A Systematic Procedure for Topological Path Identification with Raw Data Transformation in Electrical Distribution Networks

Maurizio Vassallo*, Alireza Bahmanyar†, Laurine Duchesne†, Adrien Leerschool†,
Simon Gerard‡, Thomas Wehenkel‡, Damien Ernst*§
* Department of Electrical Engineering and Computer Science, Liège, Belgium
† Intelligent Systems Solutions, Haulogy, Neupré, Belgium
‡ RESA, Liège, Belgium
§ LTCI, Télécom Paris, Institut Polytechnique de Paris, France
{mvassallo, dernst}@uliege.be
{alireza.bahmanyar, laurine.duchesne, adrien.leerschool}@haulogy.net
{simon.gerard, thomas.wehenkel}@resa.be

Abstract—This paper introduces a systematic approach to address the topological path identification (TPI) problem in power distribution networks. Our approach starts by listing the DSO's raw information coming from several sources. The raw information undergoes a transformation process using a set of transformation functions. This process converts the raw information into well-defined information exploitable by an algorithm. Then a set of hypothetical paths is generated, considering any potential connections between the elements of the power distribution system. This set of hypothetical paths is processed by the algorithm that identifies the hypothetical paths that are compatible with the well-defined information. This procedure operates iteratively, adapting the set of transformation functions based on the result obtained: if the identified paths fail to meet the DSO's expectations, new data is collected and/or the transformation functions found to be responsible for the discrepancies are modified. The systematic procedure offers practical advantages for DSOs, including improved accuracy in path identification and high adaptability to diverse network configurations, even with incomplete or inaccurate data. Consequently, it emerges as a useful tool for the construction of digital twins of power distribution networks that aligns with DSO expectations.

Index Terms—Topological path identification, raw information, active and backup paths, digital twin.

I. Introduction

The distribution of electricity within power grids is a complex process characterised by networks of interconnected paths that transmit energy from a source to individual consumers. Within these intricate webs, each customer is generally associated with two types of paths: active paths, responsible for the distribution of electricity, and backup paths, which are used in case of emergencies. The identification of customer paths is important for distribution system operators (DSOs) for several reasons. First, knowing the topological paths can serve as a vital tool to identify and inform customers about potential power outages due to network maintenance. Second, it allows DSOs to proactively anticipate and address potential power quality issues within

the network, such as overvoltages and outages. Third, knowledge of backup paths can be useful to keep power supply, especially during emergencies.

Solving the topological path identification (TPI) problem involves the identification of both active and backup paths for each customer in a given power distribution system.
DSOs try to solve the TPI problem, but unfortunately, they use case-specific techniques that might not always be practical in reality. For example, DSOs often rely on GPS data of the elements. They incorrectly assume this data to be accurate and complete. However, considering the scale of the networks and the limited precision of the GPS devices, there may be some missing values or some inaccuracies in the GPS recordings that must be taken into account. Moreover, DSOs tend to concentrate only on a single active path for each customer, despite the potential existence of multiple paths that align with their information. Additionally, DSOs often overlook the significance of backup power supply paths, ignoring them. Therefore, DSOs lack a comprehensive framework, resulting in an inaccurate or incomplete identification of the paths, with consequent waterfall effects on other network operations.

Many studies have attempted to address the TPI problem based on the data available to DSOs. Among these techniques, some strategies use the concept of distance between elements to identify the paths. For example, in [1], the authors propose an identification methodology that establishes connections between customers and the nearest feeder based on their spatial proximity. Similarly, the authors in [2], propose a work to identify the connectivity of elements of a low-voltage power network based on the distance of their coordinates, moreover the authors identify and rectify connectivity errors between the elements that may arise in the network. In [3], authors develop an algorithm to transform the data from the Open-StreetMap database into GPS information about the nodes to

reconstruct the electrical power grid. Alternative techniques take advantage of the smart meters' measurements. In [4], the authors identify a low-voltage distribution network based on GIS data, energy market data, and network measurements. The authors in [5] use the voltage and current time-series from a limited number of smart sensors to infer the network topology and the lines' impedance. Similarly, the approach in [6] explores the correlation properties of voltage measurements to estimate the power distribution network topology. Authors in [7], identify the switch states, phases, and line parameters of a low-voltage grid in German using the voltage and current measurements of a limited number of smart sensors. Despite the interesting results obtained, the prior research exhibits a dual set of limitations that tend to produce suboptimal solutions. First, they do not explicitly mention how each piece of raw information is transformed into well-defined information. This can be a problem if the output of their method is incorrect, since DSOs or data scientists may not easily identify the source of the error. Second, the mentioned works mainly focus on identifying a single active path among the many compatible ones, and they ignore the backup paths. In other words, they solve the TPI problem in an unstructured and partial way.

To address the limitations of existing approaches, this paper presents a systematic procedure for solving the TPI problem. This systematic approach is characterised by listing an initial set of the DSO's raw information coming from several sources. We then define a set of transformation functions that are used to transform the raw information into well-defined information exploitable by an algorithm. We define a set of hypothetical paths, and we narrow it down by identifying the feasible paths that are compatible with the set of well-defined information. The compatible paths are evaluated with a diagnostic function. If these feasible paths fail to meet the evaluation criteria, new data is collected and/or the transformation functions found to be responsible for the discrepancies are modified to ensure that the compatible paths align with the DSO's requirements. The systematic procedure of listing all the information and transformations can effectively complement existing methods, hence augmenting their capabilities and leading to more accurate results. Moreover, the adaptability of this procedure to any available data, from GIS data to smart meter recordings, makes it suitable for every network and different DSO context.

The rest of the paper is organised as follows. Section II presents the definition of the power network's elements used in this work. Section III defines the concept of topological paths, with particular emphasis on real, active, backup and hypothetical paths. The problem statement is presented in Section IV, outlining the objectives of the paper. Section V defines the methodology used to solve the problem. The methodology is applied to an academic example in Section VI. Section VII presents complexity and scalability considerations of the proposed approach, presenting a potential alternative solution.

Section VIII concludes the work and discusses possible future works.

II. POWER NETWORK ELEMENTS

Power distribution networks typically encompass a wide range of elements. The set of elements $\mathcal{E}$ is defined as:

$$\mathcal{E} = \{e_1, e_2, ..., e_{|\mathcal{E}|}\} \tag{1}$$

where a single element inside the set $\mathcal{E}$ is denoted as e_i and $|\mathcal{E}|$ represents the total number of elements in the set $\mathcal{E}$.

The set of attributes $\mathcal{A}$ is defined as follows:

$$\mathcal{A} = \{a_1, a_2, ..., a_{|\mathcal{A}|}\} \tag{2}$$

where a single attribute is denoted as a_i and $|\mathcal{A}|$ represents the total number of attributes in the set $\mathcal{A}$.

A single element $e \in \mathcal{E}$ is defined as a list of attributes. Formally:

$$e = (e.a_1, e.a_2, ..., e.a_{|e|}) \tag{3}$$

where the dot symbol is used to access the attributes of the element e, a_i is the i-th attribute, and $a_{|e|}$ is its last attribute.

The location attribute of the elements is identified by pairs of coordinates. Elements may have only one or more than one pair of coordinates. Given one element, $e \in \mathcal{E}$, it is possible to access its coordinates as follows:

$$c = e.coor \tag{4}$$

where c is the list of the element's coordinates. We refer to the i-th pair of list of coordinates, c, of the element e with the following notation: $e.coor_i$, with $1 \leq i \leq |c|$, and $|c|$ the total number of pairs of coordinates of c. For example, $e.coor_1$ is the first pair of coordinates of the element, e, $e.coor_i$ is the i-th pair of coordinates and $e.coor_{|c|}$ is the last one.

Each element within the network can be categorised based on its type attribute, serving as a characteristic that distinguishes it from other elements. The set of types, denoted as $\mathcal{T}$, includes all the possible types associated with the elements in $\mathcal{E}$:

$$\mathcal{T} = \{t_1, t_2, ..., t_{|\mathcal{T}|}\}. \tag{5}$$

Among the typical types, there may be $customer$ connection points, electrical $transformer$ stations, $switch$ devices and other types that are possible to find in power distribution networks.

Given an element $e \in \mathcal{E}$, its type is accessed as follows:

$$t = e.type \tag{6}$$

The $Subset()$ function is used to identify all elements of a specific type. This function takes two input parameters: the element set, $\mathcal{E}$, and a type $t \in \mathcal{T}$. The result of this function is a subset of $\mathcal{E}$, representing those elements of $\mathcal{E}$ associated with the specified type, t. Formally:

$$Subset(\mathcal{E}, t) = \{e \in \mathcal{E} \mid e.type = t\}. \tag{7}$$

The distance between two elements is calculated using a $Dist()$ function. The $Dist()$ function takes two elements

$e_i, e_j \in \mathcal{E}$ as input, and it outputs a scalar value, d, representing the minimum distance between their coordinates. The $Dist()$ function is defined as follows:

$$d = Dist(e_i, e_j) = \min_{\substack{\forall a \in e_i.coor, \\ \forall b \in e_j.coor}} ||a - b|| \qquad (8)$$

where $||\cdot||$ represents the Euclidean distance.

Moreover, given an element, $e \in \mathcal{E}$, and a generic set of elements, $\mathcal{S}$, we define a function $Closest()$. This function identifies an element within the set $\mathcal{S}$ that is closest to the element e. The $Closest()$ function can be defined as follows:

$$s^* = Closest(e, \mathcal{S}) = \arg\min_{s \in S}\big(Dist(e, s)\big) \qquad (9)$$

where s^* denotes the element of the set $\mathcal{S}$ that is the closest to e.

III. TOPOLOGICAL PATHS

This section serves as the foundation for understanding the concepts of real, active, backup and hypothetical paths within the context of this work.

A. Real paths

The flow of electricity in an electrical network generally follows a unique path from an MV/LV transformer to the customer. Between the customer and the transformer, there may be many other elements. All of these connected elements together form a topological path. The set of all real paths, denoted as $\mathcal{P}$, is expressed as follows:

$$\mathcal{P} = \{p \in \mathcal{P} \mid p \text{ is the real list of unique elements} \qquad (10)$$
$$\text{from a customer to a transformers}\}.$$

Each individual path, $p \in \mathcal{P}$, uniquely represents the real path that electricity can follow from an MV/LV transformer to the customer. Due to the large scale of the power distribution networks and limitations of the recorded data, these paths are often not clearly known for the DSOs.

The i-th element of a path, $p \in \mathcal{P}$, is referred with the following notation: p_i, with $1 \le i \le |p|$, and $|p|$ the total number of elements of the path p. For example, p_1 and $p_{|p|}$ are always a customer and the transformer, respectively.
Two elements in a path are considered connected if they directly follow each other. For a path p, p_i and p_j are connected if the condition $j = i+1$ is verified, with $i, j < |p|$.

Given an element of type $customer$, $e \in Subset(\mathcal{E}, customer)$, the set of paths $\mathcal{P}_e$, represents the subset of all paths in which the customer, e, is present as starting point. Formally, the subset $\mathcal{P}_e$ is determined using the following notation:

$$\mathcal{P}_e = \{p \in \mathcal{P} \mid p_1 = e\}. \qquad (11)$$

The length of a given path $p \in \mathcal{P}$ is calculated with the following formula:

$$l = LengthPath(p) \qquad (12)$$
$$= \sum_{i=1}^{|p|-1} \underbrace{Dist(p_i, p_{i+1})}_{i} + \underbrace{\sum_{j=1}^{|p_i.coor|-1} Dist(p_i.coor_j, p_i.coor_{j+1})}_{ii}.$$

The formula calculates the total length (l) of a given path (p) by summing the distances between consecutive elements (i) along the path and adding the length of a single element (ii). The length of a single element is determined by the sum of the distances between pairs of consecutive coordinates.

B. Active and backup paths

A single customer can have many paths to different transformers, but generally only one path is active, while all the others are backup paths used in case of emergency. The status attribute of the switches defines the activity or inactivity of a path.

The set of active paths, $\mathcal{P}_a$ is defined as follows:

$$\mathcal{P}_a = \{p \in \mathcal{P} \mid p \text{ contains no elements of} \qquad (13)$$
$$\text{type switch with status open}\}.$$

Similarly, the set of backup paths, $\mathcal{P}_b$, is formulated as follows:

$$\mathcal{P}_b = \{p \in \mathcal{P} \mid p \text{ contains at least on element of} \qquad (14)$$
$$\text{type switch with status open}\}.$$

Given the set of paths, $\mathcal{P}_a$ and $\mathcal{P}_b$, the DSOs can effectively analyse the paths for each customer and identify both its active and backup paths. In particular, given an element of type customer, $e \in Subset(\mathcal{E}, customer)$, the sets of the customer's active and backup paths are identified as follows:

$$P_{a,e} = \{p \in \mathcal{P}_a \mid p_1 = e\}, \qquad (15)$$
$$P_{b,e} = \{p \in \mathcal{P}_b \mid p_1 = e\}. \qquad (16)$$

C. Hypothetical paths

In this context, hypothetical paths refer to a set of potential paths within the distribution network. The term "hypothetical" emphasises the uncertainty in determining the topological paths due to incomplete or inaccurate information. Indeed, the hypothetical paths are constructed under the consideration that only the initial point of the paths is known, while all connections among the in-between elements to a transformer may be unknown.

The set of hypothetical paths, denoted as $\mathcal{H}$, is expressed as follows:

$$\mathcal{H} = \{h \in \mathcal{H} \mid h \text{ is any list of unique elements} \qquad (17)$$
$$\text{from a customer to a transformer}\}.$$

The size of the hypothetical path set denoted as $|\mathcal{H}|$, given a set of elements $\mathcal{E}$ is calculated with the following parameters.

- E: the total number of elements in the set $\mathcal{E}$, $E = |\mathcal{E}|$.

- C: the total number of elements of type *customer* in the set $\mathcal{E}$, $\texttt{C} = |Subset(\mathcal{E}, customer)|$.
- T: the total number of elements of type *transformer* in the set $\mathcal{E}$, $\texttt{T} = |Subset(\mathcal{E}, transformer)|$.
- $\texttt{R} = \texttt{E} - \texttt{C} - \texttt{T}$: the remaining number of elements that can be connected to each other, excluding elements of type *customer* and *transformer*.
- $^{n}P_{k}$: the permutation formula, which calculates the number of ways to arrange k items from a set of n items ($^{n}P_{k} = \frac{n!}{(n-k)!}$).

The following equation calculates the total number of possible hypothetical paths in the network:

$$|\mathcal{H}| = \sum_{i=0}^{\texttt{R}} \texttt{C} \cdot {}^{\texttt{R}}P_{i} \cdot \texttt{T}. \tag{18}$$

Equation (18) can be interpreted as follows. It starts by considering the simplest paths, which involves directly connecting a customer to a transformer in a single step ($i = 0$).

Gradually increasing the number of steps, it considers longer paths that involve more intermediate elements. The elements are considered in various permutations (${}^{\texttt{R}}P_{i}$).

The calculation ends when the maximum step ($i = R$) is reached, indicating the longest paths possible in the network.

It is worth noting that the same notation used for the real path set can be extended to the hypothetical path set. For instance, when identifying the hypothetical paths associated with a specific customer, denoted as $e \in Subset(\mathcal{E}, customer)$, they can be expressed as follows:

$$\mathcal{H}_{e} = \{h \in \mathcal{H} \mid h_{1} = e\}. \tag{19}$$

IV. PROBLEM STATEMENT

DSOs aim to identify a subset of the hypothetical paths set $\mathcal{H}$, that represent the closest approximation of the set of real paths $\mathcal{P}$, denoted as $\hat{\mathcal{P}}$. This identification process depends on a set of raw information, $\mathcal{I}$. Based on this information, it is possible to construct an approximate representation of the set of elements known to the DSO and their associated attributes, respectively denoted by $\hat{\mathcal{E}}$ and $\hat{\mathcal{A}}$.

Ideally, we would compare the set of real paths, $\mathcal{P}$, and its approximation, $\hat{\mathcal{P}}$, using a rigorous metric to measure the similarity between the two sets. This metric would evaluate the degree of similarity between the sets, resulting in a similarity score ranging from 0 to 1. The evaluation of the two sets can result in one of the three possible outcomes: a perfect match indicated by a score of 1, complete dissimilarity corresponds to a score of 0, and partial congruence represented by any score between 0 and 1.

Unfortunately, this evaluation is not possible since the set of real paths, $\mathcal{P}$, is generally unknown to DSOs. For solving the lack of a proper evaluation between sets, the solution found by our procedure is validated with a $Diagnostic()$ function. This function serves as a critical tool for assessing the validity of the solution. The $Diagnostic()$ function evaluates the found paths by checking any possible discrepancy from reality.

V. METHODOLOGY

This section outlines the systematic procedure proposed in this study to address the TPI problem in power distribution networks. The main goal of this methodology is to identify the hypothetical paths in $\mathcal{H}$ that are compatible with a set of well-defined information. This set represents the closest approximation of the set of real paths $\mathcal{P}$, denoted as $\hat{\mathcal{P}}$.

A. Raw information, transformation functions and well-defined information

Raw information refers to available information or data that comes from various sources. Usually, the raw information is provided in an unorganised way, such as text, images, and spreadsheets, that cannot be potentially exploited directly. The set of raw information is denoted as follows: $\mathcal{I} = \{i_{1}, i_{2}, ..., i_{|\mathcal{I}|}\}$.

Transformation functions are responsible for transforming the available raw information into clear knowledge that is relevant within the specific context of a power network. The set of transformation functions is denoted as follows: $\mathcal{F} = \{f_{1}, f_{2}, ..., f_{|\mathcal{F}|}\}$. Mathematically, the transformation functions take one or more pieces of raw information as input and return the transformed information as output.

Well-defined information refers to the transformation of the raw information set $\mathcal{I}$, using the transformation function set $\mathcal{F}$. This well-defined information forms the basis to provide a structured foundation, essential for identifying compatible paths. The set of well-defined information is denoted as follows: $\mathcal{I}' = \{i'_{1}, i'_{2}, ..., i'_{|\mathcal{I}'|}\}$, with a single piece well-defined information being the output of a transformation functions, i.e. $i' = f(i)$.

Given the set of well-defined information, $\mathcal{I}'$, the goal is to identify a subset of the hypothetical path set compatible with $\mathcal{I}'$, denoted as $\mathcal{H}^{\mathcal{I}'}$. This set represents the best approximation of the set of real paths, denoted as $\hat{\mathcal{P}}$, and it is considered solution for the TPI problem.

B. Methodology steps

Such methodology is performed in six steps:

1) The first step involves listing the information available to the DSO, denoted as $\mathcal{I}$.
2) Then establish an initial set of transformation functions, denoted as $\mathcal{F}$.
3) Extract the relevant well-defined information, denoted as $\mathcal{I}'$, from the raw available data $\mathcal{I}$, using the transformation functions $\mathcal{F}$.
4) With the well-defined information $\mathcal{I}'$, use an algorithm to identify a subset of the hypothetical paths compatible with the set of well-defined information, $\mathcal{I}'$, denoted as $\mathcal{H}^{\mathcal{I}'}$.
5) Evaluate the solution found with the $Diagnostic()$ function.

- If any issue is detected, new data is collected and/or the transformation functions are modified. Return to step 3.
- Otherwise, if no issue is observed, the set $\mathcal{H}^{\mathcal{I}'}$ is considered as the solution for the problem, and it represents the closest representation of the set of real paths $\mathcal{P}$, denoted as $\hat{\mathcal{P}}$. Continue to step 6.

6) Once the solution $\hat{\mathcal{P}}$ is found, identify the customers' active and backup paths.

These six simple steps collectively constitute a systematic and adaptive procedure that enables the identification of both active and backup paths, ensuring compatibility with the well-defined information.

VI. ACADEMIC EXAMPLE

To illustrate the methodology used to solve the TPI problem, an academic example of a power distribution network with a limited number of elements is considered. This example represents a network where two main feeders are connected to two MV/LV transformers, supplying some customers through active and backup paths.

The example, along with the associated data and code, is publicly available on the GitHub repository *https://github.com/TPIproblem/TPIpaper*, featuring an open-source Apache v2 licence.

A. Real network

The real configuration of the network under consideration is illustrated in Fig. 1a. It includes different kinds of elements such as transformers, lines, switches, and customers. Each element's position is defined by one or more pairs of coordinates that localise them within a 2-D plane, as it is possible to see for some representative elements: e_2, e_{13} and e_{21}. Switch elements are visually differentiated by an empty square or a filled square, to indicate their status: open or closed, respectively. More details about the coordinates of the elements can be found in the Appendix.

B. Raw information

We assume the set of raw information available to the DSO, denoted as $\mathcal{I}$, is provided by some pieces of information:

i_1: Documents containing the DSO's list of elements, their coordinates, and their types. However, some elements are missing, and some coordinates are inaccurate, as illustrated in Fig. 1b.

i_2: The DSO lacks the information about the customers connections to the network, but their general practice is to minimise the cost of connecting customers to the grid.

i_3: Information collected from a technical meeting with the DSO's experts where they stated that, for economic considerations (cost per metre) and efficiency concerns (energy loss per metre), the connections between two elements should not be too long.

i_4: DSO's general practice to reduce energy losses by minimising the total path length of each customer.

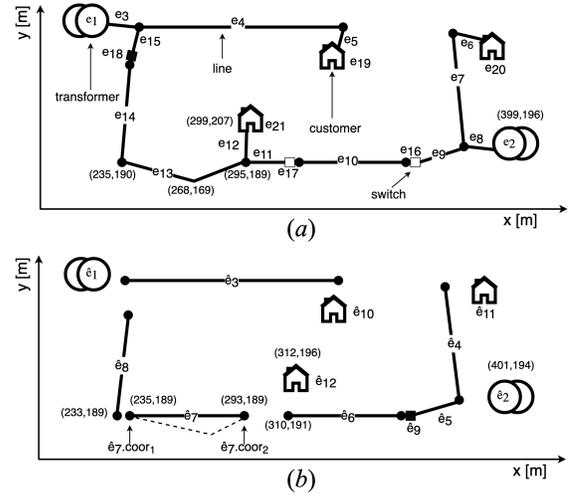i_5: The knowledge that each customer is supplied by exactly one transformer.

Fig. 1. Real network, Fig. 1a, and the approximate network when considering only the elements known by the DSO, Fig. 1b. It is possible to note some inaccuracies in the GPS location and some missing elements.

C. Well-defined information

To identify accurate paths, our methodology uses a set of transformation functions, denoted as $\mathcal{F}$. These transformation functions convert the set of information available to the DSO, denoted as $\mathcal{I}$, into well-defined information, denoted as $\mathcal{I}'$. This well-defined information is used by an algorithm to find an approximation of the set of real paths $\mathcal{P}$, expressed as $\hat{\mathcal{P}}$.

For example, the transformation functions transform the raw information, given by database recordings, spreadsheet files or technical meetings, into well-defined information, like power distribution network elements and their type, GPS coordinates and business rules, as shown in Fig. 2.
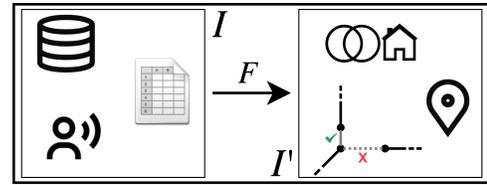

Fig. 2. Transformation functions transforming different kinds of data into well-defined information.

In this case, the well-defined information $\mathcal{I}'$, obtained as a result of transforming the raw information $\mathcal{I}$ with the transformation functions $\mathcal{F}$, is presented as follows.

$i'_1 = f_1(i_1)$: set of elements, $\hat{\mathcal{E}}$, and their attributes, $\hat{\mathcal{A}}$.

$i'_2 = f_2(i_2)$: elements of type *customer* are connected to the closest element of type *line*.
Therefore, for every path $h \in \mathcal{H}^{\mathcal{I}'}$, the condition $h_2 = Closest\big(Subset(\hat{\mathcal{E}}, line), h_1\big)$ must be satisfied.

$i'_3 = f_3(i_3)$: elements of type *transformer*, *line* and *switch* are connected to the closest element only if the distance calculated using the elements GIS coordinates is smaller than a maximum value R.

Therefore, for every path $h \in \mathcal{H}^{\mathcal{I}'}$ and for any $i < |h| - 1$ the condition $h_{i+1} = Closest(\mathcal{E}, h_i)$ AND $Dist(h_i, h_{i+1}) < R$ must be satisfied.

$i'_4 = f_4(i_4)$: the total length of a path, defined as the sum of the lengths of the single lines, is at most L.

Therefore, for every path $h \in \mathcal{H}^{\mathcal{I}'}$ the condition $LengthPath(h) < L$ must be satisfied.

Given the well-defined information i'_1, it is possible to define the set of elements known to the DSO, $\hat{\mathcal{E}}$, and their attributes $\hat{\mathcal{A}}$. Generally $\hat{\mathcal{E}}$ is a subset of $\mathcal{E}$ as there might be additional elements in the actual network that are not present in the data available to the DSO.

In such an academic example, the set $\hat{\mathcal{E}}$ is given by the elements $\hat{e}_1$ to $\hat{e}_{12}$ as shown in Fig. 1b. The set of types, $\hat{\mathcal{T}}$, is composed of: $transformer$ (elements $\hat{e}_1$ and $\hat{e}_2$), $line$ ($\hat{e}_3$ to $\hat{e}_8$), $switch$ ($\hat{e}_9$) and $customer$ ($\hat{e}_{10}$ to $\hat{e}_{12}$).

Figure 1b illustrates the geographical position of the elements in the network, as defined by i'_1. In particular, it is possible to note that the position of the elements of type $transformer$ and $customer$ are identified by one pair of coordinates; elements of type $line$ are identified by two pair of coordinates; elements of type $switch$ are also identified by a single pair of coordinates, and they are represented differently, either as empty or filled, to better visualise their status attribute. An empty representation corresponds to a switch whose status is open, while a filled representation corresponds to a switch whose status is closed.

Table I summarises the attributes of each element.

TABLE I
ELEMENTS AND THEIR ATTRIBUTES

Element	Attributes		
	Coordinates	Type	Status
$\hat{e}_1$ and $\hat{e}_2$	One pair	$transformer$	-
$\hat{e}_3$ to $\hat{e}_8$	Two	$line$	-
$\hat{e}_9$	One pair	$switch$	Open or close
$\hat{e}_{10}$ to $\hat{e}_{12}$	One pair	$customer$	-

As mentioned, the data may not always be accurate or there may be some wrong or missing coordinates. For example, the customer $\hat{e}_{12}$, in the data available to DSO is closer to the element $\hat{e}_6$ while in reality the customer is closer to the element e_{13} (corresponding to the element $\hat{e}_7$). Furthermore, the element $\hat{e}_7$ is missing a pair of coordinates in the information available to the DSO (highlighted in the dashed line in Fig. 1b). Indeed, for lines, DSOs have only the endpoints (illustrated as $\hat{e}_7.coor_1$ and $\hat{e}_7.coor_2$), while intermediate points may not be available. Therefore, in the reconstructed network, the line is modelled as a straight line between its two endpoints. In such a case, this approximation introduces the possibility that the length of the reconstructed line may deviate from the actual length in the real network.
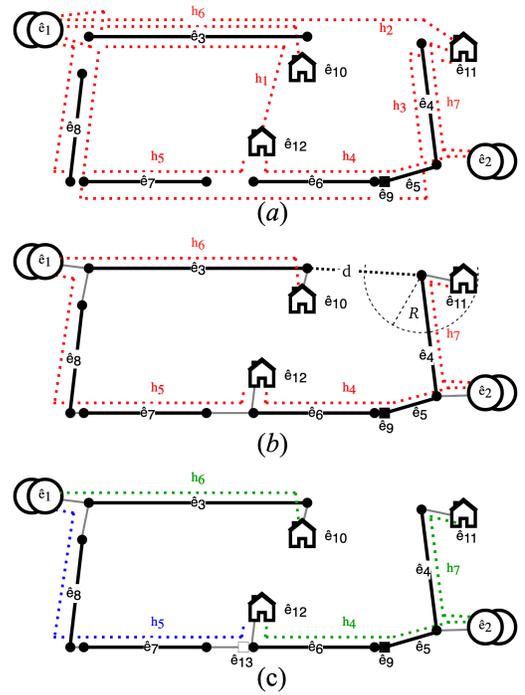

Fig. 3. Application of the systematic procedure to an academic power distribution network. Some hypothetical paths are shown in Fig. 3a. Figure 3b shows the network situation after excluding the paths not compatible with the well-defined information from the hypothetical path set. Figure 3c shows the network situation after modifying the transformation functions, $\mathcal{F}$, to take into account some identified issues. After finding a solution to the TPI problem, it is possible to identify the active and backup paths, as illustrated in Fig. 3c.

D. Hypothetical paths compatible with the well-defined information

The exact number of hypothetical paths, represented by the set $\mathcal{H}$, cannot be determined without knowing the set of elements $\mathcal{E}$ present in the power distribution network. However, the size of a subset of $\mathcal{H}$, denoted as $\mathcal{H}^{i'_1}$, representing the hypothetical paths compatible with the well-defined information i'_1 can be determined.

To calculate the size of the set $\mathcal{H}^{i'_1}$ we need the set of elements $\hat{\mathcal{E}}$ and the set of types $\hat{\mathcal{T}}$. These two sets are obtained using the well-defined information i'_1. Therefore, in this scenario, considering the set of elements $\hat{\mathcal{E}}$ and the set of attributes $\hat{\mathcal{T}}$, the size of the hypothetical paths set, $\mathcal{H}^{i'_1}$, is calculated using Eq. (18) with $\mathtt{E} = 12$, $\mathtt{T} = 2$, $\mathtt{C} = 3$ and $\mathtt{R} = 7$.

$$|\mathcal{H}^{i'_1}| = 82200. \tag{20}$$

Some hypothetical paths for certain customers are shown in Fig. 3a. For example, the hypothetical path h_1 composed by the customer $\hat{e}_{12}$ connected to the element $\hat{e}_3$ and the element $\hat{e}_3$ is connected to the transformer $\hat{e}_1$; or the hypothetical path h_3 composed by the customer $\hat{e}_{11}$ connected to the element $\hat{e}_4$, and following in order: $\hat{e}_5$, $\hat{e}_9$, $\hat{e}_6$, $\hat{e}_7$, $\hat{e}_8$, $\hat{e}_3$ and then the transformer $\hat{e}_1$.

To identify the set of hypothetical paths compatible with the well-defined information, $\mathcal{H}^{\mathcal{I}'}$, the process of reduction is

performed as follows. Starting from the set of hypothetical paths $\mathcal{H}^{i'_1}$, the process sequentially excludes the paths that are not compatible with the remaining components of the well-defined information, $i'_j \in \mathcal{I}', 1 < j < |\mathcal{I}'|$.

Therefore, we determine the set $\mathcal{H}^{i'_1,i'_2}$ by excluding the paths $h \in \mathcal{H}^{i'_1}$ that are not compatible with the condition of customer's next connection as specified by the well-defined information i'_2.

For example, the hypothetical path h_1, in Fig. 3a, is excluded since the customer, $\hat{e}_{12}$, next connection is the element $\hat{e}_6$, and not the element $\hat{e}_3$. The size of the set $\mathcal{H}^{i'_1,i'_2}$ is given by:

$$|\mathcal{H}^{i'_1,i'_2}| = 11742. \tag{21}$$

Following, we determine the set $\mathcal{H}^{i'_1,i'_2,i'_3}$ by excluding paths $h \in \mathcal{H}^{i'_1,i'_2}$ that are not compatible with the condition of proximity as specified by the well-defined information i'_3.

For example, the path h_2, in Fig. 3a, is excluded since the distance between the elements $\hat{e}_3$ and $\hat{e}_4$ is greater than R. The value of R, that represents the maximum distance for the connection between two elements, is set to 20 metres. In this case, the value of $R = 20$ metres is chosen based on the assumption that electrical connections between elements in the network are typically within close proximity.

$$|\mathcal{H}^{i'_1,i'_2,i'_3}| = 6 \tag{22}$$

Finally, we determine the set $\mathcal{H}^{i'_1,i'_2,i'_3,i'_4}$ excluding the paths, $h \in \mathcal{H}^{i'_1,i'_2,i'_3}$ that are not compatible with the condition of length of path as specified by the well-defined information i'_4. For example, the path h_3, in Fig. 3a, is excluded because the length of the path is greater than the L. The value of L, that to represent the maximum length of a path, is set to 400 metres. The value $L = 400$ metres is chosen to ensure that selected paths are relatively direct and minimise energy loss due to excessive total lengths.

$$|\mathcal{H}^{i'_1,i'_2,i'_3,i'_4}| = 4 \tag{23}$$

After the process, the set of hypothetical paths is reduced to the paths compatible with the well-defined information. Therefore, the set of hypothetical paths compatible with the well-defined information for the different customers is given by:

$$\mathcal{H}^{\mathcal{I}'} = \mathcal{H}^{\mathcal{I}'}_{\hat{e}_{10}} \cup \mathcal{H}^{\mathcal{I}'}_{\hat{e}_{11}} \cup \mathcal{H}^{\mathcal{I}'}_{\hat{e}_{12}}, \text{ with} \tag{24}$$

$$\mathcal{H}^{\mathcal{I}'}_{\hat{e}_{10}} = \{h_6\} \tag{25}$$

$$\mathcal{H}^{\mathcal{I}'}_{\hat{e}_{11}} = \{h_7\} \tag{26}$$

$$\mathcal{H}^{\mathcal{I}'}_{\hat{e}_{12}} = \{h_4, h_5\}. \tag{27}$$

E. Diagnostic function

The found solution, $\mathcal{H}^{\mathcal{I}'}$, is evaluated with a $Diagnostic()$ function. This function can process both raw and well-defined information, and it assesses the set $\mathcal{H}^{\mathcal{I}'}$ for any potential issues that may be present.

One possible validation involves checking the number of active paths for each element of type $customer$ as specified by the raw information i_5. This validation is performed by examining the paths for each customer and taking into account the diverse combinations of switch(es) status. This involves examining all possible combinations of switch(es) status and verifying that there is only one active path for each customer, regardless of the switch(es) configuration(s).

In this case, during the analysis of the identified paths, it is observed that there are multiple active paths for some customers, as is possible to see in the set $\mathcal{H}^{\mathcal{I}'}_{\hat{e}_{12}}$ in Eq. (27). To address this issue, one solution is to collect more data, for example with a technical meeting with the DSO. Therefore, after the meeting with the DSO, a new important piece of information is collected and added to the set of information. This new piece of information is expressed as follows.

i_6 : one element of type $switch$ is needed for the customer $\hat{e}_{12}$ to disconnect it from the element $\hat{e}_1$.

A new transformation function, f_5, is added to process the new information, such that the new well-defined information i'_5 is given as follows.

$i'_5 = f_5(i_6)$: a new element, $\hat{e}_{13}$, of type $switch$ is added to the set of elements $\hat{\mathcal{E}}$. The location of this new element is chosen to be between the element $\hat{e}_6$ and $\hat{e}_7$ with status open. In this way, the customer $\hat{e}_{12}$ has only one active path.

Given the new well-defined information i'_5, the entire procedure is executed one more time, resulting in a new approximation of hypothetical paths, $\mathcal{H}^{\mathcal{I}'}$ is found. Following the modification of the transformation functions, the solution is re-evaluated with the $Diagnostic()$ function: the solution is found acceptable, and there is no more information available, so the algorithm is stopped.

The set $\mathcal{H}^{\mathcal{I}'}$ is considered as the best possible approximation of the set of real paths $\mathcal{P}$, given the information available to the DSO. The approximation found denoted as $\hat{\mathcal{P}}$ is considered the solution for the TPI problem.

F. Active and backup paths

Once the solution for the TPI problem is identified, denoted as $\hat{\mathcal{P}}$, it is possible to extract the active and backup paths for each customer. For example, considering the customer $\hat{e}_{12}$ we denote its active and backup path as $\hat{\mathcal{P}}_{a,\hat{e}_{12}}$ and $\hat{\mathcal{P}}_{b,\hat{e}_{12}}$ respectively. These paths are highlighted in Fig. 3c with the green colour representing the active path and the blue colour representing its possible backup path. More details about the customers paths can be found in Appendix Sections C and D.

VII. SCALABILITY DISCUSSION

The methodology proposed in Section V for accurately identifying paths in distribution networks relies on the generation of a large set of hypothetical paths and subsequently excluding the paths that are not compatible with the well-defined information. Even if this approach is effective in identifying the customer paths for small-scale networks, it becomes computationally expensive for larger networks due to the exponential growth in the number of hypothetical paths. The total number of hypothetical paths is given by Eq. 18. The presence of a factorial term in the equation

indicates a rapid growth in the number of hypothetical paths to consider, resulting in an increase of computational cost as the number of network elements increases.

To address the scalability challenges, we propose an alternative approach that we call the expanding path set (EPS) technique. This technique focuses on recursively expanding the set of hypothetical paths: starting with a small set of hypothetical paths and iteratively growing the set based on the available well-defined information. A possible implementation of this alternative approach is detailed by the algorithm given in Appendix Section E.

While both algorithms, the excluding and expanding one, come up with the same solution, the algorithm growing the set is more efficient since only the paths that are compatible with the well-defined information are considered.

VIII. CONCLUSION

In conclusion, this paper introduced a systematic procedure to address the complex challenge of topological path identification (TPI) within power distribution networks. By leveraging a methodology that integrates raw information, transformation functions that transform the raw information into well-defined information, and hypothetical paths compatible with the well-defined information, the proposed approach effectively identifies the customer paths within the power distribution network. This systematic procedure bridges the gap between data scientists and network experts, facilitating their collaboration in resolving the TPI problem in a consistent and efficient manner. As a result, this work significantly contributes to enhancing the efficiency and reliability of power distribution network modelling and management.

The future works for this paper are to apply our procedure to a real distribution network in Belgium. Additionally, future research could explore methods for identifying more efficient ways to determine parameters such as the maximum distance R and the total path length L, potentially considering distributions of values rather than single values, employing suitable probability distribution functions. This would allow for a more systematic and adaptable approach to path identification in real-world scenarios.

REFERENCES

[1] A. Navarro-Espinosa and L. F. Ochoa, "Reconstruction of low voltage distribution networks: From GIS data to power flow models," in *23rd International Conference on Electricity Distribution*, Jun. 2015.

[2] A. Guzmán, A. Argüello, J. Quirós-Tortós, and G. Valverde, "Processing and correction of secondary system models in geographic information systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3482–3491, 2019. DOI: 10.1109/TII.2018.2876356.

[3] J. Kays, A. Seack, T. Smirek, F. Westkamp, and C. Rehtanz, "The generation of distribution grid models on the basis of public available data," *IEEE Transactions on Power Systems*, vol. 32, no. 3, pp. 2346–2353, 2017. DOI: 10.1109/TPWRS.2016.2609850.

[4] G. Zepu, L. Yongjian, X. Ziwei, Y. Yilan, and Z. Lianmei, "Knowledge graph-based method for identifying topological structure of low-voltage distribution network," *The Journal of Engineering*, vol. 2020, no. 12, pp. 1177–1184, 2020. DOI: https://doi.org/10.1049/joe.2019.1319.

[5] A. Benzerga, D. Maruli, A. Sutera, A. Bahmanyar, S. Mathieu, and D. Ernst, "Low-voltage network topology and impedance identification using smart meter measurements," in *2021 IEEE Madrid PowerTech*, 2021, p. 6. DOI: 10.1109/PowerTech46648.2021.9495093.

[6] S. Bolognani and L. Schenato, "Identification of power distribution network topology via voltage correlation analysis," in *52nd IEEE Conference on Decision and Control*, 2013, pp. 1659–1664. DOI: 10.1109/CDC.2013.6760120.

[7] S. de Jongh, F. Mueller, F. Osterberg, C. A. Cañizares, T. Leibfried, and K. Bhattacharya, "Data-driven topology and parameter identification in distribution systems with limited measurements," *arXiv preprint arXiv:2308.09521*, Aug. 2023.

APPENDIX

A. Real network elements

Table II provides a detailed overview of all the information regarding the elements in the real network, including their type, coordinates, and status. It is possible to note that some elements of type $line$ can have more than two pairs of coordinates, as in the case of element e_{13}.

TABLE II
REAL NETWORK ELEMENTS AND THEIR ATTRIBUTES

ID	Coordinates	Type	Status
e_1	(232, 299)	transformer	-
e_2	(399, 196)	transformer	-
e_3	(232, 299), (241, 298)	line	-
e_4	(241, 298), (327, 297)	line	-
e_5	(327, 297), (324, 287)	line	-
e_6	(388, 295), (393, 292)	line	-
e_7	(388, 295), (390, 193)	line	-
e_8	(390, 193), (399, 196)	line	-
e_9	(390, 193), (376, 191)	line	-
e_{10}	(376, 191), (315, 191)	line	-
e_{11}	(315, 191), (295, 189)	line	-
e_{12}	(295, 189), (299, 207)	line	-
e_{13}	(295, 189), (268, 169), (235, 190)	line	-
e_{14}	(235, 190), (239, 282)	line	-
e_{15}	(239, 282), (241, 298)	line	-
e_{16}	(376, 191)	switch	open
e_{17}	(315, 191)	switch	open
e_{18}	(239, 282)	switch	close
e_{19}	(324, 287)	customer	-
e_{20}	(393, 292)	customer	-
e_{21}	(299, 207)	customer	-

B. DSO network elements

Table III shows all the details about the elements initially available to the DSO. It is possible to note that elements of type $line$ have only two pairs of coordinates, since intermediate points are generally not known by DSOs.

C. Active paths

Table IV shows all the active paths for each customer. Moreover, Fig. 4, 5 and 6 show the hypothetical active paths compatible with the well-defined information for the customers $\hat{e}_{10}$, $\hat{e}_{11}$ and $\hat{e}_{12}$ respectively.

D. Backup paths

Table V showing all the backup paths for each customer. Fig. 7 shows the hypothetical backup paths compatible with the well-defined information for the customer $\hat{e}_{12}$.

ID	Coordinates	Type	Status
$\hat{e}_1$	(230, 297)	transformer	-
$\hat{e}_2$	(401, 194)	transformer	-
$\hat{e}_3$	(240, 297), (328, 296)	line	-
$\hat{e}_4$	(388, 295), (391, 193)	line	-
$\hat{e}_5$	(390, 193), (374, 189)	line	-
$\hat{e}_6$	(374, 189), (310, 191)	line	-
$\hat{e}_7$	(293, 189), (235, 189)	line	-
$\hat{e}_8$	(233, 189), (242, 282)	line	-
$\hat{e}_9$	(374, 189)	switch	close
$\hat{e}_{10}$	(321, 289)	customer	-
$\hat{e}_{11}$	(393, 292)	customer	-
$\hat{e}_{12}$	(320, 196)	customer	-

TABLE IV

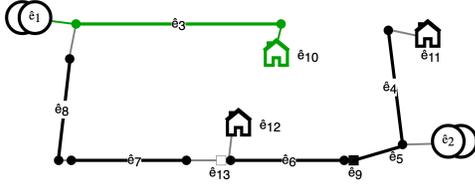| Customer ID	Paths
$\hat{e}_{10}$	$(\hat{e}_{10}, \hat{e}_3, \hat{e}_1)$
$\hat{e}_{11}$	$(\hat{e}_{11}, \hat{e}_4, \hat{e}_2)$
$\hat{e}_{12}$	$(\hat{e}_{12}, \hat{e}_6, \hat{e}_9, \hat{e}_5, \hat{e}_4, \hat{e}_2)$

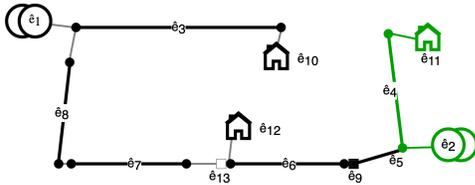Fig. 4. Active path for customer $\hat{e}_{10}$

Fig. 5. Active path for customer $\hat{e}_{11}$

Fig. 6. Active path for customer $\hat{e}_{12}$

TABLE V

Customer ID	Paths
$\hat{e}_{10}$	None
$\hat{e}_{11}$	None
$\hat{e}_{12}$	$(\hat{e}_{12}, \hat{e}_6, \hat{e}_{13}, \hat{e}_7, \hat{e}_8, \hat{e}_3, \hat{e}_1)$

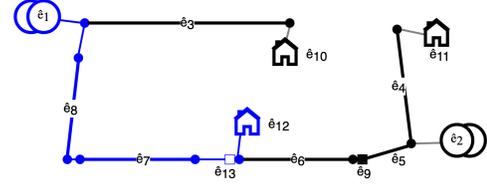
Fig. 7. Backup path for customer $\hat{e}_{12}$

E. Expansion path set algorithm

The expansion path set (EPS) algorithm works as follows. For each customer, the $IdentifyPaths()$ function initialises an empty path and iteratively explores the paths considering adjacent elements. Paths that fail to meet the constraints specified by the well-defined information of the academic example explained in Section VI (distance between elements, total path length limits) are not further expanded. If an element of type $transformer$ is reached, the algorithm is stopped, and the path is returned. Otherwise, the algorithm recursively explores new paths invoking the same function with the updated parameters.

The recursiveness of the algorithm operates similar to an expansion tree: as the algorithm examines each element, it branches out to explore all possible connections that lead to paths that are compatible with the set of well-defined information.

Function: $IdentifyPaths(current\_element, current\_path, R, L)$
  Inputs:
- $current\_element$: the element to consider as a starting point to find its path(s) to the transformer(s).
- $current\_path$: the current path being considered.
- R: the maximum distance criterion for path compatibility.
- L: the maximum path length criterion for path compatibility.

  Output:
- $candidate\_paths$: list of paths compatible with the set of well-defined information.

1: $candidate\_paths$ = []
2: **for all** $next\_element$ in $\left(\hat{\mathcal{E}} - Subset(\hat{\mathcal{E}}, customer)\right)$ **do**
3: **if** $next\_element$ not in $current\_path$ **then**
4:   $distance = Dist(current\_element, next\_element)$
5:   $new\_path = current\_path + [next\_element]$
6:   $length = LengthPath(new\_path)$
7:   **if** $(distance < R$ and $length < L)$ **then**
8:    **if** $next\_element.type = transformer$ **then**
9:     $candidate\_paths.add(new\_path)$
10:   **else**
11:    $paths = IdentifyPaths(next\_element, new\_path, R, L)$
12:    $candidate\_paths.add(paths)$
13: **return** $candidate\_paths$