. Jr. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, Feb. 1956.
- [31] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, Jun. 2006.
- [32] M. E. Baran and F. F. Wu, "Network reconfiguration in distribution systems for loss reduction and load balancing," *IEEE Transactions on Power Delivery*, vol. 4, no. 2, pp. 1401–1407, Apr. 1989.
- [33] L. Bobo, A. Venzke, and S. Chatzivasileiadis, "Second-order cone relaxations of the optimal power flow for active distribution grids: comparison of methods," *International Journal of Electrical Power & Energy Systems*, vol. 127, pp. 106625, May 2021.
- [34] F. Luo et al., "HPDS133-bus: A Benchmark Test System for Distribution Systems with High Penetration of Distributed Generation," in *CSEE Journal of Power and Energy Systems*, vol. 9, no. 3, pp. 963–977, May 2023, doi: 10.17775/CSEEJPES.2020.01890.
- [35] C. for Energy Regulation (CER), "CER smart metering project electricity customer behaviour trial, 2009–2010," *Irish Social Sci. Data Arch.*, vol. 1, no. 1, 2012.
- [36] (2106). "Adres-dataset," Inst. Energy Syst. Elect. Drives, Vienna Univ. Technol., Vienna, Austria, [Online]. Available: http://www.ea.tuwien.ac.at/projects/adres_concept/EN/
- [37] M. R. Dorostkar-Ghamsari, M. Fotuhi-Firuzabad, M. Lehtonen, and A. Safdarian, "Value of distribution network reconfiguration in presence of renewable energy resources," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1879–1888, May 2016.
- [38] National Electrical Manufacturers Association, "American national standard for electricity meter-0.1, 0.2, and 0.5 accuracy classes," American National Standards Institute, Inc., Washington, DC, USA, ANSI C12.20–2015, Feb. 17, 2017.



Yanli Liu received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from Tianjin University, Tianjin, China, in 2007, 2009, 2014 respectively, where she is currently working as a Professor with School of Electrical and Information Engineering, Tianjin University. Her research interests include smart grid architecture, smart grid situational awareness, resilience of power cyber-physical system, and the application of artificial intelligence and data analytics in power systems.



Peng Wang received the B.Eng. degree from North China Electric Power University, Beijing, China, in 2017 and the M.Eng. degree from Tianjin University, Tianjin, China, in 2022. He joined State Grid Chongqing Electric Power Company in 2022. His research interests include distribution system topology identification, and big data analysis of distribution system.