Multisource Data Fusion Outage Location in Distribution Systems via Probabilistic Graphical Models

Yuxuan Yuan, *Member, IEEE*, Kaveh Dehghanpour, Zhaoyu Wang, *Senior Member, IEEE*, and Fankun Bu, *Graduate Student Member, IEEE*

***Abstract*—Efficient outage location is critical to enhancing the resilience of power distribution systems. However, accurate outage location requires combining massive *evidence* received from diverse data sources, including smart meter (SM) last gasp signals, customer trouble calls, social media messages, weather data, vegetation information, and physical parameters of the network. This is a computationally complex task due to the high dimensionality of data in distribution grids. In this paper, we propose a multi-source data fusion approach to locate outage events in partially observable distribution systems using Bayesian networks (BNs). A novel aspect of the proposed approach is that it takes multi-source evidence and the complex structure of distribution systems into account using a probabilistic graphical method. Our method can radically reduce the computational complexity of outage location inference in high-dimensional spaces. The graphical structure of the proposed BN is established based on the network's topology and the causal relationship between random variables, such as the states of branches/customers and evidence. Utilizing this graphical model, accurate outage locations are obtained by leveraging a Gibbs sampling (GS) method, to infer the probabilities of de-energization for all branches. Compared with commonly-used exact inference methods that have exponential complexity in the size of the BN, GS quantifies the target conditional probability distributions in a timely manner. A case study of several real-world distribution systems is presented to validate the proposed method.**

***Index Terms*—Approximate inference, Bayesian networks, data fusion, outage location, partially observable distribution system.**

I. INTRODUCTION

FREQUENT power outages are becoming a critical issue in the U.S. In 2018, the Department of Energy estimates that outages are costing the U.S. economy $150 billion annually [1]. 1.9 million customers in Midwest were affected by 1.4 million outages between August 10 and 13, 2020 [2]. Outage detection in distribution grids is an immediate and indispensable task after service disruptions, without which utilities cannot obtain needed situational awareness for initiating repair and restoration. This suggests an urgent need of efficient approaches to shorten the time of lateral-level outage location. Traditionally, outage location inference has been done based on manual outage mapping, which in addition to voltage and current components measured only at the substations, has mainly depended on customers' trouble calls. However, trouble calls alone are not a reliable source for outage location inference. It is estimated that only one-third of customers report the events in the first hour of outages, which might prolong the location determination process [3]. Also, customers might contact utilities due to temporary and individual problems rather than system-level outage events, which can mislead the location process and result in additional truck rolls to verify power outages.

One way of avoiding these problems is to rely on advanced metering infrastructure (AMI)-based techniques, which can send outage notifications at the grid-edge by leveraging the bidirectional communication function of smart meters (SMs). Researchers have dedicated great efforts to this topic. In [4], a hierarchical generative model is proposed that employs SM error count measurements to detect anomalies. In [5], a multi-label support vector machine model is developed that utilizes the state of customers' SMs to identify states of distribution lines. In [6], a two-stage method is presented to detect non-technical losses and outage events using real-time consumption data from SMs. In [7], a framework that combines the use of optimally deployed power flow sensors and load forecasts is proposed to detect outage events. In [8], a hypothesis testing-based outage location method is developed that combines the power flow measurements and SM-based load forecasts of the nodes. In [9], by using data from SMs and fault indicators, a multiple-hypothesis method with an extended protection tree is presented to detect a fault and identify the activated protective devices. The main challenge is that most AMI-based methods require *full observability* for distribution grids, i.e., SM installation for all customers. This assumption is not necessarily applicable to practical distribution systems, mostly due to utilities' budgetary limitations. To perform outage detection in partially observable systems, we have proposed a generative adversarial network (GAN)-based method to efficiently identify outage region [10]. Although this method is guaranteed to

Manuscript received May 8, 2021; revised September 6, 2021 and November 5, 2021; accepted November 14, 2021. This work was supported in part by the National Science Foundation under Grant EPCN 2042314, and in part by Advanced Grid Modeling Program at the U.S. Department of Energy Office of Electricity under Grant DE-OE0000875. Paper no. TSG-00724-2021. *(Corresponding author: Zhaoyu Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011 USA (e-mail: yuanyx@iastate.edu; wzy@iastate.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2021.3128752.

Digital Object Identifier 10.1109/TSG.2021.3128752

TABLE I
AVAILABLE LITERATURE ON DATA-DRIVEN OUTAGE DETECTION IN DISTRIBUTION SYSTEMS

Reference	Approach	Data source	Pros and Cons
[4]	Hierarchical generative model	Smart meter data	(+) Using hierarchical structure of the network and multivariate counts data, (-) Ignore interdependence between data sources and branches/customer status, accuracy decline for poor observable systems
[5]	Support vector machine		(+) Fast and accurate, (-) Fully observable system assumption
[6]	Fuzzy petri network		(+) Using real-time consumption data from smart meters, (-) Fully observable system assumption
[7]	Maximum a-posteriori method		(+) Optimal line flow sensor placement with load forecasts, (-) Additional cost
[8]	Hypothesis testing approach		(+) Combining power flow measurements and smart meter-based load forecasts to handle poor observability, (-) Lossless system assumption, fixed branch failure probability assumption
[9]	Multiple-hypothesis method		(+) Robustness for missing outage reports and fault indicators, (-) Assuming most two concurrent events can occur in a scenario, full observable system assumption
[10]	GAN-based method		(+) Capturing maximum amount of information on outage location from smart meter measurements, (-) Zone-based outage location
[11]	Polling method	Non-smart meter data	(+) Integration the operation of SCADA and smart meters, (-) Fully observable system assumption
[12]	Distributed approach		(+) Following a distributed manner to address scalability, (-) Requiring sensor (both power flow and smart meter) measurements and nodal load forecast statistics
[13]	Ensemble learning approach		(+) Using public weather information data to handle poor observability, (-) System-level outage analysis
[14]	Natural Language Processing approach		(+) Identifying outage-related tweets to handle poor observability, (-) System-level outage analysis, accuracy decline for rural systems
[15]	Mixed-integer linear program		(+) Simultaneously estimating the operation topology and outage sections, (-) Requiring line flow measurements and forecasted load data
[16]	Multi-layer perception neural network		(+) Using social sensors to handle poor observability, (-) System-level outage analysis, accuracy decline for rural systems
[17]	Dynamic-programming-based method		(+) Optimal line and nodal sensor placement for outage detection, (-) Additional cost, specific assumption for nodal sensors
[18]	State estimation-based method		(+) Well-developed method (-) Requiring data redundancy or high-confidence pseudo-measurement

capture the maximum amount of information on outage location, it does not provide granular outage location estimation at the branch level due to the limitations of the single data source. This issue is further exacerbated considering that SM signal communication to the utilities' data centers can fail due to hardware/software malfunctions and tampering [4].

Rather than using SM data, an alternative solution is to utilize other grid-independent data sources to identify outage events in real-time. In [11], an AMI-based polling method is proposed to enhance outage detection. In [12], a distributed outage detection algorithm is proposed with the primary objective of addressing scalability and communication bottleneck concerns. In [13], weather information data is used to detect outages in overhead distribution systems employing an ensemble learning approach. In [14], a data-driven outage identification approach is proposed that extracts textural and spatial information from social media. In [15], a mixed-integer linear program (MILP) is formulated to identify the topology under both outage and normal operating conditions using line flow measurements, forecasted load data, and ping measurements from a limited set of SMs. In [16], a modified approach of Kleinberg's burst detection algorithm is proposed to ensure the prompt detection of power outages. In [17], a dynamic programming-based minimum cost sensor placement solution is proposed for outage detection in distribution systems. In [18], the classical distribution system state estimation tool is extended to infer the status of switches. Nonetheless, the considerable uncertainty of these data sources can lead to erroneous outage location and additional costs for utilities. For example, only a part of SM last gasp signals can be delivered to the utility's data center due to hardware and software issues. Thus, to handle the limitations and uncertainties of individual data sources, this paper proposes a multi-source data fusion strategy to combine outage-related information from diverse sources for accurate outage location. A summary of the literature is shown in Table I.

One fundamental challenge in multi-source outage location is the computational complexity of the problem: first, outage location inference is the process of computing the probabilities of topology candidates after disrupting events by leveraging available information received by utilities. Estimating these probability values requires obtaining the *joint probability distribution function (PDF)* of the unknown state variables and the evidence, which is a high-dimensional mathematical object. Considering that outage data sources and branches/customer status are interdependent, directly quantifying this joint distribution requires enumerating probabilities of all possible combinations of variables, which is computationally infeasible in actual distribution systems. In addition, outage data sources have heterogeneous characteristics such as accuracy levels and reporting rates. Further, they may provide inconsistent and contrary information. How to integrate these data sources is a challenge. In [19], a probabilistic method is proposed for fault location by combining the measurements from digital relays at substations, intelligent electric devices along primary feeders, SCADA sensors in the feeder circuit, and smart meters. Statistics of historical fault location data are used to estimate fault location errors with probability in real time. The difficultly we face in this work, is to effectively integrate data from non-metered data sources (i.e., trouble calls, social media messages, and weather data), which makes the construction of a data fusion outage location framework challenging.

To address these challenges and the shortcomings of the previous works in the literature, a multi-source data fusion

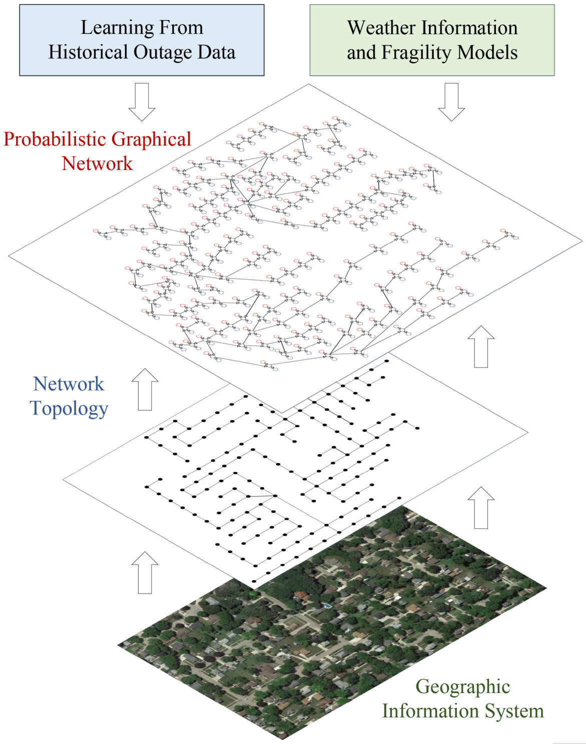


Fig. 1. Graphical approach towards outage location inference.

method is presented to identify and locate the lateral-level outage events in partially observable distribution systems. To achieve this, we have adopted a *probabilistic graphical modeling* approach towards data fusion to reduce the computational complexity of representing high-dimensional joint PDF of the system. The basic idea of this methodology is to use a graph-based representation as the foundation for encoding the joint distribution. Specifically, we first investigate statistical relationships among outage data sources and branches/customer status to build a Bayesian network (BN) for each distribution feeder. System topology in normal operations and context data, such as weather data and vegetation information, from geographic information system are used to design the architecture of the BN, as shown in Fig. 1. The graph parameters are learned empirically from historical outage data. It should be noted that the proposed method does not consider information of distributed energy sources. The rationale behind this is that most customer-level rooftop photovoltaics are integrated into distribution systems at behind-the-meter. Also, use of customer-level batteries in distribution systems has not become prevalent, which hinders utilities from using distributed energy data to detect power outages. By utilizing the proposed BN-based method, the high-dimensional joint PDF of the system is decomposed into a set of more manageable probabilistic *factors*. Then, the conditional PDF of the state of network branches and the connectivity of customer switches can be inferred by solving a probabilistic inference over the BN given the observed evidence in real time. This inference task is solved by leveraging a Gibbs sampling (GS) method. As a Markov chain Monte Carlo (MCMC)-based algorithm, GS can provide a full characterization of the distribution of unknown variables by generating a sequence of samples. We have used multiple real-world distribution systems from our utility partners to validate the performance of the proposed method. The main contributions of this paper can be summarized as follows.

- A probabilistic graphical model-based approach is proposed to seamlessly integrate heterogeneous outage-related data sources. The statistics of historical outage data are used to explicitly model the uncertainties of different data sources by graph parameterization. By utilizing this method, different data sources can complement each other to increase the amount of outage information, thus addressing low smart device coverage or customer report rates in actual grids.
- Multiple conditional independencies are explored to simplify the probabilistic graphical modeling. Meanwhile, a fragility model is integrated with the graph to formulate the conditional independence between the branch state and context data. These strategies can reduce the overfitting risk in the graph parameterization caused by outage data scarcity.
- An MCMC-based method is utilized to simplify the multi-dimensional summation in the outage location inference, which leads to an exponential reduction in detection and location time. This method can provide a good representation of a PDF by leveraging random variable instantiations, without knowing all the distribution's mathematical properties. The proposed technology determines the outage location by estimating the states of all the branches and customers.

The rest of this paper is constructed as follows. In Section II, the statement of the outage location problem is described. Section III presents the proposed BN-based data fusion model, along with structure selection and parameter learning schemes. An MCMC approximate inference algorithm is given in Section IV. The numerical results are analyzed in Section V. Section VI concludes the paper with major findings.

II. Outage Location Problem Statement

Considering that outage events cause topological changes in the grid, outage location is the process of inferring the probabilities of post-event operational topology candidates. In general, the accuracy of outage location depends on the completeness of outage information. Compared to traditional outage detection using only customer calls, combining different outage-related information, including SM last gasp signals, customer trouble calls, social media messages, wind speed, vegetation information, and physical parameters of the grid will greatly improve the accuracy and speed of outage detection. Different data sources can complement each other to increase the amount of outage information, thus addressing low SM coverage or customer report rates. It should be noted that this combination means integrating data from diverse sources as well as different customers. Hence, the proposed method aims to take full advantage of all available data in actual grids without the need to install additional metering devices for accurate outage detection and location.

This ensures the practicability of the proposed method for real-world applications. Specifically, SM last gasp signals and customer trouble calls are generally available in the distribution systems [4]–[6]. As demonstrated concretely in [14], most customers are already actively engaged in social media such as Facebook and Twitter in this information age. By applying suitable natural language processing methods, social data can be converted into binary outage evidence, similar to customer trouble calls and last gasp signals. The rationale behind the use of wind speed and vegetation information is that 87% of major power outages happen because trees are blown into power lines, or poles are destroyed by high intense winds [5]. To estimate the impact of these information, physical grid parameters, including the number of conductor wires and distribution poles, are necessary.

These data sources can be easily obtained after a power outage has occurred. Specifically, SM will automatically send the last gasp signal to the head-end system of the AMI after power disruptions. Trouble calls and social media messages are reported by customer's phones and Twitter. Wind speed and the physical parameters of the grid can be found from neighboring land-based station and grid model, respectively. Note that the proposed method does not have specific requirements for the range of wind speeds. Our method follows the line of fragility analysis using 3-s gust wind speed and grid physical parameters to calculate the probability of failure of the individual branch when the neighboring upper-stream branch is energized [20]. This fragility analysis is applicable to both normal and extreme weather. Regarding the vegetation evidence, the tree coverage data adjacent to power lines is utilized. Utilities can add or remove data sources in probabilistic graphical model according to their situations. For example, for systems lacking extreme weather events, vegetation information and wind speed can be removed to reduce the complexity of the model, as these two data sources may not have a significant impact on outage detection and location during normal weather. After data collection, last gasp signals, customer trouble calls, wind speed, vegetation information, and physical parameters can be directly transformed into outage evidence as input to the proposed model. For social media messages, a natural language processing tool is required to extract outage-related words, as proposed in our previous work [14]. Then, social media messages are converted into binary outage evidence, similar to customer trouble calls and last gasp signals. Note that all formulations in the paper are implicitly phase-based, meaning that separate equations should be written and applied to each phase of the distribution system to consider the multi-phase and unbalanced nature of the grid into account. With this in mind, and for the sake of clarity and tractability, phase-related notations/signs are dropped from all equations.

Regarding notation, vectors/matrices are represented with bold letters. Uppercase letters refer to random and evidence variables. Lowercase letters are the assignment of values to the related variables. For example, for a random variable X, let x denotes its realization. Given the multi-source evidence, $\boldsymbol{E}$, the inference process is mathematically formulated using the *Bayes estimator* [21], where the conditional PDF of network topology, Y, given the set of evidence is represented as $P(Y = y|\boldsymbol{E} = \boldsymbol{e})$ and calculated in terms of the joint distribution of Y and $\boldsymbol{E}$, denoted by $P(Y = y, \boldsymbol{E} = \boldsymbol{e})$. The most probable candidate topology, which also determines the location of the outage event, is obtained by maximizing this conditional PDF, as:

$$y^* = \underset{y}{\operatorname{argmax}}\, P(Y = y|\boldsymbol{E} = \boldsymbol{e}) = \frac{P_{Y,\boldsymbol{E}}(y, \boldsymbol{e})}{P_{\boldsymbol{E}}(\boldsymbol{e})} \tag{1}$$

where, y^* is the most likely network topology after the outage. Y is a multinomial variable which is represented in terms of the states of primary network branches ($\boldsymbol{D}$) and the connection of customer switches ($\boldsymbol{C}$), as $Y = \{\boldsymbol{D}, \boldsymbol{C}\}$. Here, $\boldsymbol{D} = [D_1, \ldots, D_k]$, where k is the number of branches in the feeder and D_i is a binary variable representing the connectivity state for the i'th branch in the feeder: $D_i = 0$ means that the branch is *energized*. In other words, there is an uninterrupted path between the branch and the substation. $D_i = 1$ indicates that the branch is *de-energized*. Similarly, $\boldsymbol{C} = [\boldsymbol{C_1}, \ldots, \boldsymbol{C_k}]$, with $\boldsymbol{C_i}$ representing the set of connection states for all the customers that are supplied by the i'th branch. Hence, $\boldsymbol{C_i} = [C_i^1, \ldots, C_i^{z_i}]$, where z_i is the total number of customers that are connected to the i'th branch, and C_i^j is the state of the j'th customer: $C_i^j = 0$ means that the customer is energized, and $C_i^j = 1$ implies that the customer is de-energized. Note that the pre-outage topology is determined by assigning 0 to all the state variables (i.e., all branches are energized and customers are energized). Thus, $P(Y = y|\boldsymbol{E} = \boldsymbol{e})$ in (1) can be rewritten in terms of the joint PDF of the newly-defined variables, $P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})$, as follows [22]:

$$P(Y = y|\boldsymbol{E} = \boldsymbol{e}) = P_{\boldsymbol{D},\boldsymbol{C}|\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}|\boldsymbol{e}) = \frac{P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})}{P_{\boldsymbol{E}}(\boldsymbol{e})}. \tag{2}$$

Using (2), the maximization over topology candidates can be conveniently transformed into finding the best values for the individual branch/customer states belonging to $\{\boldsymbol{D}, \boldsymbol{C}\}$ using their conditional PDFs, $P_{D_i|\boldsymbol{E}}(d_i|\boldsymbol{e})$ and $P_{C_i^j|\boldsymbol{E}}(c_i^j|\boldsymbol{e})$. These conditional PDFs are obtained $\forall i, j$ using a marginalization process over the joint PDF, as follows [23]:

$$P_{D_i|\boldsymbol{E}}(d_i|\boldsymbol{e}) = \sum_{\{\boldsymbol{d},\boldsymbol{c}\}\setminus d_i} P_{\boldsymbol{D},\boldsymbol{C}|\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}|\boldsymbol{e}) = \sum_{\{\boldsymbol{d},\boldsymbol{c}\}\setminus d_i} \frac{P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})}{P_{\boldsymbol{E}}(\boldsymbol{e})} \tag{3}$$

$$P_{C_i^j|\boldsymbol{E}}\left(c_i^j|\boldsymbol{e}\right) = \sum_{\{\boldsymbol{d},\boldsymbol{c}\}\setminus c_i^j} P_{\boldsymbol{D},\boldsymbol{C}|\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}|\boldsymbol{e}) = \sum_{\{\boldsymbol{d},\boldsymbol{c}\}\setminus c_i^j} \frac{P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})}{P_{\boldsymbol{E}}(\boldsymbol{e})} \tag{4}$$

where, $A \setminus B$ represents all the elements in A that specifically are not in the set B.

In general, the goal of the proposed work is to solve (3)-(4) in real time. However, considering the complexity of distribution grids, obtaining the explicit representation of the joint PDF, $P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})$, is unmanageable for two reasons: (I) a complete description of $P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})$ induces an exponential complexity in the order of $2^r - 1$, where r is the total cardinality of all the unknown variables, $r = |\boldsymbol{D}| + |\boldsymbol{C}|$. Hence, modeling this joint PDF using brute-force search over all possible combinations of branch/customer states is computationally infeasible for large-scale distribution systems. (II) Due

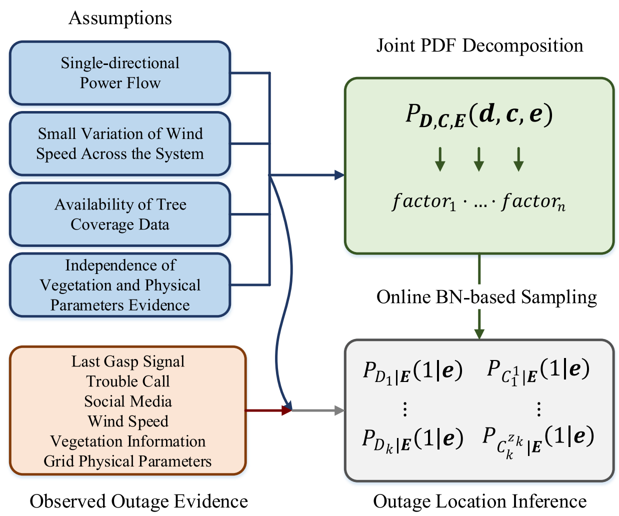


Fig. 2. Assumptions of the proposed method.

to the outage data scarcity in distribution grids, it is impossible to acquire enough historical data to robustly estimate the massive number of parameters of this joint distribution. One solution is to use *naive classification* by assuming full independence among all evidence and unknown state variables [23]. However, this assumption is not applicable to practical distribution systems and may lead to severe misclassification due to overfitting.

III. BN-Based Data Fusion Model

To counter computational complexity and overfitting in the outage location inference, we propose a BN-based method. A unique feature of our method is a seamless integration of heterogeneous data sources by leveraging conditional independencies inherent in the grid and data. These conditional independencies enable a scalable and compact graphical representation of different data and enhance outage inference efficiency. More precisely, by using the proposed method, the joint PDF $P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d},\boldsymbol{c},\boldsymbol{e})$ is decomposed into a set of *factors* with significantly smaller size. Using this computationally efficient BN-based approach, we can infer the conditional PDF of the state of each primary branch and the customer switch given outage-related evidence from various data sources in real time, shown in (3)-(4), to rapidly identify the location of lateral-level outage events. Given the unbalanced nature of distribution networks, the proposed algorithm is applied to each phase separately. Specifically, for three-phase unbalanced systems, we build three different Bayesian networks based on the information regarding which customers are connected to which service transformers or phases. In rare systems without this knowledge, the previous customer grouping methods can be applied before establishing the graphical models [24]–[26].

As shown in Fig. 2, this work is based on several assumptions, which are listed below.

- The proposed method only considers distribution networks with single-directional power flows. Otherwise, the conditional independencies regarding the state of the upstream and downstream branches will become ambiguous.
- The vegetation data adjacent to power lines is assumed to be available for utilities. In rare cases without such records, the tree coverage data in the census tract including the power lines can be used [27].
- All the branches are assumed to be subjected to the maximum wind speed at the middle point of the system in this work. The rationale behind this is that the variation of wind speed across the distribution system is minimal. This assumption is consistent with the previous fragility analysis [20].
- The vegetation and physical parameter evidence for each specific branch is assumed to be independent of those in other branches. Relaxation of this assumption will be further investigated in future works.

A. Factorization of the Joint PDF and BN Representation

The main idea of a BN-based representation is to use conditional independencies, encoded in a graph structure, to compactly break down high-dimensional joint PDFs with a set of factors. Here, a factor refers to a low-dimensional and more manageable conditional PDF that is determined by two components: a *child* variable, such as D_i and a number of *parent* variables denoted by $Pa(\cdot)$, such as $Pa(D_i)$. Parent variables represent the direct causal sources of influence for a child variable. In other words, each child is a stochastic function of its parents [23]. Thus, if the values of the parents are known, then the child variable becomes conditionally independent of random variables that do not directly influence it in a causal manner. It can be shown that by using chain rule over these conditional independencies, defined by parent-child relationships, the joint PDF of a set of random variables can be simplified as the multiplication of the identified factors [23]. In the outage location problem, this factorization leads to the following data fusion representation for the joint PDF:

$$
\begin{aligned}
P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d},\boldsymbol{c},\boldsymbol{e}) = &\left(\prod_{i=1}^{k} P_{D_i|Pa(D_i)}(d_i|Pa(d_i))\right)\\
&\times\left(\prod_{i=1}^{k}\prod_{j=1}^{z_i} P_{C_i^j|Pa\left(C_i^j\right)}\left(c_i^j|Pa\left(c_i^j\right)\right)\right)\\
&\times\left(\prod_{i=1}^{u} P_{E_{i,j}^h|Pa\left(E_{i,j}^h\right)}\left(e_{i,j}^h|Pa\left(e_{i,j}^h\right)\right)\right)\\
&\times\left(\prod_{i=1}^{u} P_{E_{i,j}^m|Pa\left(E_{i,j}^m\right)}\left(e_{i,j}^m|Pa\left(e_{i,j}^m\right)\right)\right) \quad (5)
\end{aligned}
$$

where, $u = |\boldsymbol{E}|$, and the factors are $P_{D_i|Pa(D_i)}(d_i|Pa(d_i))$, $P_{C_i^j|Pa(C_i^j)}(c_i^j|Pa(c_i^j))$, $P_{E_{i,j}^h|Pa(E_{i,j}^h)}(e_{i,j}^h|Pa(e_{i,j}^h))$, and $P_{E_{i,j}^m|Pa(E_{i,j}^m)}(e_{i,j}^m|Pa(e_{i,j}^m))$, $\forall i,j$. $E_{i,j}^h$ denotes the human-based evidence from the customer-side, including trouble calls and social media messages; $E_{i,j}^m$ represents meter-based evidence from customer-side, such as smart meter last gasp signals. When an outage occurs, utilities can determine the values of $E_{i,j}^h$ and $E_{i,j}^m$, according to the information received. For example, if one customer calls to report a power outage, this customer's human evidence is identified as 1; otherwise, it should be 0. Compared with the original model in (2) that requires $2^r - 1$ independent parameters, the new representation in (5) only needs $\sum_{i=1}^{k} 2^{|Pa(D_i)|} + \sum_{i=1}^{k}\sum_{j=1}^{z_i} 2^{|Pa(C_i^j)|} + \sum_{i=1}^{n} 2^{|Pa(E_{i,j}^h)|} + \sum_{i=1}^{n} 2^{|Pa(E_{i,j}^m)|}$ parameters. It can be observed that the number of parameters in the new representation is a function of size of parents for each variable. Considering that the number of variables' parents is typically small, the new representation achieves a radical complexity reduction in outage location inference.

As a directed acyclic graph, BN offers a convenient way of representing the factorization (5). Accordingly, the random variables, $\{\boldsymbol{D}, \boldsymbol{C}, \boldsymbol{E}\}$, are represented as the *vertices* of the BN. Using the identified factors in (5), the vertices of the BN are connected by drawing *directed edges* that start from parent vertices and end in child vertices. Specifically, BN encodes the conditional independencies defined by the factors as follows: any vertex, X, is conditionally independent of its *non-descendant* vertices in the graph, $Nd(X)$, if the values of its parents are known. This is symbolically denoted by $(X \perp Nd(X)|Pa(X))$ [28]. $Nd(X)$ is the set of the vertices of the BN, excluding parents of X, to which no directed path exists originating from X. $A \perp B$ means that A and B are marginally independent.

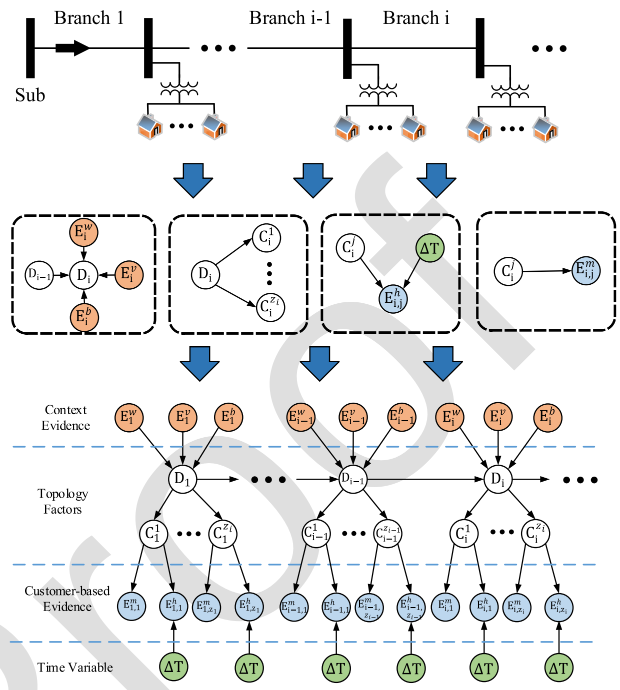


Fig. 3. BN of a typical radial distribution system.

B. BN Structure Development and Parameterization

Developing a BN requires discovering the structure of the graph and the parameters of the conditional PDFs. To do this, a knowledge discovery-based method is utilized in this paper. An inherent feature of radial grids is their tree-like structure, resulting in a unique one-directional path between all nodes. If this path is disrupted at any branch, then the states of all downstream branches can be inferred as de-energized without a need for further search. Based on this feature, the parent-child variables of each factor in (5) can be described as follows.

(1) Factor $P_{D_i|Pa(D_i)}(d_i|Pa(d_i))$ represents the conditional independencies of the form $D_i \perp Nd(D_i)|Pa(D_i)$. The parents of branch state variable are selected as $Pa(D_i) = \{D_{i-1}, E_i^w, E_i^v, E_i^b\}$, as shown in Fig. 3. Here, D_{i-1} is the state of the neighboring upper-stream branch. $\{E_i^w, E_i^v, E_i^b\}$ are the evidence for the i'th branch. Specifically, E_i^w denotes 3-s gust wind speed collected by local land-based station. The value of E_i^w is determined by the maximum wind speed at the middle point of the system. E_i^v refers to vegetation information, which contains vegetation constants and diameters of the trees adjacent to each branch. E_i^b represents the i'th branch's physical parameters, including the length of conductors and the number of poles of each branch. Based on this parent selection scheme for branch state variables, $Nd(D_i)$ includes all the variables that are not downstream of the i'th branch in the feeder (see Fig. 3). To show the direct causal influences of these four variables on D_i, two cases are described: $D_{i-1} = 1$ and $D_{i-1} = 0$.

In the first case, when the parent branch is de-energized, then $D_i = 1$ with *probability 1*. Consequently, all variables on the path from the substation to D_{i-1}, represented with $\{D_1, \ldots, D_{i-2}\}$, are conditionally independent from $\{D_i\}$ given $D_{i-1} = 1$. The intuition behind this is that in radial networks there is only one unique path between the substation and each branch; if this path is interrupted at any arbitrary point in $\{D_1, \ldots, D_{i-2}\}$, we can automatically conclude $D_{i-1} = 1$ regardless of the location of outage in the path. Hence, considering the binary nature of variable D_i, the conditional PDF, $P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}(d_i|1, e_i^w, e_i^v, e_i^b)$, can be formulated as:

$$P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}\left(1|1, e_i^w, e_i^v, e_i^b\right) = 1$$
$$P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}\left(0|1, e_i^w, e_i^v, e_i^b\right) = 0. \tag{6}$$

In the second case, if the neighboring upper-stream branch is energized, then all upstream branches of the i'th branch are also energized with probability 1, and have not been impacted by outage, $\{D_1 = 0, \ldots, D_{i-2} = 0\}$. In this case, $D_i = 1$ will only occur when this branch is damaged. As demonstrated concretely in [27], the majority of branch damage is caused by tree contacts to power lines and broken poles due to high wind speed. Thus, three context variables E_i^w, E_i^v and E_i^b are serve as causal evidence for the i'th branch state to estimate the probability of outage at the i'th branch. The conditional PDF, $P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}(d_i|0, e_i^w, e_i^v, e_i^b)$, can be formulated as a Bernoulli distribution as follows:

$$P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}\left(d_i|0, e_i^w, e_i^v, e_i^b\right) = \begin{cases} P_l^i & \text{for } d_i = 1 \\ 1 - P_l^i & \text{for } d_i = 0 \end{cases} \tag{7}$$

where, the probability of failure for branch i, denoted as P_l^i, is a function of e_i^w, e_i^v, and e_i^b. To formulate this function, a fragility model is leveraged. Basically, the fragility model is a series model with the fragility analysis of each pole and conductor within the branch:

$$P_l^i = 1 - \prod_{d=1}^{L}\left(1 - \phi\left(\frac{\ln\left(\frac{e_i^w}{\chi}\right)}{\xi}\right)\right)\prod_{f=1}^{K}\left(1 - P_f\left(e_i^w, e_i^v\right)\right) \quad (8)$$

where, L is the number of distribution poles used for supporting branch i, K is the number of conductor wires between two neighboring poles at the i'th branch, ϕ is the standard normal probability integral, χ is the median of the fragility function, ξ is the logarithmic standard deviation of intensity measure, and $P_f(e_i^w, e_i^v)$ represents the failure probability for conductor f of branch i which is modeled as follows:

$$P_f\left(e_i^w, e_i^v\right) = (1 - p_u)\max\left\{\min\left\{\frac{F_{wind,f}\left(e_i^w\right)}{F_{no,f}\left(e_i^w\right)}, 1\right\}, \alpha \cdot P_t\left(e_i^v\right)\right\} \quad (9)$$

where, p_u is the probability of conductor f being underground, $F_{wind,f}(e_i^w)$ represents the wind force loading on the conductor and $F_{no,f}(e_i^w)$ demonstrates the maximum perpendicular force of the conductor wire determined as shown in [20]. α describes the average tree-induced damage probability of overhead conductor, and $P_t(e_i^v)$ is the fallen tree-induced failure probability of conductor f computed as in [27]. Hence, for the case $D_{i-1} = 0$, equations (8) and (9) are utilized to estimate the probability of outage for branch i given the values of the context variables E_i^w, E_i^v, and E_i^b. To summarize, the conditional PDFs given in equations (6) and (7) fully determine the factors of the form $P_{D_i|Pa(D_i)}(d_i|Pa(d_i))$.

(2) Factor $P_{C_i^j|Pa(C_i^j)}(c_i^j|Pa(c_i^j))$ represents the conditional PDF of the status of customer j given parent variables. The parent of customer state variable is selected as $Pa(C_i^j) = \{D_i\}$ (see Fig. 3). Here, D_i is the state of the immediate upper-stream branch that supplies the j'th customer. To show the casual relationship between C_i^j and D_i, two cases are considered: $D_i = 1$ and $D_i = 0$.

In the first case, if the primary branch is de-energized, the probability of $C_i^j = 1$ is 1 due to the radial structure of the feeder. Utilizing this deterministic relationship, $P_{C_i^j|D_i}(c_i^j|d_i)$ can be written as follows:

$$\begin{aligned} P_{C_i^j|D_i}(1|1) &= 1 \\ P_{C_i^j|D_i}(0|1) &= 0. \end{aligned} \quad (10)$$

In the second case, if the primary branch is energized, then the path between the substation and the i'th branch is active. Hence, customer outage, $C_i^j = 1$, can only be caused by overloading/faults at the customer-side occurring with probability π_2. This case is represented using a Bernoulli distribution adopted from statistical outage information [29]:

$$P_{C_i^j|D_i}\left(c_i^j|0\right) = \begin{cases} \pi_2 & \text{for } c_i^j = 1 \\ 1 - \pi_2 & \text{for } c_i^j = 0. \end{cases} \quad (11)$$

To account for the uncertainty of parameter π_2, a beta distribution is defined with user-defined hyper-parameters α_2 and β_2:

$$\pi_2 \sim Beta(\alpha_2, \beta_2) = \gamma_2 \pi_2^{\alpha_2 - 1}(1 - \pi_2)^{\beta_2 - 1} \quad (12)$$

where, γ_2 is a normalizing constant and defined as $\gamma_2 = \Gamma(\alpha_2 + \beta_2)$ with $\Gamma = \int_0^\infty t^{x-1}e^{-t}dt$ [23].

(3) Factor $P_{E_{i,j}^h|Pa(E_{i,j}^h)}(e_{i,j}^h|Pa(e_{i,j}^h))$ represents the conditional independencies $E_{i,j}^h \perp Nd(E_{i,j}^h)|Pa(E_{i,j}^h)$. The parents of human-based evidence, $E_{i,j}^h$, are selected as $Pa(E_{i,j}^h) = \{C_i^j, \Delta T\}$, as shown in Fig. 3. ΔT refers to the time elapsed after the outage occurrence. More precisely, ΔT embodies the time period that utilities need to wait before outage reports are issued [30]. It is clear that there is a trade-off between the amount of human-based evidence and waiting time of outage location inference. For example, when feeder observability is extremely low, utilities may increase ΔT to receive more human-based evidence for outage location inference. Within the ΔT period, the time at which the human-based evidence is received, T, after outage occurrence at time, T_0, is distributed according to an exponential distribution as shown in [31]:

$$f\left(T = t|T_0 = t_0, C_i^j = 1\right) = \lambda_1 e^{-\lambda_1(t - t_0)}. \quad (13)$$

Thus, given Δt, the probability of $P(E_{i,j}^h = 1|C_i^j = 1, T - T_0 \leq \Delta t)$ can be calculated as:

$$\begin{aligned} &P(E_{i,j}^h = 1|C_i^j = 1, T - T_0 \leq \Delta t) \\ &\quad = \int_0^{\Delta t}\lambda_1 e^{-\lambda_1 t'}\mathrm{d}t' = -e^{-\lambda_1\Delta t} + 1. \end{aligned} \quad (14)$$

Hence, the factor $P_{E_{i,j}^h|C_i^j,\Delta T}(e_{i,j}^h|c_i^j, \Delta t)$ is obtained as follows:

$$P_{E_{i,j}^h|C_i^j,\Delta T}\left(e_{i,j}^h|c_i^j, \Delta t\right) = \begin{cases} -e^{-\lambda_1\Delta t} + 1 & \text{for } e_{i,j}^h = 1, c_i^j = 1 \\ e^{-\lambda_1\Delta t} & \text{for } e_{i,j}^h = 0, c_i^j = 1 \\ \pi_3 & \text{for } e_{i,j}^h = 1, c_i^j = 0 \\ 1 - \pi_3 & \text{for } e_{i,j}^h = 0, c_i^j = 0 \end{cases} \quad (15)$$

where, π_3 denotes a small user-defined value to take into account the possibility of false positives, such as illegitimate trouble call and social media data processing errors.

(4) Factor $P_{E_{i,j}^m|Pa(E_{i,j}^m)}(e_{i,j}^m|Pa(e_{i,j}^m))$ is the conditional independencies $E_{i,j}^m \perp Nd(E_{i,j}^m)|Pa(E_{i,j}^m)$. Compared to the human-based signals $E_{i,j}^h$, AMI-based notification mechanism will be delivered almost instantaneously to the utilities. Thus, the parent of meter-based evidence is selected as $Pa(E_{i,j}^m) = \{C_i^j\}$ (see Fig. 3). When the state of customer switch is known, $E_{i,j}^m$ becomes conditionally independent of the remaining variables, as encoded by the factor:

$$P_{E_{i,j}^m|C_i^j}\left(e_{i,j}^m|c_i^j\right) = \begin{cases} \pi_4 & \text{for } e_{i,j}^m = 1, c_i^j = 1 \\ 1 - \pi_4 & \text{for } e_{i,j}^m = 0, c_i^j = 1 \\ \pi_5 & \text{for } e_{i,j}^m = 1, c_i^j = 0 \\ 1 - \pi_5 & \text{for } e_{i,j}^m = 0, c_i^j = 0 \end{cases} \quad (16)$$

where, π_4 and π_5 represent the AMI communication reliability and the SM malfunction probability values, respectively. For concreteness, π_4 is the probability that the last gasp can be delivered to the utilities correctly for outage notification. π_5 is the probability that the SM loses power due to its own failure and sends a last gasp signal. In this work, the values of these

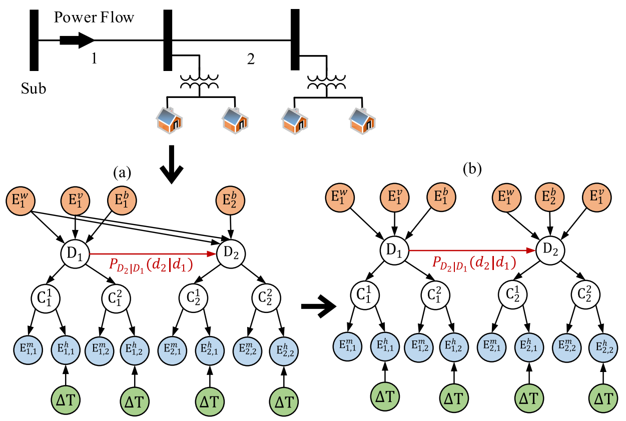


Fig. 4. 3-node lateral and matching BN graph.

two parameters are determined based on the historical outage reports. Considering the size of the historical data is limited, beta distributions are used to model the uncertainty of these two parameters as follows:

$$\pi_4 \sim Beta(\alpha_4, \beta_4) = \gamma_4 \pi_4^{\alpha_4 - 1}(1 - \pi_4)^{\beta_4 - 1}$$
$$\pi_5 \sim Beta(\alpha_5, \beta_5) = \gamma_5 \pi_5^{\alpha_5 - 1}(1 - \pi_5)^{\beta_5 - 1}. \quad (17)$$

To help the reader understand how a Bayesian network is built, an example is shown in Fig. 4. This toy system includes 3 nodes and 4 customers. First, since the state of each branch is directly impacted by weather, vegetation information, and physical parameters, $E^w_{1,1}$, $E^v_{1,1}$, and $E^b_{1,1}$ are modeled as parent nodes for D_1. Then, given the tree-like structure of the system, the state of the branch 1 serves as the immediate casual source of influence for the states of its immediate downstream branch and customers (i.e., D_2, C^1_1, C^2_1). When the state of the customer, C^1_1, is known, outage evidences from this customer become conditionally independent from D_1. Further, if the utility knows that C^1_1 is in outage, probabilities of receiving SM last gasp signals and trouble calls from that customer are uncorrelated. Hence, C^1_1 is modeled as parent node for $E^m_{1,1}$ and $E^h_{1,1}$ in the graph. This exemplary system can be treated a block cell for any radial feeder in general, which means that the proposed method can be generalized to any radial distribution system. Also, some high-level context evidence, including weather information and vegetation information, affect multiple neighboring branches in the same region, as shown in Fig. 4 (a). However, the size of the region is impacted by several factors (i.e., the geographic location of weather station and the grid infrastructure) and is hard to quantify and draw. Therefore, to avoid misunderstanding, two assumptions are utilized to build a more general BN graph, as shown in Fig. 4 (b). The details of the assumptions can be found at the beginning of Section III. In sum, the evidence from the branch-side (i.e., wind speed, vegetation information, and the physical parameters) is causal sources of branch states, which is formulated as a fragility model. When the branch state is observed, the branch-side evidence becomes independent from the states of the connected customers. In contrast, the evidence from the customer-side (i.e., human- and meter-based evidence) is independent from the rest of state and evidence variables, if the state of upstream customer is known, which is denoted as conditional independendency. Furthermore, if the utility knows that a customer is in an outage, the probabilities of receiving SM last gasp signals and human-based evidence will become uncorrelated. In this case, customer states are causal sources of the evidence. Thus, customer states are modeled as parent nodes for these data sources.

IV. BN-Based Outage Location Inference Using GS

The data fusion outage location process is transformed into a probabilistic inference over the graphical model. After construction and parameterization of the BN, $P_{\boldsymbol{D},\boldsymbol{C},\boldsymbol{E}}(\boldsymbol{d}, \boldsymbol{c}, \boldsymbol{e})$ has been simplified. However, solving (3)-(4) still requires calculating computationally expensive summation operations $P_{\boldsymbol{E}}(\boldsymbol{e})$ over all nodes of the graph simultaneously, which is not scalable for large-scale distribution grids [23]. To address this, a GS algorithm is used to perform the inference task over the BN [32].

A. GS Algorithm

GS is an MCMC-based approximate inference method,[1] which allows one to provide a good representation of a PDF by leveraging random variable instantiations, without knowing the distribution's mathematical properties [32]. The key advantage of this method is that it employs univariate conditional distributions for sampling, which eliminates the dependency on the dimension of the random variable space. Thus, compared to the commonly-used exact inference methods, such as variable elimination and clique trees, GS is insensitive to the size of BN [22]. This indicates that the GS method is especially beneficial for complex real-world applications.

When an outage occurs, the de-energization probabilities of branches/customers are inferred using the GS algorithm and the BN structure. To do this, first, all the outage evidence from the customer-side, $\{E^h_{1,1}, \ldots, E^h_{z_k,k}, E^m_{1,1}, \ldots, E^m_{z_k,k}\}$, is collected after ΔT has elapsed: if utilities receive trouble call/tweet or last gasp signal from the j'th customer at branch i, the corresponding evidence $E^h_{i,j}$ or $E^m_{i,j}$ is set to 1. In contrast, if the trouble call/tweet or last gasp signal is missing, the $E^h_{i,j}$ or $E^m_{i,j}$ is set to 0. Also, the branch-level evidence, $\{E^w_1, \ldots, E^w_k, E^v_1, \ldots, E^v_k, E^b_1, \ldots, E^b_k\}$, is set to the local wind speed, vegetation data, and i'th branch's physical parameters, respectively. After transferring these data to outage evidence, arbitrary initial samples are randomly assigned to all the unknown state variables $\{\boldsymbol{D}, \boldsymbol{C}\}$: $[D_1 = d^{(0)}_1, \ldots, D_k = d^{(0)}_k, C^1_1 = c^{1,(0)}_1, \ldots, C^{z_k,(0)}_k]$. Then, an arbitrary state variable is selected as the sampling starting point, e.g., D_i. At iteration $\tau + 1$ of GS, following the structure of the BN, the assigned samples to the parents and children of D_i are inserted into a local Bayesian estimator [22], as shown in (20), to approximate the conditional PDF of D_i given the latest

[1]MCMC is a subset of Monte Carlo methods. Unlike the common Monte Carlo methods that generate independent data samples from a specific distribution, MCMC methods generate samples where the next sample is dependent on the existing sample.

samples:

$$P_\Phi\left(d_i|\boldsymbol{d_{-i}}^{(\tau)}\right) = \frac{P_{D_i|Pa(D_i)}(d_i|Pa(d_i))P_{Ch(D_i)|PC(D_i)}(Ch(d_i)|PC(d_i)}{\sum_{d_i} P_{D_i|Pa(D_i)}(d_i|Pa(d_i))P_{Ch(D_i)|PC(D_i)}(Ch(d_i)|PC(d_i))} \tag{18}$$

where, $\boldsymbol{d_{-i}}^{(\tau)}$ is all the latest samples except for d_i, including values of evidence variables, and:

$$P_{D_i|Pa(D_i)}(d_i|Pa(d_i)) = P_{D_i|D_{i-1},E_i^w,E_i^v,E_i^b}\left(d_i|d_{i-1}^{(\tau)}, e_i^w, e_i^v, e_i^b\right) \tag{19}$$

$$P_{Ch(D_i)|PC(D_i)}(Ch(d_i)|PC(d_i)) = P_{D_{i+1}|D_i,E_i^w,E_i^v,E_i^b}\left(d_{i+1}^{(\tau)}|d_i, e_i^w, e_i^v, e_i^b\right)\prod_{j=1}^{z_i} P_{C_i^j|D_i}\left(c_i^{j,(\tau)}|d_i\right). \tag{20}$$

Hence, $P_\Phi(d_i|\boldsymbol{d_{-i}}^{(\tau)})$ can be directly calculated using the determined factors, (6)-(17), in Section III-B. Note that because $P_\Phi(d_i|\boldsymbol{d_{-i}}^{(\tau)})$ is a PDF over a single random variable given the samples assigned to all the others, this computation can be performed efficiently. Utilizing $P_\Phi(d_i|\boldsymbol{d_{-i}}^{(\tau)})$, a new sample $d_i \leftarrow d_i^{(\tau+1)}$ is drawn using the inverse transform method [23] to replace $d_i^{(\tau)}$. Then, the algorithm moves to a next non-evidence variable of BN to perform the local sampling process (see (20)). When all the unknown variables of the BN have been sampled once, one iteration of GS is complete. This process is able to propagate the information across the BN and combine the data from diverse sources to infer the location of outage efficiently. The sampling process is repeatedly applied until a sufficient number of random samples are generated for the unknown variables, $\{\boldsymbol{D}, \boldsymbol{C}\}$. It has been theoretically proved that the approximate PDFs, $P_\Phi(\cdot)$, are guaranteed to approach the target conditional PDFs, $P_{D_i|\boldsymbol{E}}(d_i|\boldsymbol{e})$ and $P_{C_i^j|\boldsymbol{E}}(c_i^j|\boldsymbol{e})$, defined in (3)-(4) [23]. Thus, $P_{D_i|\boldsymbol{E}}(d_i|\boldsymbol{e})$ and $P_{C_i^j|\boldsymbol{E}}(c_i^j|\boldsymbol{e})$ can be estimated by counting the samples generated by the GS algorithm. As an example, $P_{D_i|\boldsymbol{E}}(1|\boldsymbol{e})$ is estimated as follows:

$$P_{D_i|\boldsymbol{E}}(1|\boldsymbol{e}) \approx \frac{\sum_{\tau=0}^{M} d_i^\tau}{M} \tag{21}$$

where, M is the number of iterations. After the GS process, the most likely value of each branch/customer state is determined based on the obtained approximated conditional PDFs to solve (1). To achieve this, due to the binary nature of the state variables, a 0.5 threshold is used, e.g., $P_{D_i|\boldsymbol{E}}(1|\boldsymbol{e}) \leq 0.5$ indicates branch i is energized. After the connectivity states of all the branches/customers are inferred, the location of outage events are obtained by selecting the nearest de-energized branch to the substation. See Algorithm 1 for details.

Algorithm 1 Outage Location Inference Using GS

Require: : BN G; iteration number M; evidence $\boldsymbol{E}$;
Randomly generate i.i.d. samples $\boldsymbol{x}^{(0)} \leftarrow \{D_i = d_i^{(0)}, \ldots, C_i^j = c_i^{j,(0)}, \forall i, j\}$ from uniform distribution; $\boldsymbol{x}^{(0)} \leftarrow \boldsymbol{x}^{(0)} \cup \boldsymbol{E}$
for $\tau = 0, \ldots, M$ **do**
  for $i = 1, \ldots, |\boldsymbol{D}+\boldsymbol{C}|$ **do**
    Select one random variable $X_i \in \{\boldsymbol{D}, \boldsymbol{C}\}$
    $\boldsymbol{x_{-i}}^{(\tau)} \leftarrow \boldsymbol{x}^{(\tau)} - x_i^{(\tau)}$
    Obtain $Pa(X_i)$ and $Ch(X_i)$ from G
    $\frac{P_{X_i|Pa(X_i)}(x_i|Pa(x_i))P_{Ch(X_i)|X_i}(Ch(x_i)|x_i)}{\sum_{x_i} P_{X_i|Pa(X_i)}(x_i|Pa(x_i))P_{Ch(X_i)|X_i}(Ch(x_i)|x_i)} \rightarrow P_\Phi$
    Draw a new sample, $x_i^{(\tau+1)} \sim P_\Phi$
    $x_i^{(\tau+1)} \leftarrow x_i^{(\tau)}$
  end for
end for
Return sample vectors: $\boldsymbol{d_i} = \{d_i^{(0)}, \ldots, d_i^{(M)}\}$ and $\boldsymbol{c_i^j} = \{c_i^{j,(0)}, \ldots, c_i^{j,(M)}\}$, $\forall i, j$
$P_{D_i|\boldsymbol{E}}(1|\boldsymbol{e}) \leftarrow \frac{\sum_{\tau=0}^{M} d_i^{(\tau)}}{M}, \forall i$
$P_{C_i^j|\boldsymbol{E}}(1|\boldsymbol{e}) \leftarrow \frac{\sum_{\tau=0}^{M} c_i^{j,(\tau)}}{M}, \forall i, j$
If $P_{D_i|\boldsymbol{E}}(1|\boldsymbol{e}) \leq 0.5 \implies d_i = 1, \forall i$; if $P_{C_i^j|\boldsymbol{E}}(1|\boldsymbol{e}) \leq 0.5 \implies c_i^j = 1, \forall i, j$
Select the nearest de-energized branch as the outage location

---

B. GS Calibration Process

One challenge in GS is how to determine the number of iterations, M. In general, if the iterations have not proceeded long enough, the sampling may grossly misrepresent the target distributions, thus decreasing the inference accuracy. In contrast, if the value of M is large enough, the theory of MCMC guarantees that the stationary distribution of the samples generated using the GS algorithm [22]. However, such a strategy leads to high computational time, which increases outage duration and cost. Hence, by using GS, a trade-off exists between the accuracy and computational time of outage location. To find a reasonable maximum iteration number for a specific BN, a *potential scale reduction* factor, R, is utilized to diagnose the convergence of the GS at different numbers of iterations [33]. The basic idea is to measure between- and within-sequence variances of generated sample sequences. Specifically, for each M, we start with n sample sequences produced by the GS for each unknown variable in the BN. After discarding the samples generated in the warm-up period, each sequence is divided into two halves of the same size, m, and used to complement the original sequences. All sample sequences are concatenated into a matrix of size $2n \times m$, denoted as $\boldsymbol{\theta}$. Utilizing this matrix, the between-sequence and within-sequence variances are calculated as follows:

$$B_i = \frac{m}{2n-1}\sum_{j=1}^{2n}\left(\bar{\boldsymbol{\theta}}_{.j} - \bar{\boldsymbol{\theta}}_{..}\right)^2 \tag{22}$$

$$V_i = \frac{1}{2n}\sum_{j=1}^{2n} s_j^2 \tag{23}$$

where, B_i is the between-sequence variance of variable i, V_i is the within-sequence variance of variable i, $\bar{\boldsymbol{\theta}}_{.j}$ is the within-sequence means that can be calculated using $\bar{\boldsymbol{\theta}}_{.j} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}_{ij}$. $\bar{\boldsymbol{\theta}}_{..}$ is the overall mean that can be computed using $\bar{\boldsymbol{\theta}}_{..} = \frac{1}{2n}\sum_{j=1}^{2n}\bar{\boldsymbol{\theta}}_{.j}$. s_j^2 denotes the j'th sample sequence variance obtained as $s_j^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{\theta}_{ij} - \bar{\boldsymbol{\theta}}_{.j})^2$. Utilizing V_i and B_i, R_i is defined and computed as [22]:

$$R_i = \sqrt{\frac{\frac{n-1}{n}V_i + \frac{1}{n}B_i}{V_i}}. \tag{24}$$

In theory, the value of R_i equals 1 as $2m \to \infty$. $R_i \gg 1$ indicates that either estimate of the variance can be further decrease by more iterations. In other words, the generated sequences have not yet made a full tour of the target PDF. Alternatively, if $R_i \approx 1$, the sequences are close to the target PDF. Here, following the previous work [22], a threshold $R_\psi = 1.1$ is adopted to select the value of M. Thus, $M \leftarrow 2m$ is set as the number of iterations that satisfy $R_i \leq R_\psi, \forall i$ for the BN. To have the same level of R, the number of iterations M is different for systems with different scales and evidence. In general, the number of M is determined by the size of variables ($|\boldsymbol{D}| + |\boldsymbol{C}| + |\boldsymbol{E}|$). It should be note that $|\boldsymbol{D}| + |\boldsymbol{C}| + |\boldsymbol{E}|$ is not equivalent to the system scale. For example, urban systems can have the similar number of primary nodes as rural systems, but with a significant difference in the number of customers and evidence (both human-based and meter-based evidence).

C. Application Challenges

As detailed below, we discuss some application challenges:

- In actual grids, utilities may have incomplete information regarding secondary topology. This lack of knowledge inhibits the development and parameterization of BN structure. One solution is to apply field inspection or data-driven methods for secondary network topology identification.
- The graphical structure of the proposed BN is established based on the network's topology in normal operations. However, the distribution system often undergoes reconfiguration, which can impact the topology of the grid. Thus, before running the proposed outage detection and location method, previous state estimation-based methods can be utilized to update the topology in normal operations.
- Directed probabilistic graphs alone cannot capture conditional independencies when there are multi-directional power flows caused by meshed topology or high DER penetration. The future work will be done to meet this gap by investigating hybrid graphs.

V. Numerical Results

This section explores the practical effectiveness of the proposed data fusion outage location method. Three real-world distribution feeders are utilized in this case study, which are publicly available online [34]. The topological information is shown in Fig. 5. For each test system, we have evaluated the proposed method under three different observability levels, 25%, 50%, 75%. Note that the observability level is calculated as the ratio of customers with SMs to those without SMs. To validate the average performance of the proposed method, a Monte Carlo approach has been utilized to generate 1500 outage scenarios for each case (a total of 9 cases). In each scenario, the outage location is randomly chosen. All

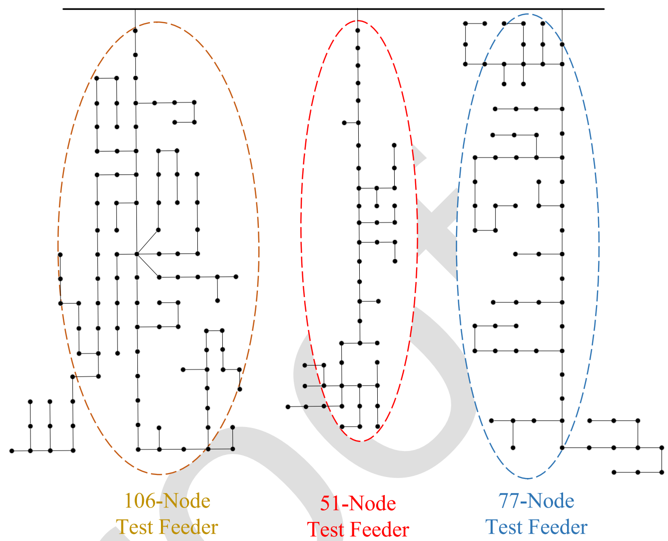


Fig. 5. Three test feeders with different sizes.

aforementioned evidence, including trouble calls, social media messages, last gasp signal, vegetation information, and wind speed, are utilized to perform outage detection and location using the proposed method. Specifically, a portion of customers are randomly selected to install SMs. When a customer is assumed to have the SM, this indicates that the customer is likely to send a last gasp signal when an outage occurs. Based on the historical data, this probability that refers to AMI communication reliability is assigned as 82% in this work. The amount and location of meter-based evidence in each scenario is therefore determined by pre-defined system observability, the geographical distribution of SMs and the location of simulated outages. For the customer trouble calls and social media messages, the human-based evidence is generated using an exponential PDF given ΔT. Note that the parameter of this PDF is considerably different from that of (14) to simulate the uncertainty of the BN parameterization in real-world applications. Consequently, in the outage inference task, we do not know the PDF used to generate evidence and the conditional PDF of the outage location. Basically, in each scenario, the amount and location of the human-based evidence is determined by the total number of customers, the locations of simulated outages, and ΔT. For all scenarios, the value of ΔT is assigned as 10 minutes, which indicates that only a fraction of customers are active in making trouble calls or posting social media messages. For each test system, the vegetation information and the branch's physical parameters are provided by our utility partners. For some unknown parameters, such as tree diameter, we refer to the previous work [27]. Further, depending on the geographical locations of the available systems, the wind speed data is obtained from national oceanic and atmospheric administration (NOVAA) [35]. Since vegetation information and weather data can affect multiple neighboring branches in the same region, the related evidence of the branches in the region is considered to be the same. Moreover, to simulate real-world power outages, 10%, 15%, and 3% of total evidence is assumed to be wrong to simulate

TABLE II
OUTAGE LOCATION OBSERVABILITY SENSITIVITY ANALYSIS

System Name	Observability	Branch-level Accuracy	Branch-level Precision	Branch-level Recall	Branch-level F_1	System-level Accuracy
51-Node Test Feeder	25%	99.05%	86.48%	99.56%	90.65%	69.73%
	50%	99.65%	92.77%	99.82%	95.07%	83.93%
	75%	99.89%	98.38%	100%	98.93%	96.33%
77-Node Test Feeder	25%	98.7%	83.47%	98.88%	88.05%	69.5%
	50%	99.41%	92.43%	98.86%	94.32%	86.6%
	75%	99.60%	92.82%	99.89%	95.24%	88.1%
106-Node Test Feeder	25%	98.92%	83.91%	99.05%	88.61%	69.6%
	50%	99.58%	91.11%	99.54%	94.1%	80.9%
	75%	99.92%	98.19%	100%	98.88%	92.6%

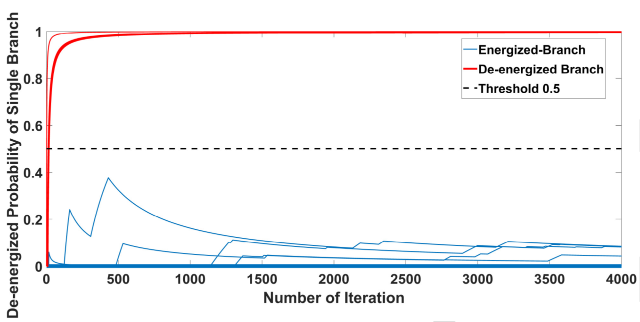


Fig. 6. Branch de-energization probabilities for one outage case.

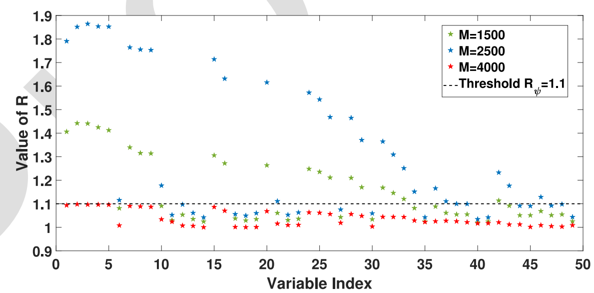


Fig. 7. GS algorithm calibration results for the 51-node system.

the illegitimate calls, natural language processing errors, and AMI communication failure.

A. GS Calibration Results

Basically, the GS calibration is a trial and error process using a specific index, R. Hence, in each test feeder, we have generated 500 sample sequences for each unknown variable in the BN at different sampling iterations, M. Fig. 7 shows the values of R_i in the 51-node test feeder. As can be seen, by increasing the number of M, the values of R_i's tend to converge to 1. By selecting $M = 4000$, all R_i's drop below the user-defined calibration threshold, $R_\psi = 1.1$, which indicates that GS has reached a reasonable number of iterations in this BN. Note that GS calibration is a offline process; as a result, the high computational burden of the trial and error process does not impact the real-time performance of the proposed method.

B. Performance of the Proposed Data Fusion Model

Fig. 6 shows the GS-based inferred dis-connectivity probability values of primary branches in the 51-node test feeder in single outage scenario. As can be seen, for branches downstream of the outage location, these probabilities converge to significantly higher values compared to the branches that are not impacted by the outage event. By using the threshold, the energized branches and the de-energized branches can be easily distinguished to locate the outage. This demonstrates that the BN-based outage location inference method is able to correctly determine the state of the system. Note that there are many blue lines overlapping with the x-axis (with zero de-connectivity probability).

To evaluate the performance of the proposed outage location method for 1500 generated outage cases in the test systems, several statistical metrics are applied among all primary branches and customers, including accuracy, precision, recall, and F_1 score [36], [37]. These indexes are determined as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (25)$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad (26)$$

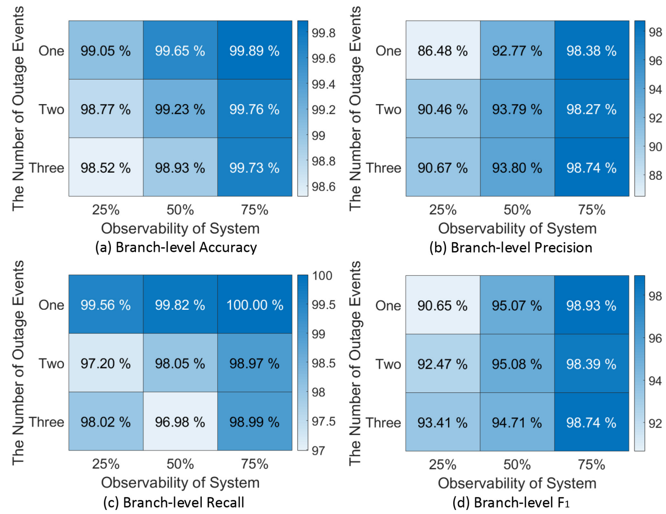
$$Recall = \frac{(TP)}{(TP + FN)} \quad (27)$$

$$F_1 = \frac{\left(\beta^2 + 1\right) * Prec * Recall}{\left(\beta^2 * Prec + Recall\right)} \quad (28)$$

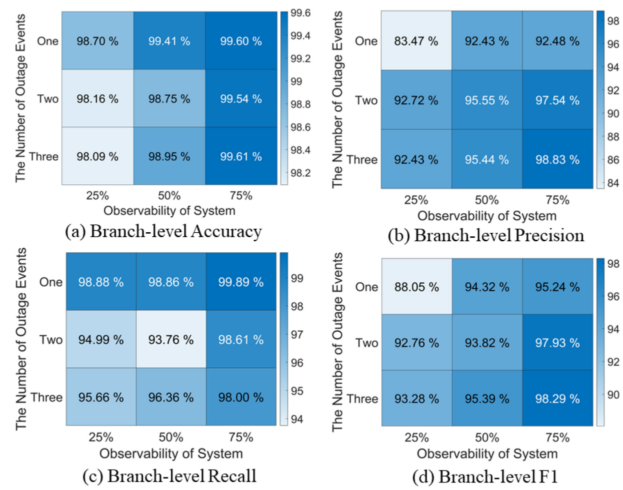
where, TP is the true positive (i.e., state of branch is inferred as de-energized while its actual state is also de-energized), TN is the true negative (i.e., state of branch is considered as an energized while its true state is also energized), FP is the false positive (i.e., state of branch is inferred as de-energized while its actual state is energized), FN is the false negative (i.e., state of branch is inferred as energized while its actual state is de-energized), P and N are the numbers of total positives and negatives, and β is the precision weight which is selected to be 1 in this paper. The average values of these indexes are presented in Table II for the three different test feeders with various observability levels. In all cases, the lowest accuracy, precision, recall, and F_1 score are 98.7%, 83.47%, 98.88%, and 88.05%, respectively. For 50% and 75% observability cases, all branch-level indexes reach values over 0.9. Also, the system-level accuracy is calculated for all cases. Specifically, the system-level accuracy refers to the percentage of times that the states of all the branches/customers have been inferred correctly in outage scenarios. In other words, even though the outage location is inferred correctly, the system-level accuracy may fail because of one misclassified branch. For example, for 77-node test feeder, our method can accurately infer the states of all the branches/customers for about 1300 of the 1500 outage scenarios when the observability level is 50%. In this case, the system-level accuracy is around 86.6%. As shown in the table, when the observability is 25%, the system-level accuracy is about 70%. This could be due to the evidence scarcity. We have analyzed the failed scenarios. In more than 80% of these scenarios, the proposed method can infer the actual location of the outage but misjudged the status of one or two branches. For the cases that have 75% observability, the system-level accuracy is about 90%. This result is not surprising since we have assigned false positive and false negative alarms in each scenario. Such alarms reduce the completeness of outage information. By comparing the results of the three feeders, it can be concluded that the performance of the proposed outage location method improves as the observability increases, due to the high confidence levels of meter-based evidence. Also, the proposed algorithm shows almost the same level of performance over the different test feeders. This result demonstrates that the BN-based outage location method is nearly insensitive to the topology of the underlying network.

To further evaluate the performance of our method, coinciding multiple outage events are generated in three test systems. Note that coinciding outage events refer to multiple simultaneous outages that take place at different locations that are randomly selected. For concreteness, we have also calculated the accuracy under 25%, 50%, and 75% observability levels. Fig. 8 shows the performance indexes as a function of observability level and the number of outages for the three systems. As can be seen, almost in all cases, higher observability improves the performance indexes regardless of the number of coinciding outage events. In all cases, even though the system observability is only 25%, almost all statistical indices are above 90%. When the system observability is 75%, almost all statistical indices are higher than 98%. Also, the indexes have nearly similar values in cases with single and multiple outages. Hence, we can conclude that the method has a stable performance for multiple outages.

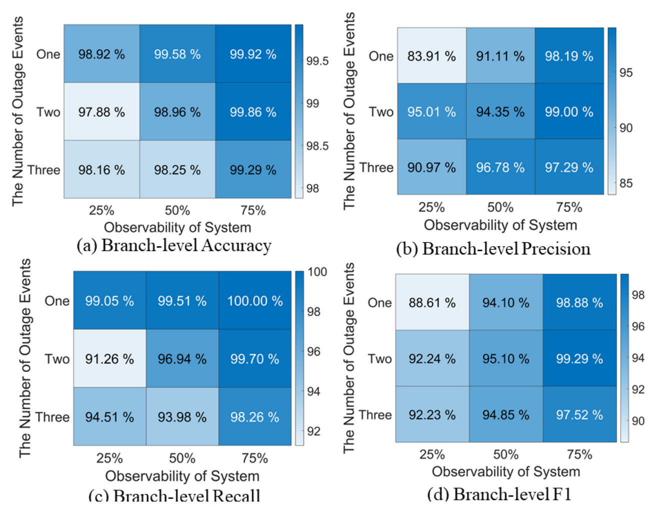
To explore the impact of information on the performance, two more extreme cases are simulated. In the first case, all human-based evidence is removed in the Bayesian network. In the second case, the uncertainty of meter-based evidence is manually increased. Specifically, by changing the values of α_4 and β_4 (see (16) and (17)), the probability that the last gasp can be delivered to the utilities correctly for outage notification is substantially set to 50%. Hence, when a customer is assumed to have the smart meter, there is only 50% probability that the meter will send a last gasp signal when an outage occurs. Using the three real-world test feeders, different scenarios are simulated, and the results for system-level location accuracy are summarized in Fig. 9. Testing results show that the performance of the proposed method is



(a) Results of the 51-node test system with coinciding multi-outage events

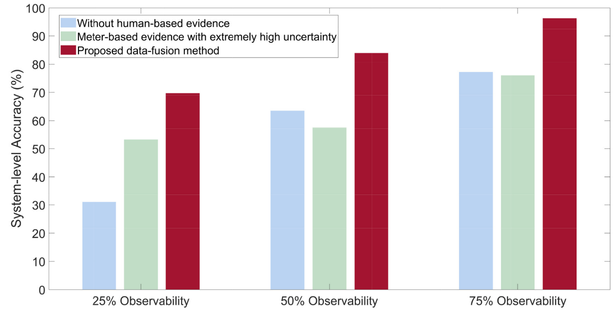


(b) Results of the 77-node test system with coinciding multi-outage events

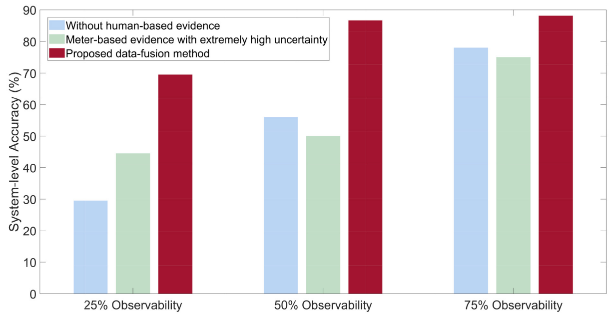


(c) Results of the 106-node test system with coinciding multi-outage events

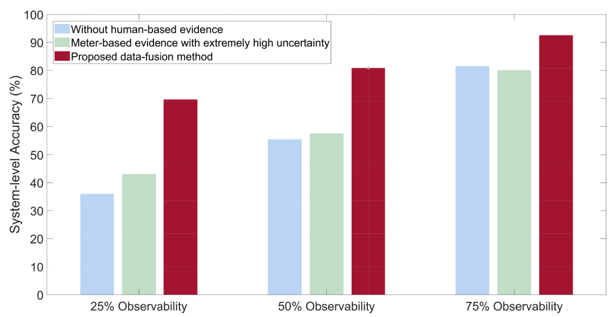
Fig. 8. Sensitivity analysis with coinciding multi-outage events.



(a) Results of the 51-node test system under different evidence scenarios

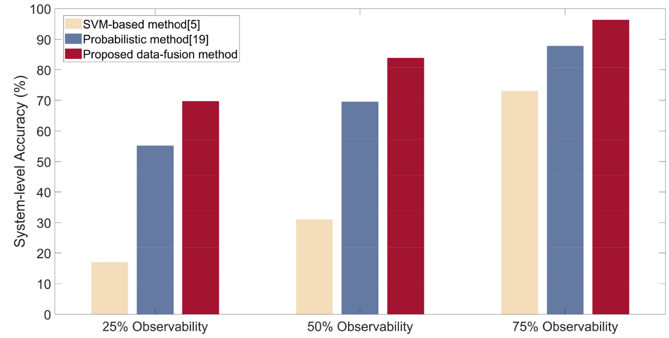


(b) Results of the 77-node test system under different evidence scenarios

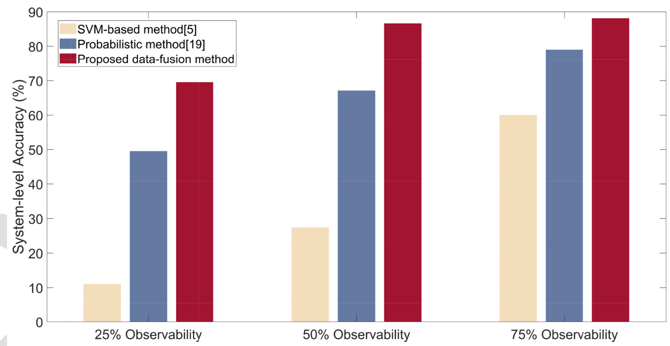


(c) Results of the 106-node test system under different evidence scenarios

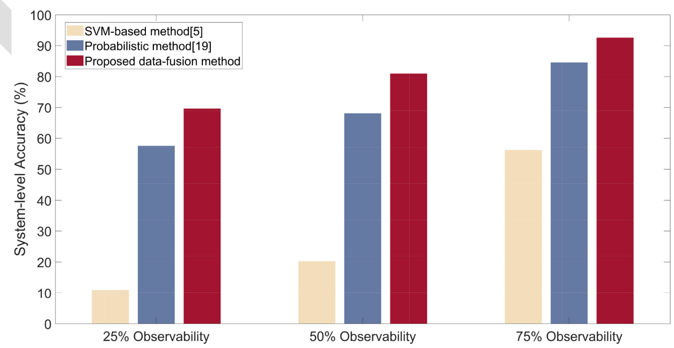
Fig. 9. Performance of the proposed method under different evidence scenarios.



(a) Comparison results of the 51-node test system



(b) Comparison results of the 77-node test system



(c) Comparison results of the 106-node test system

Fig. 10. Comparison of outage location results with two previous methods.

impacted by the amount of outage information. By comparing the results among the three cases, it is clear that incorporating non-metered information (i.e., customer trouble calls and social media messages) is critical for distribution systems with low observability. For the systems with high observability, the uncertainty of the SM last gasp signals can limit the performance of the proposed method.

C. Method Comparison

We have conducted numerical comparisons with two existing outage location methods, a support vector machine (SVM) based approach [5] and a probabilistic approach [19]. Specifically, in [5], smart meter last gasp signals have been utilized to train a SVM mode, one of the state-of-the-art classification models, for estimating the outage location. In [19], the measurements from digital relays at substations and smart meter signals have been incorporated for probabilistic diagnosis. Note that since there are no remote fault indicators installed in the test systems, two constraints (i.e., constraint (4) and (5) in the [19]) are ruled out in the simulations. To ensure a fair comparison among the three methods, the accuracy of all three was assessed based on the same branch-level criteria. The comparison results are demonstrated in Fig. 10. It can be observed that [19] and the proposed method generally outperform [5], especially when the system has low observability. This indicates that our method and [19] can achieve good outage location accuracy with smaller number of smart meters by integrating heterogeneous outage-related data sources, which makes it a suitable method in most distribution grids that are only partially observable. Among the data-fusion-based methods, our method performs slightly better than [19]. The difference between these two approaches is that the proposed method not only uses data from smart meters, but also effectively combines data from non-metered data sources (i.e., trouble calls, social media messages, and weather data).

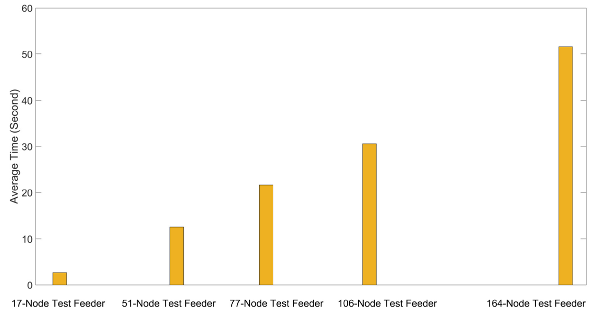


Fig. 11. Average simulation time for the five test feeders.

D. Computational Complexity Analysis

The case study is conducted on a standard PC with an Intel Xeon CPU running at 4.10GHZ and with 64.0GB of RAM and an Nvidia Geforce GTX 1080ti 11.0GB GPU. To provide a comprehensive computational complexity analysis, the proposed method is conducted on two additional real-world distribution feeders: a 17-node and 164-node feeders. The detailed information of these feeders can be found in [10]. Fig. 11 shows the average computational time of outage inference for the test feeders. As described in the figure, by using our standard PC, the average computational time for outage location inference in five test feeders are $\{2.7s, 12.58s, 21.64s, 30.14s, 51.59s\}$, respectively. Also, the proposed model does not infer outage location in a system-wide fashion, but performs feeder-level location estimation. This strategy enables parallel computation of different feeders to further reduce the computational time. These salient features can facilitate the application of practical distribution systems.

VI. Conclusion

In this paper, we have presented a novel multi-source data fusion approach to detect and locate outages in partially observable distribution networks. The problem is cast as the process of inferring the probabilities of post-event operational topology candidates. Our method encodes the network's topology and the causal relationship between outage evidence and branch states into BNs by leveraging the conditional independence inherent in distribution grids. By constructing the BNs, the proposed method is able to infer the connectivity probability of individual primary branches with nearly linear complexity in the size of the network. Moreover, this method exploits data redundancy to reduce the impact of data uncertainty, and is suitable for arbitrary radial distribution systems. Based on simulation results on real-world networks, the proposed method can accurately detect and locate outage events within a short time.

Future study will seek to extend the proposed method in meshed grids with high penetration distributed energy resources. BNs alone cannot fully capture conditional independencies when there are multi-directional power flows. Hence, we plan to explore hybrid graphs that consist of both directed BNs and fully undirected Markov networks. Further, a joint Boltzmann distribution function will be investigated to embody graph parameters.

References

[1] "Industry's First Complete Accident Tolerant Fuel Assembly in Operation at Commercial U.S. Reactor." Office of Nuclear Energy. 2018. [Online]. Available: https://www.energy.gov/ne/office-nuclear-energy AQ3

[2] "August 10, 2020 Derecho." U.S. Department of Commerce, NOAA. 2020. [Online]. Available: https://www.weather.gov/dmx/2020derecho

[3] G. Kumar and N. M. Pindoriya, "Outage management system for power distribution network," in *Proc. Int. Conf. Smart Elect. Grid (ISEG)*, Guntur, India, Sep. 2014, pp. 1–8.

[4] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 5820–5830, Nov. 2018.

[5] Z. S. Hosseini, M. Mahoor, and A. Khodaei, "AMI-enabled distribution network line outage identification via multi-label SVM," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 5470–5472, Sep. 2018.

[6] S.-J. Chen, T.-S. Zhan, C.-H. Huang, J.-L. Chen, and C.-H. Lin, "Nontechnical loss and outage detection using fractional-order self-synchronization error-based fuzzy petri nets in micro-distribution systems," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 411–420, Jan. 2015.

[7] Y. Zhao, R. Sevlian, R. Rajagopal, A. Goldsmith, and H. V. Poor, "Outage detection in power distribution networks with optimally-deployed power flow sensors," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Vancouver, BC, Canada, 2013, pp. 1–5.

[8] R. A. Sevlian, Y. Zhao, R. Rajagopal, A. Goldsmith, and H. V. Poor, "Outage detection using load and line flow measurements in power distribution systems," *IEEE Trans. Power Syst.*, vol. 33, no. 2, pp. 2053–2069, Mar. 2018.

[9] Y. Jiang, C.-C. Liu, M. Diedesch, E. Lee, and A. K. Srivastava, "Outage management of distribution systems incorporating information from smart meters," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 4144–4154, Sep. 2016.

[10] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "Outage detection in partially observable distribution systems using smart meters and generative adversarial networks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5418–5430, Nov. 2020.

[11] S. T. Mak and N. Farah, "Synchronizing SCADA and smart meters operation for advanced smart distribution grid applications," in *Proc. IEEE PES Innovat. Smart Grid Technol. (ISGT)*, 2012, pp. 1–7.

[12] A. N. Samudrala, M. H. Amini, S. Kar, and R. S. Blum, "Distributed outage detection in power distribution networks," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5124–5137, Nov. 2020.

[13] P. Kankanala, S. Das, and A. Pahwa, "AdaBoost^{+}: An ensemble learning approach for estimating weather-related outages in distribution systems," *IEEE Trans. Power Syst.*, vol. 29, no. 1, pp. 359–367, Jan. 2014.

[14] H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2516–2524, Sep. 2016.

[15] A. Gandluru, S. Poudel, and A. Dubey, "Joint estimation of operational topology and outages for unbalanced power distribution systems," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 605–617, Jan. 2020.

[16] S. S. Khan and J. Wei, "Real-time power outage detection system using social sensing and neural networks," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Anaheim, CA, USA, 2018, pp. 927–931.

[17] A. N. Samudrala, M. H. Amini, S. Kar, and R. S. Blum, "Sensor placement for outage identifiability in power distribution networks," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 1996–2013, May 2020.

[18] A. Primadianto and C.-N. Lu, "A review on distribution system state estimation," *IEEE Trans. Power Syst.*, vol. 32, no. 5, pp. 3875–3883, Sep. 2017.

[19] Y. Jiang, "Data-driven probabilistic fault location of electric power distribution systems incorporating data uncertainties," *IEEE Trans. Smart Grid*, vol. 12, no. 5, pp. 4522–4534, Sep. 2021.

[20] A. M. Salman, Y. Li, and M. G. Stewart, "Evaluating system reliability and targeted hardening strategies of power distribution systems subjected to hurricanes," *Rel. Eng. Syst. Safety*, vol. 144, pp. 319–333, Dec. 2015.

[21] C. Fu, Z. Yu, and D. Shi, "Bayesian estimation based load modeling report," 2018, *arXiv:1810.07675*.

[22] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL, USA: CRC Press, 2013.

[23] D. Koller, N. Friedman, and B. F, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009. AQ4

[24] W. Luan, J. Peng, M. Maras, J. Lo, and B. Harapnuk, "Smart meter data analytics for distribution network connectivity verification," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 1964–1971, Jul. 2015.

[25] W. Wang, N. Yu, B. Foggo, J. Davis, and J. Li, "Phase identification in electric power distribution systems by clustering of smart meter data," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Anaheim, CA, USA, 2016, pp. 259–265.

[26] B. Foggo and N. Yu, "Improving supervised phase identification through the theory of information losses," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2337–2346, May 2020.

[27] M. Ouyang and L. Dueñas-Osorio, "Multi-dimensional hurricane resilience assessment of electric power systems," *Struct. Safety*, vol. 48, pp. 15–24, May 2014.

[28] N. Bassamzadeh and R. Ghanem, "Multiscale stochastic prediction of electricity demand in smart grids using Bayesian networks," *Appl. Energy*, vol. 193, pp. 369–380, May 2017.

[29] "Smart Meters Can Reduce Power Outages and Restoration Time." National Electrical Manufacturers Association. 2021. [Online]. Available: https://www.nema.org/storm-disaster-recovery/smart-grid-solutions/smart-meters-can-reduce-power-outages-and-restoration-time

[30] Y. Jiang, "Data-driven fault location of electric power distribution systems with distributed generation," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 129–137, Jan. 2020.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 851–860.

[32] C. Fu *et al.*, "Bayesian estimation based parameter estimation for composite load," 2019, *arXiv:1903.10695*.

[33] P.-C. Bürkner, "Advanced Bayesian multilevel modeling with the R package brms," 2017, *arXiv:1705.11123*.

[34] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, "A time-series distribution test system based on real utility datd," in *Proc. North Amer. Power Symp. (NAPS)*, 2019, pp. 1–6.

[35] "Climate Data Online." National Oceanic and Atmospheric Administration. 2021. [Online]. Available: https://https://www.ncdc.noaa.gov/cdo-web/

AQ5 [36] N. Sokolova, M. Japkowicz and S. Szpakowicz, *Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation*. Heidelberg, Germany: Springer, 2006.

[37] Y. Zhang, J. Liang, Z. Yun, and X. Dong, "Knowledge-based system for distribution system outage locating using comprehensive information," *IEEE Trans. Power Del.*, vol. 32, no. 6, pp. 2398–2407, Dec. 2017.

Yuxuan Yuan (Member, IEEE) received the B.S. degree in electrical and computer engineering from Iowa State University, Ames, IA, USA, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include distribution system state estimation, synthetic networks, data analytics, and machine learning.

Kaveh Dehghanpour received the B.Sc. and M.S. degrees in electrical and computer engineering from the University of Tehran in 2011 and 2013, respectively, and the Ph.D. degree in electrical engineering from Montana State University in 2017. He is currently a Postdoctoral Research Associate with Iowa State University. His research interests include application of machine learning and data-driven techniques in power system monitoring and control.

Zhaoyu Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Shanghai Jiaotong University, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology. He is the Northrop Grumman Endowed Associate Professor with Iowa State University. His research interests include optimization and data analytics in power distribution systems and microgrids. He was a recipient of the National Science Foundation CAREER Award, the Society-Level Outstanding Young Engineer Award from IEEE Power and Energy Society (PES), the Northrop Grumman Endowment, College of Engineering's Early Achievement in Research Award, and the Harpole-Pentair Young Faculty Award Endowment. He is the Principal Investigator for a multitude of projects funded by the National Science Foundation, the Department of Energy, National Laboratories, PSERC, and Iowa Economic Development Authority. He is the Chair of IEEE PES PSOPE Award Subcommittee, the Co-Vice Chair of PES Distribution System Operation and Planning Subcommittee, and the Vice Chair of PES Task Force on Advances in Natural Disaster Mitigation Methods. He is an Associate Editor for IEEE TRANSACTIONS ON POWER SYSTEMS, IEEE TRANSACTIONS ON SMART GRID, IEEE OPEN ACCESS JOURNAL OF POWER AND ENERGY, IEEE POWER ENGINEERING LETTERS, and *IET Smart Grid*.

Fankun Bu, photograph and biography is not available at the time of publication.