Highlights

An Optimization Algorithm for Customer Topological Paths Identification in Electrical Distribution Networks

VASSALLO Maurizio,LEERSCHOOL Adrien,BAHMANYAR Alireza,DUCHESNE Laurine,GERARD Simon,WEHENKEL Thomas,ERNST Damien

- The paper proposes an ILP-based approach for identifying LV network topological paths.
- Only static data is used such as GIS data of the elements in the network and customer-transformer connection info.
- Addresses GIS data inaccuracies and identifies accurate customer topological paths.
- Provides a solution to support DSOs in digitalizing distribution networks, suitable in the case of a lack of smart meter measurements.

An Optimization Algorithm for Customer Topological Paths Identification in Electrical Distribution Networks

VASSALLO Maurizio[a,*,1], LEERSCHOOL Adrien[b,*,1], BAHMANYAR Alireza[b], DUCHESNE Laurine[b], GERARD Simon[c], WEHENKEL Thomas[c] and ERNST Damien[a,b]

[a] *Department of Electrical Engineering and Computer Science, University of Liège, Belgium*
[b] *Intelligent Systems Solutions, Haulogy, Neupré, Belgium*
[c] *RESA, Liège, Belgium*

ARTICLE INFO

Keywords:
Topological path identification
Electrical distribution network
Low network observability
Integer linear programming
Digital twin.

ABSTRACT

Accurate identification of network topologies in electrical distribution networks is critical for distribution system operators (DSOs) to perform accurate grid management and planning. However, this task is generally challenging due to inaccurate and limited data available to the DSOs, such as incorrect geographic information system (GIS) records and the lack of advanced metering infrastructure (AMI). This paper proposes a novel integer linear programming (ILP) optimization framework for identifying customer topological paths using only static data, including GIS record and customer-to-transformer connections. The proposed approach is demonstrated on both an academic and a real-world distribution network, showing its ability to handle data inaccuracies. Results indicate that the proposed method provides a robust and practical solution for DSOs to develop accurate digital twins and improve grid visibility, even in the absence of AMI data. This work offers a valuable tool for the digitalization of power distribution systems.

1. Introduction

Power distribution networks are the infrastructure responsible for delivering electricity to consumers. As distribution networks grow in complexity and scale, driven by factors such as the accommodation of new customers and the increased integration of distributed energy resources (DERs), the need for improved network planning and operation strategies becomes essential. A significant challenge in this context is identifying the specific routes electricity takes to reach customers, a task referred to as topological path identification (TPI). Accurate knowledge of network topologies and customer paths is often unavailable to distribution system operators (DSOs) due to incomplete, outdated, or erroneous data records.

Solving the TPI problem provides DSOs with a clear knowledge about the connectivity between customers and medium-voltage/low-voltage (MV/LV) substations. However, achieving accurate identification of customer topological paths remains a challenging task due to several factors. Firstly, DSOs frequently rely on incomplete or inaccurate geographic information system (GIS) data, which may contain errors derived from outdated recordings or limitations in GPS technologies ([1]). Secondly, while advanced metering infrastructure (AMI) devices are increasing the integration in power systems, many European countries still lack comprehensive AMI coverage ([2]), limiting the feasibility of measurement-based methods. Finally, scalability poses a significant obstacle; as networks grow in size and complexity, traditional approaches often fail to provide efficient and scalable solutions.

Despite these challenges, the accurate identification of customer topological paths is fundamental to numerous applications in modern power systems. It serves as a cornerstone for building a digital twin of the network, enabling DSOs to perform advanced simulations and analyses without risking disruptions to the physical infrastructure. For instance, digital twins can be used to simulate scenarios such as hosting capacity—the maximum integration of technologies like DERs—allowing DSOs to evaluate and prioritize investment and operational decisions effectively.

[*] Authors contributed equally to the paper

mvassallo@uliege.be (V. Maurizio); adrien.leerschool@haulogy.net (L. Adrien); alireza.bahmanyar@haulogy.net (B. Alireza); laurine.duchesne@haulogy.net (D. Laurine); simon.gerard@resa.be (G. Simon); thomas.wehenkel@resa.be (W. Thomas); dernst@uliege.be (E. Damien)

ORCID(s): 0009-0009-1577-4256 (V. Maurizio)

[1] Corresponding author

1.1. Literature review

In power distribution networks, various methodologies have been proposed to address the network topology identification or the TPI problem specifically. These methods can be characterized into two main categories: methods that only rely on static data such as geographic information system (GIS) data of the elements in the network, and methods that rely on AMI data such as smart meter recordings. Static data-based methods generally focus on establishing relationships and connections between elements. Relevant studies include [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. For example, in [3] and [4] the authors identify the network topology of different networks using publicly available geographical data and connecting the elements together based on their distances. The authors in [5] focus on network topology identification using Euclidean distances and breadth-first search to identify clusters in the network, followed by refining the topology by connecting nodes to the most probable lines. The paper [6] presents a kd-tree algorithm for identifying clusters of elements, followed by connecting the elements within each cluster based on a predefined distance. In [7], the authors present a new procedure to exploit graph theory and data structure properties to efficiently detect and correct errors in models of radial secondary systems. The paper [8] uses knowledge graphs to reflect the existing relationship among the elements in low-voltage distribution networks. The knowledge graphs are then used to identify the correct network topology.

The main advantage of these methods is that they operate without relying on AMI measurements, making them suitable for DSOs with limited advanced infrastructure. However, their major limitation is that GIS data can be incomplete, inaccurate, or missing for certain areas of the networks.

Dynamic data-based solutions rely on AMI measurements from the network. Relevant studies include [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. For example, in [13], the authors use line current sensor measurements and a mixed-integer linear programming (MILP) optimization algorithm to identify the topology of a power distribution network in California. The authors in [14] introduce an iterative methodology for reconstructing network topology and cable parameters of low-voltage three-phase networks with limited smart meter coverage. The authors in the paper [15] propose a similar approach using mixed-integer quadratic programming (MIQP) to identify the topology of power distribution networks using smart sensor recordings. While AMI-based methods can estimate topology and line parameters (e.g., impedance), they often require extensive sensor coverage, limiting their applicability in networks with incomplete AMI deployment. This limitation restricts their applicability in networks with incomplete AMI coverage, as evidenced by some studies using limited smart meter data [24, 25, 26].

1.2. Paper contribution

This paper presents an optimization algorithm to assist DSOs in addressing the TPI problem using only the available static data. Our methodology relies exclusively on GIS data of network elements and customer connections to MV/LV transformers, making it particularly suitable for networks where GIS data is incomplete or inaccurate and AMI coverage is limited. The problem is formulated as an integer linear programming (ILP) optimization algorithm. The objective function of the optimization problem focuses on maximizing the number of customers connected to the correct MV/LV transformer, while some constraints ensure that the solution satisfies the DSO's expectations. The proposed methodology effectively addresses data inaccuracies and incompleteness.

Moreover, a repository with the code of the proposed methodology is released. The repository is publicly available at the following link: https://github.com/TPIproblem/OptimalTPI.

1.3. Paper organization

The rest of the paper is organized as follows: Section 2 presents the definition of the power network elements considered in this work. Section 3 defines the concept of topological paths. The problem statement is presented in Section 4. Section 5 defines the methodology used to solve the problem. The methodology is applied to an academic example in Section 6 and a real Belgian power distribution network in Section 7. Section 8 concludes the work and discusses possible future works.

2. Power distribution modeling

Power distribution networks typically encompass a wide range of elements [27]. The set of elements $\mathcal{E}$ is defined as:

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

where $|\mathcal{E}|$ represents the total number of elements in the set $\mathcal{E}$ and a single element is denoted as e_k, with $k \in \{1, ..., |\mathcal{E}|\}$.

Each element can have some attributes, such as its type and coordinates. The set of all element attributes $\mathcal{A}$ is defined as follows:

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

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

$$e = \{e.a_1, ..., e.a_k, ..., e.a_{|e|}\} \tag{3}$$

where the dot notation is used to access the attributes of the element e, and a_k represents, for example, the type of the element e, its coordinates, or other relevant properties.

Each element within the network can be categorized 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 $\mathcal{E}$:

$$\mathcal{T} = \{t_1, ..., t_{|\mathcal{T}|}\}. \tag{4}$$

Among the typical types, there may be *customer* connection elements, MV/LV *transformer* substations.
Given an element $e \in \mathcal{E}$, its type, $t \in \mathcal{T}$, is accessed as follows:

$$t = e.type. \tag{5}$$

2.1. Network topological functions

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 $t \in \mathcal{T}$. Formally:

$$Subset(\mathcal{E}, t) = \{e \in \mathcal{E} \mid e.type = t\}. \tag{6}$$

In particular, three useful subsets are defined:

- $\mathtt{C}$: represents the set of customers in the network, $\mathtt{C} = Subset(\mathcal{E}, customer)$.
- $\mathtt{T}$: represents the set of terminal elements, generally $\mathtt{T} = Subset(\mathcal{E}, transformer)$. Note that the terminal element can vary depending on the context. For example, in this paper, we consider the terminal element to be the feeder junction where the feeder starts in the MV/LV substation, but it could also be extended to the high-voltage/MV substation or any other element in the network. The methodology is designed to be flexible.
- $\mathtt{R}$: represents the set of all elements, excluding customers and terminal elements, $\mathtt{R} = \mathcal{E} - \mathtt{C} - \mathtt{T}$.

The $Dist()$ function takes two elements $e_k, e_m \in \mathcal{E}$ as input, and it outputs a scalar value, d, representing the distance between them. The $Dist()$ function is defined as follows:

$$d = Dist(e_k, e_m) = ||e_k.coor,\ e_m.coor||_2 \tag{7}$$

where $||\bullet||_2$ represents the Euclidean distance and where $e_k.coor$ and $e_m.coor$ give the coordinates of the element e_k and, e_m respectively.

The $Connections()$ function identifies all elements in the set $\mathcal{E}$ that can be connected to a given element, $e \in \mathcal{E}$, considering some specific conditions. The conditions for connectivity can depend on the DSOs' requirements such as distance between the elements, the element types, and so on.
Therefore, the $Connections()$ function takes three inputs: $\mathcal{E}$, e and some conditions, returning a subset $\mathtt{K} \subset \mathcal{E}$. Formally:

$$\mathtt{K} = \text{Connections}(\mathcal{E}, e, \text{conditions}) = \{e_m \in \mathcal{E} \mid e_m \text{ can be directly connect to } e \text{ if the conditions hold}\}.$$

3. Topological paths

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

3.1. Paths

A customer topological path is represented as an ordered list of elements:

$$p = (p_1, p_2, ..., p_k, ..., p_{|p|-1}, p_{|p|}) \tag{8}$$

where p is a generic path, p_1 represents the initial element in the path which is a customer; p_k represents the k-element in the path, and $p_{|p|}$ represents the terminal element, generally an MV/LV transformer.
Therefore, $p_1 \in \mathtt{C}$, and $p_k \in \mathtt{R}$ with $k \in \{2, ..., |p| - 1\}$ and $p_{|p|} \in \mathtt{T}$.

3.2. Hypothetical paths

Hypothetical paths refer to routes where only the initial elements are known, while the connections between intermediate elements up to the terminal point are unknown. These paths represent potential ways that electricity might follow to supply a customer.
The set of hypothetical paths, denoted as $\mathcal{H}$. Each hypothetical path, $h \in \mathcal{H}$, represents a possible supply route for a customer. The number of hypothetical paths, given the set of elements $\mathcal{E}$, is finite, and it is given by:

$$|\mathcal{H}| = |\mathtt{C}| \cdot \sum_{k=1}^{|\mathtt{R}|} {}^{|\mathtt{R}|}P_k \cdot |\mathtt{T}| \tag{9}$$

where ${}^{n}P_m$ is the permutation formula, which calculates the number of ways to arrange k items from a set of n items (${}^{n}P_m = \frac{n!}{(n-m)!}$).

3.3. Real paths

The set of all real paths, denoted as $\mathcal{P}$, is a subset of the hypothetical path set $\mathcal{H}$. Each path $p \in \mathcal{P}$ represents the actual sequence of network elements that connect a customer to an MV/LV transformer.
In radial networks, the number of real paths is equal to the number of customers, since one customer is supplied by only one MV/LV transformer at any given moment. Therefore, the following condition holds $|\mathcal{P}| = |\mathtt{C}|$, where $\mathtt{C}$ represents the set of elements of type customers.

3.4. Estimated paths

The estimated paths, denoted by $\hat{\mathcal{P}}$, form a subset of the hypothetical paths, $\mathcal{H}$. This set of estimated paths, $\hat{\mathcal{P}}$, is an approximation to the set of real paths within the network, $\mathcal{P}$.
The set $\hat{\mathcal{P}}$ holds the property that the number of estimated paths in the set is equal to the number of customers considered in the network. Therefore, the following condition holds $|\hat{\mathcal{P}}| = |\mathcal{P}| = |\mathtt{C}|$.

3.5. Paths visualization

Figure 1 illustrates a simplified network, highlighting hypothetical, real, and estimated paths.
As shown in Fig. 1a the data about network elements are often incomplete, making it difficult to accurately identify the customer paths.
Figure 1b shows some of all hypothetical paths. In the absence of prior information, any path could potentially represent the real one. This highlights the challenge that with limited information, only limited conclusions can be drawn about whether a hypothetical path corresponds to its real one. Many hypothetical paths for each customer may be available.
Figures 1c and 1d illustrate the real and estimated paths for each customer, respectively. Each customer is associated with exactly one real path and one estimated path, both of which are subsets of the hypothetical path set. Therefore: $\mathcal{P} \subset \mathcal{H}$ and $\hat{\mathcal{P}} \subset \mathcal{H}$.
Figure 2 presents a Venn diagram illustrating the relationships among the path sets. The set of hypothetical paths, $\mathcal{H}$, encompasses all other sets, representing all possible paths within the network. Notably, the sets of real paths, $\mathcal{P}$, and estimated paths, $\hat{\mathcal{P}}$, overlap to a certain degree, highlighting similarities and discrepancies between the estimation and ground truth. The overlap between $\mathcal{P}$ and $\hat{\mathcal{P}}$, indicates that some estimated paths match the real ones while some others do not. For example, the paths $\hat{p}_1$ and p_1 for customer e_1, match perfectly. However, data inaccuracies can result in discrepancies, for example, the estimated path $\hat{p}_2$ deviates from the real path p_2 for the customer e_2.

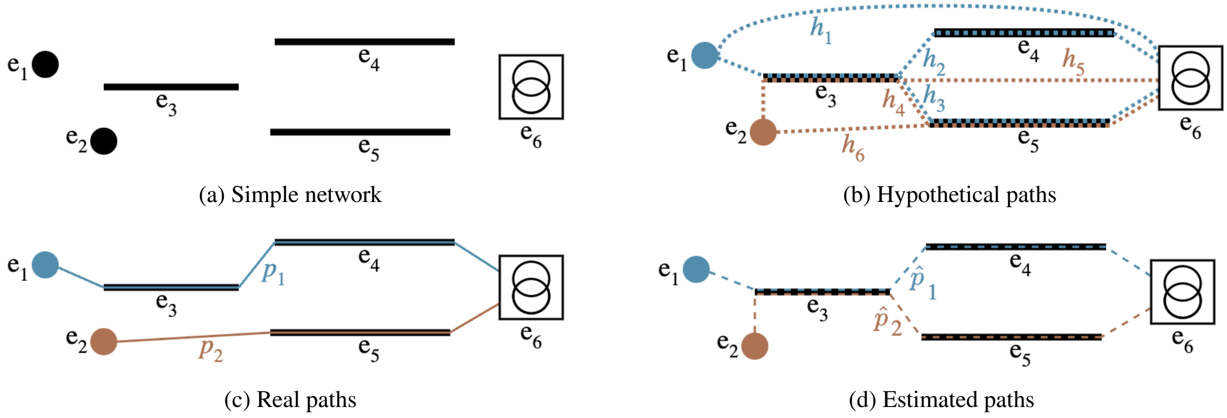


Figure 1: Visualization of the different sets of paths.

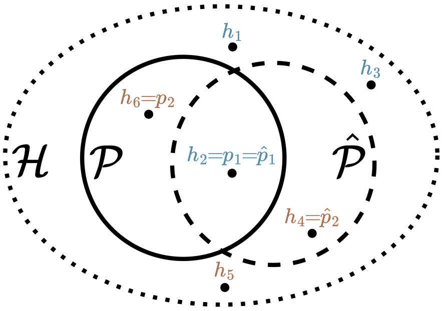


Figure 2: Representation of the different sets and their relationships.

4. Problem statement

Generally, DSOs possess different types of information. The information available to the DSO is denoted as $\mathcal{I}$. This data can include, for example, files about the GIS data of the network elements, customers' information like annual power consumption, network configuration rules explaining how the elements are connected, and more.
The goal of the DSOs is, given all the raw information available to them, to estimate the customer topological paths that are as close as possible to the real paths. Denoting the set of real paths as the set $\mathcal{P}$ and the set of estimated paths as $\hat{\mathcal{P}}$, the objective is to identify two sets to be as close as possible and in the best case equal ($\hat{\mathcal{P}} = \mathcal{P}$).
Therefore, the problem is to identify a methodology, $\mathfrak{M}()$, that, starting from the set of raw information, $\mathcal{I}$ can identify the best approximation of the real paths $\mathcal{P}$. Formally:

$$\hat{\mathcal{P}} = \mathfrak{M}(\mathcal{I}) \quad \text{s.t.} \quad \hat{\mathcal{P}} \simeq \mathcal{P}. \tag{10}$$

5. Methodology

This section outlines the step-by-step process of the proposed methodology, $\mathfrak{M}()$, designed to identify customer topological paths in electrical distribution networks. A visual representation of the methodology is provided in Fig. 3.

5.1. Raw information listing

Raw information refers to available information or data that come from various sources. The set of raw information is listed as follows: $\mathcal{I} = \{i_1, ..., i_k, ..., i_{|\mathcal{I}|}\}$, with i_k ($k \in \{1, ..., |\mathcal{I}|\}$) denoting a single piece of raw information.

For this paper, we use only static data, and in particular, the data needed is the GIS coordinates of the different elements in the network, their type, and the feeder terminal junction to which each customer is connected. Generally, this kind

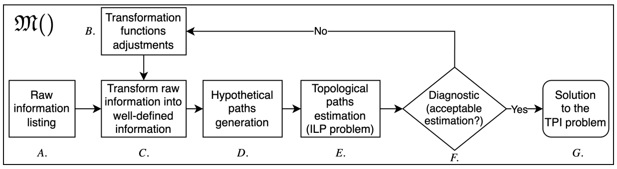

$\mathfrak{M}()$
B.
Transformation functions adjustments
No
Raw information listing
Transform raw information into well-defined information
Hypothetical paths generation
Topological paths estimation (ILP problem)
Diagnostic (acceptable estimation?)
Yes
Solution to the TPI problem
A.
C.
D.
E.
F.
G.

Figure 3: Flowchart of the steps proposed by our methodology, $\mathfrak{M}()$, to identify the customer topological paths in electrical distribution networks.

of information is available to the DSOs, even if it may be incomplete and/or inaccurate.
Formally, the data needed for this paper is:

- The set of elements, $\mathcal{E}$.
- The set of attributes, $\mathcal{A}$, contains at least the coordinates of the elements, the types, and the feeder terminal junction to which each customer is connected.
- The set of types $\mathcal{T}$, with $\mathcal{T}$ containing at least customer, line, junction, and transformer.

Therefore, $e.coor$, $e.type$ must be known for all the elements, $e \in \mathcal{E}$ and $c.junction$ is known for all the customers, $c \in \mathbb{C}$.

5.2. Transformation functions adjustments

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: $\mathcal{F} = \{f_1, ..., f_m, ..., f_{|\mathcal{F}|}\}$, with f_m $(m \in \{1, ..., |\mathcal{F}|\})$ denoting a single transformation function.
Mathematically, the transformation functions take one piece of raw information as input and return the transformed information as output.

5.3. Well-defined information

The well-defined information forms the basis to provide a structured foundation to identify customer paths that are as close as possible to reality. The set of well-defined information is denoted as follows: $\mathcal{I}' = \{i'_1, ..., i'_n, ..., i'_{|\mathcal{I}'|}\}$, with i'_n $(n \in \{1, ..., |\mathcal{I}'|\})$ denoting a single piece of well-defined information. Each element in the set $\mathcal{I}'$ is the result of applying a transformation function on a piece of raw information, i.e. $i'_n = f_m(i_k)$.

5.4. Hypothetical paths generation

Given the set of well-defined information, $\mathcal{I}'$, the set of hypothetical path compatible with the set $\mathcal{I}'$, denoted as $\mathcal{H}^{\mathcal{I}'}$, is constructed. Each piece of information in $\mathcal{I}'$ allows to reduce the size of the hypothetical path set $\mathcal{H}^{\mathcal{I}'}$.
The idea is that the more data is available, the closer to reality the paths inside the set $\mathcal{H}^{\mathcal{I}'}$ are.

5.5. Topological paths estimation

After generating the hypothetical paths compatible with the well-defined information, $\mathcal{H}^{\mathcal{I}'}$, the next step of the methodology is to identify the paths in the set $\mathcal{H}^{\mathcal{I}'}$ that are an approximation of the real paths in the actual network, therefore identifying the set of estimated paths $\hat{\mathcal{P}}$.
While various methods can solve the TPI problem, we present an ILP approach designed to find the solution $\hat{\mathcal{P}}$ and to meet the DSO requirements.

5.5.1. Matrices generation

The following sections describe the main components of our ILP formulation, including the matrices and the optimization problem.

- Hypothetical paths matrix:
  The set of hypothetical paths, $\mathcal{H}^{\mathcal{I}'}$ can be equivalently interpreted as a binary matrix with a number of rows equivalent

to $|\mathcal{H}^{\mathcal{I}'}|$ and a number of columns equivalent to $|\mathtt{C}| + |\mathtt{R}| + |\mathtt{T}|$.
Therefore, the hypothetical paths matrix, denoted as **H**, can be written as follows:

$$\mathbf{H} = \left(\begin{array}{c|cccc|cccc|cccc|}
 & c_1 & c_2 & \cdots & c_{|\mathtt{C}|} & r_1 & r_2 & \cdots & r_{|\mathtt{R}|} & t_1 & t_2 & \cdots & t_{|\mathtt{T}|} \\ \hline
h_1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 \\
h_2 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
h_k & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
h_{|\mathcal{H}^{\mathcal{I}'}|} & 0 & 0 & \cdots & 1 & 0 & 1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \hline
 & \multicolumn{4}{c}{\mathbf{H}_{\mathtt{C}}} & \multicolumn{4}{c}{\mathbf{H}_{\mathtt{R}}} & \multicolumn{4}{c}{\mathbf{H}_{\mathtt{T}}}
\end{array}\right) \tag{11}$$

Each row h_k in the matrix **H** represents a hypothetical path compatible with the well-defined information. A value of 1 or 0 in this row indicates whether the corresponding element, $e \in \mathtt{C} \cup \mathtt{R} \cup \mathtt{T}$, is present (1) or not (0) in path h_k. A single element of the matrix is accessed as $\mathbf{H}_{(k,m)}$ where m represents the row ($k \in \{1, ..., |\mathcal{H}^{\mathcal{I}'}|\}$) and m represents the column ($m \in \{1, ..., |\mathtt{C}| + |\mathtt{R}| + |\mathtt{T}|\}$).

For clarity, we divide the **H** matrix into three sub-matrices:

- The $\mathbf{H}_{\mathtt{C}}$ matrix, denoted as customers elements matrix, indicates the presence of a customer $c \in \mathtt{C}$ in any path, $h \in \mathcal{H}^{\mathcal{I}'}$. Therefore, the $\mathbf{H}_{\mathtt{C}}$ matrix has dimension $|\mathcal{H}^{\mathcal{I}'}| \times |\mathtt{C}|$.
  Since one and only one customer can be present in one path, the following condition must be satisfied:

$$\sum_{k}^{|\mathtt{C}|} \mathbf{H}_{\mathtt{C}\,(m,k)}, \forall m \in \{1, ..., |\mathcal{H}^{\mathcal{I}'}|\}. \tag{12}$$

- The $\mathbf{H}_{\mathtt{R}}$ matrix, denoted as remaining elements matrix, indicates the presence of elements $r \in \mathtt{R}$ in the paths, $h \in \mathcal{H}^{\mathcal{I}'}$. Therefore, the $\mathbf{H}_{\mathtt{R}}$ matrix has dimension $|\mathcal{H}^{\mathcal{I}'}| \times |\mathtt{R}|$.
  No particular condition is imposed on the matrix $\mathbf{H}_{\mathtt{R}}$.
- The $\mathbf{H}_{\mathtt{T}}$ matrix, referred as terminal elements matrix, indicates the presence of a terminal node, $t \in \mathtt{T}$, in the paths, $h \in \mathcal{H}^{\mathcal{I}'}$. Therefore, the $\mathbf{H}_{\mathtt{T}}$ matrix has dimension $|\mathcal{H}^{\mathcal{I}'}| \times |\mathtt{T}|$.
  Since one and only one terminal element can be present in one path, the following condition must be satisfied:

$$\sum_{k}^{|\mathtt{T}|} \mathbf{H}_{\mathtt{T}\,(m,k)}, \forall m \in \{1, ..., |\mathcal{H}^{\mathcal{I}'}|\}. \tag{13}$$

- Terminal association matrix:
  We define a binary matrix, referred to as the terminal association matrix, denoted as $\mathbf{T}_{\mathtt{R}}$, which indicates the association of elements $r \in \mathtt{R}$ to a specific terminal element $t \in \mathtt{T}$.
  The $\mathbf{T}_{\mathtt{R}}$ matrix has dimension $|\mathtt{T}| \times |\mathtt{R}|$ and can be represented as follows:

$$\mathbf{T}_{\mathtt{R}} = \begin{array}{c|ccccccc|}
 & r_1 & r_2 & \cdots & r_m & \cdots & r_{|\mathtt{R}|-1} & r_{|\mathtt{R}|} \\
t_1 & 0 & 1 & \cdots & 0 & \cdots & 0 & 0 \\
t_2 & 1 & 0 & \cdots & 0 & \cdots & 1 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
t_k & 0 & 0 & \cdots & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
t_{|\mathtt{T}|} & 0 & 0 & \cdots & 0 & \cdots & 0 & 1
\end{array} \tag{14}$$

where 0 and 1 represent respectively whether an element $r \in \mathtt{R}$ is assigned to a terminal element $t \in \mathtt{T}$ or not.

- Estimated paths matrix:
  The solution of the TPI problem, the set $\hat{\mathcal{P}}$, can be also conceptualized as a binary matrix that represents the hypothetical paths estimated to be the real paths. This path matrix, denoted as $\hat{\mathbf{P}}$, has dimension $1 \times |\mathcal{H}^{\mathcal{I}'}|$.

$$\hat{\mathbf{P}} = \begin{array}{c} \begin{array}{ccccc} h_1 & \cdots & h_k & \cdots & h_{|\mathcal{H}^{\mathcal{I}'}|} \end{array} \\ \begin{pmatrix} 1 & \cdots & 1 & \cdots & 0 \end{pmatrix} \end{array} \tag{15}$$

where 0 and 1 represent whether the path $h_k \in \mathcal{H}^{\mathcal{I}'}$ is an estimated optimal path or not.

5.5.2. ILP optimization problem

The optimization problem is presented in Eqs. 16a–e.

$$\max_{\hat{\mathbf{P}},\mathbf{T}_{\mathtt{R}}} \; \omega \cdot \left(\sum_{k}^{|\mathcal{H}^{\mathcal{I}'}|} \hat{\mathbf{P}}_{(k)} \right) - \sum_{m}^{|\mathtt{T}|} \sum_{n}^{|\mathtt{R}|} \mathbf{T}_{\mathtt{R}\,(m,\,n)} \tag{16a}$$

$$\text{s.t.} \quad \sum_{k}^{|\mathtt{R}|} \mathbf{H}_{\mathtt{R}\,(m,\,k)} \cdot \hat{\mathbf{P}}_{(m)} \cdot \mathbf{H}_{\mathtt{T}\,(m,\,n)} \leq \left(\left(\mathbf{T}_{\mathtt{R}} \times \mathbf{H}_{\mathtt{R}}^{\mathtt{T}} \right) \cdot \mathbf{H}_{\mathtt{T}}^{\mathtt{T}} \right)_{(n,m)}, \quad \forall m \in \{1, ..., |\mathcal{H}^{\mathcal{I}'}|\}, \; \forall n \in \{1, ..., |\mathtt{T}|\} \tag{16b}$$

$$\left(\hat{\mathbf{P}} \times \mathbf{H}_{\mathtt{C}} \right)_{(k)} \leq 1, \quad \forall k \in \{1, ..., |\mathtt{C}|\} \tag{16c}$$

$$\sum_{k}^{|\mathtt{T}|} \mathbf{T}_{\mathtt{R}\,(k,\,m)} \leq 1, \quad \forall m \in \{1, ..., |\mathtt{R}|\} \tag{16d}$$

$$\hat{\mathbf{P}}_{(k)} \in \{0,1\}, \;\; \forall k \in \{1, ..., |\mathcal{H}^{\mathcal{I}'}|\}, \quad \mathbf{T}_{\mathtt{R}\,(m,\,n)}, \in \{0,1\} \;\; \forall m \in \{1, ..., |\mathtt{T}|\}, \; \forall n \in \{1, ..., |\mathtt{R}|\} \tag{16e}$$

The following section presents each component of our ILP formulation, as presented in Eq.s (16)a–e:

- Equation 16a aims to maximize the number of customers connected to their respective terminal elements by maximizing the number of estimated paths, $h \in \hat{\mathbf{P}}$, with value 1. Additionally, a penalty term on the terminal association matrix $\mathbf{T}_{\mathtt{R}}$ is included, proportional to the number of elements assigned to each terminal element, to prevent unnecessary assignments.
  The weight ω represents the trade-off between how important is to identify a path for a customer over how many elements are used in such a path.

- The constraint in Eq. 16b verifies the validity of the paths included in the matrix $\hat{\mathbf{P}}$. It ensures that for every estimated path, $h \in \hat{\mathcal{P}}$, all its elements $e \in h$ are assigned to the same terminal element, $t \in \mathtt{T}$.

- The constraint in Eq. 16c ensures that each customer, $c \in \mathtt{C}$, has at most one estimated path in $\hat{\mathbf{P}}$ with a value equal to 1. This constraint is designed to prevent a customer from being associated with more than one estimated path.

- Equation 16d guarantees that each element $r \in \mathtt{R}$ is associated to at most one specific terminal element $t \in \mathtt{T}$. Therefore, this constraint is designed to prevent the same elements from being associated with multiple terminal elements.

5.6. Diagnostic function

The set of estimated paths, $\hat{\mathcal{P}}$, identified is validated with a $Diagnostic()$ function. This function is an abstract function that can be customized to accommodate various validation criteria and methodologies and it serves as a critical tool for assessing the validity of the solution. The $Diagnostic()$ function evaluates the found paths by checking any possible discrepancies from reality.

The diagnostic function outputs a list of issues encountered in the solution proposed, if any. If some issues are identified, it is possible to adjust the transformation functions to address these issues. Subsequently, the methodology is rerun to incorporate these changes until the results are acceptable.

5.7. Solution of the TPI problem

If no issues are identified or the DSO considers the solution to be acceptable, then the set $\hat{\mathcal{P}}$ is regarded as the final solution to the TPI problem.

6. 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.

6.1. 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 may be missing, and some coordinates may be inaccurate.

i_2: MV/LV transformer cabin manuals provide instructions on cabin setup. This includes details about the organization of connection boards and how feeder lines connect to specific terminal elements or feeder terminal junctions on these boards.

i_3: Customers' information about to which feeder terminal junction in the MV/LV transformer they are connected.

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

i_5: Information that LV networks are operated radially.

6.2. Transformation functions

The set $\mathcal{F}$ enumerates the transformation functions. Given that there are five pieces of raw information, the number of transformation functions is also five ($|\mathcal{F}| = |\mathcal{I}| = 5$).

6.3. Well-defined information

The set of well-defined information is given by:

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

$i'_2 = f_2(i_2)$: The elements of two types $junction$ and $transformer$ are directly connected with no intermediate element. Since each feeder terminal junction has a known associated transformer, we can simplify the path identification problem. Instead of needing to track the path from the customer to the transformer, we only need to consider the path from the customer to the MV/LV transformer of the feeder junction terminal.

$i'_3 = f_3(i_3)$: For each element of type $customer$, $e \in \mathcal{E}$, its attribute $junction$, $c.junction$, where it is connected to is known.

$i'_4 = f_4(i_4)$: An element can be connected to another only if its distance is less than a given distance D. Moreover, the total length of a path must not exceed a maximum length L.

$i'_5 = f_5(i_5)$: Since the network has a radial structure and does not have loops, each element can only appear once in each path.

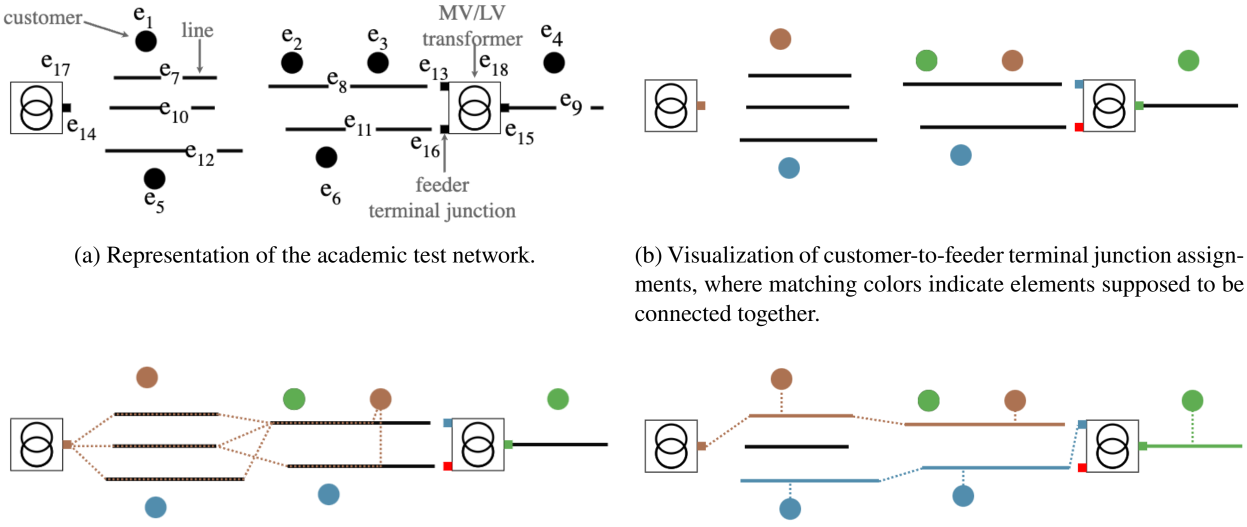
The set of elements known to the DSO is given by:

$$\mathcal{E} = \{e_1, ..., e_{18}\}. \tag{17}$$

The set of attributes, $\mathcal{A}$, is given by:

$$\mathcal{A} = \{type, coordinate, junction\}. \tag{18}$$

The set of types, $\mathcal{T}$, is given by: $customer$ (elements e_1 to e_6), $line$ (e_6 to e_{13}), $junction$ (e_{13} to e_{16}) and $transformer$ (e_{17} and e_{18}) as shown in Fig. 4a.



(a) Representation of the academic test network.

(b) Visualization of customer-to-feeder terminal junction assignments, where matching colors indicate elements supposed to be connected together.

(c) Image showing some hypothetical paths for the customer e_3.

(d) Final network configuration after customer path identification.

Figure 4: Illustration of the process of identifying customer paths in the academic network considered.

6.4. Hypothetical paths compatible with the well-defined information

The first two pieces of information are used to generate an initial set of hypothetical paths, $\mathcal{H}^{i'_1,i'_2}$. Therefore, the size of the set of hypothetical paths, compatible with the first two pieces of the well-defined information, i'_1 and i'_2, is calculated using Eq. 9:

$$|\mathcal{H}^{i'_1,i'_2}| = 61607. \tag{19}$$

The calculation is performed with the set of elements considered $\mathcal{E}$, the customer set $\mathtt{C} = Subset(\mathcal{E}, customer)$, the terminal points set $\mathtt{T} = Subset(\mathcal{E}, junction)$ and the remaining elements set $\mathtt{R} = \mathcal{E} - \mathtt{C} - \mathtt{T} - Subset(\mathcal{E}, transformer)$.

Each remaining piece of well-defined information is used to exclude the paths in $\mathcal{H}^{i'_1,i'_2}$ that are not compatible with the well-defined information. Therefore, considering all the remaining pieces of well-defined information, i'_3 to i'_5, the total number of hypothetical paths compatible with the well-defined information is $|\mathcal{H}^{\mathcal{I}'}|$.
Detailed information about the elements in each hypothetical path can be found in the online repository (see Section 1.2).

6.5. Results paths estimation

The matrices $\mathbf{H}_{\mathtt{C}}$, $\mathbf{H}_{\mathtt{R}}$, $\mathbf{H}_{\mathtt{T}}$, $\hat{\mathbf{P}}$ and $\mathbf{T}_{\mathtt{R}}$ are used in the optimization problem as shown in Eqs. 16a–e.
In particular, the matrices $\mathbf{H}_{\mathtt{C}}$, $\mathbf{H}_{\mathtt{R}}$ and $\mathbf{H}_{\mathtt{T}}$ are used as the parameters of the ILP problem, while $\hat{\mathbf{P}}$ and $\mathbf{T}_{\mathtt{R}}$ are the decision variables and represent a proposed solution to the optimization problem.
The matrix $\hat{\mathbf{P}}$ contains the estimated paths are considered as an estimation of the real paths; while the matrix $\mathbf{T}_{\mathtt{R}}$ represents which elements belong to each feeder terminal junction.

The estimated optimal paths for this academic network are:

$$\hat{\mathbf{P}} = \begin{array}{cccccccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ \end{array}$$
$$\hat{\mathbf{P}} = \left(\begin{array}{cccccccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array} \right) \tag{20}$$

The matrix $\mathbf{T}_{\mathtt{R}}$ is given by:

$$\mathbf{T}_{\mathtt{R}} = \begin{array}{c} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{array} \begin{array}{c} \begin{array}{cccccc} e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{21}$$

6.6. Results explanation

We will now examine the results obtained from applying the optimization method to the academic example. This analysis will clarify the meaning and role of each component in Eqs. 16a–e.

6.6.1. Maximization term

The maximization term, Eq. 16a, contains two operations:

- Maximizing estimated paths: the first term, $\sum_k^{|\mathcal{H}^{\mathcal{I}'}|} \hat{\mathbf{P}}_{(k)}$, focuses on maximizing the number of estimated optimal paths, therefore paths whose have a value of 1.
- Penalty for unnecessary assignments: the second term, $\sum_m^{|\mathtt{T}|} \sum_n^{|\mathtt{R}|} \mathbf{T}_{\mathtt{R}\,(m,\,n)}$, introduces a penalty term related to the terminal association matrix. This term discourages assigning elements to a feeder terminal junction when unnecessary.

The value of ω is important to determine the relative importance of finding more valid paths over minimizing the number of elements assigne to the path. A larger value of ω encourages the optimization to find more valid paths, while a smaller value puts more emphasis on minimizing the elements used.
In this case, ω is set to $\omega = 10$.

6.6.2. Validity path constraint

The left-hand side of the constraint term in Eq. 16b is calculated in three operations:

(i) $\sum_k^{|\mathtt{R}|} \mathbf{H}_{\mathtt{R}\,(\bullet,\,k)}$. This operation counts the number of elements present in each hypothetical path, $h \in \mathcal{H}^{\mathcal{I}'}$. The resulting matrix, of dimensions $|\mathcal{H}^{\mathcal{I}'}| \times 1$, is presented in transposed form for space efficiency as follows:

$$\begin{array}{c} \begin{array}{cccccccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \end{array} \\ \begin{pmatrix} 1 & 2 & 2 & 2 & 3 & 3 & 2 & 1 & 2 & 1 \end{pmatrix} \end{array} \tag{22}$$

(ii) The result, Eq. 22, is multiplied element-wise by $\hat{\mathbf{P}}$. This operation assures the verification of the constraint only for the estimated optimal paths. The resulting matrix, still of dimensions $|\mathcal{H}^{\mathcal{I}'}| \times 1$, is transposed and presented as follows:

$$\begin{array}{c} \begin{array}{cccccccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 2 & 1 \end{pmatrix} \end{array} \tag{23}$$

(iii) Finally, the result, Eq. 23 is multiplied by the matrix $\mathbf{H}^{\top}$. The result of this operation still has dimensions $|\mathcal{H}^{\mathcal{I}'}| \times |\mathtt{T}|$ and the transposed form is shown as follows:

$$\begin{array}{c} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{array} \begin{array}{c} \begin{array}{cccccccccc} h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{24}$$

The element-wise multiplication of matrices in Eq. 23 and $\mathbf{H}^\top$ is handled through a process known as broadcasting. This technique allows matrices of different dimensions to be multiplied element-wise without explicitly resizing them. Broadcasting follows these key principles: matrices do not need to have identical dimensions for element-wise operations, and the resulting matrix adopts the dimensions of the input matrix with the largest dimensions.
Formally, given a matrix $\mathbf{A}$ of dimensions $M \times 1$ and a matrix $\mathbf{B}$ of dimensions $M \times N$, with $M, N \in \mathbb{N}^+$ and the result of the product $\mathbf{A} \cdot \mathbf{B}$ is a matrix $\mathbf{C}$ of dimensions $M \times N$, where each element is given by:

$$\mathbf{C}_{(m,n)} = \mathbf{A}_{(m,1)} \cdot \mathbf{B}_{(m,n)}, \ \forall m \in M, \forall n \in N. \tag{25}$$

Therefore, the resulting matrix in Eq. 24 has dimensions $|\mathcal{H}^{\mathcal{I}'}| \times |\mathtt{T}|$.

Similarly, the right-hand side of the constraint term in Eq. 16b is calculated in two operations:

(i) the matrix multiplication of $\mathbf{T}_\mathtt{R}$ and $\mathbf{H}_\mathtt{R}^\top$. This operation calculates the number of elements that are present at the same time in a feeder terminal junction $t \in \mathtt{T}$ and in a path $h \in \mathcal{H}^{\mathcal{I}'}$. The resulting matrix has dimensions $|\mathtt{T}| \times |\mathcal{H}^{\mathcal{I}'}|$ and is given as follows:

$$\begin{array}{c|cccccccccc|} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ e_{13} & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 2 & 1 \\ e_{14} & 1 & 1 & 1 & 1 & 2 & 1 & 2 & 0 & 0 & 0 \\ e_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ e_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{26}$$

For example, considering the value at position (e_{14}, h_7) of the matrix in Eq. 26. The value is 2 since the same elements e_7 and e_8 are at the same time assigned to the feeder terminal junction e_{14} (second row of the junction association matrix in Eq. 21) and are present in the path h_7 (hypothetical path h_7 composed by the elements: $h_7 = (e_3, e_7, e_8, e_{14})$).

(ii) Similarly, for the left-hand side, the result of Eq. 26 is multiplied by the matrix $\mathbf{H}_\mathtt{T}$. This assures that the comparison on both sides of the inequality is performed only on the elements that are assigned to the same feeder terminal junction. The result of this operation has dimensions $|\mathtt{T}| \times |\mathcal{H}^{\mathcal{I}'}|$ and is expressed as follows:

$$\begin{array}{c|cccccccccc|} & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} \\ e_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 \\ e_{14} & 1 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ e_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ e_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \tag{27}$$

Finally, by comparing Eq. 24 and Eq. 27, the condition Eq. 24 ≤ Eq. 27 is verified for each correspondent value of the matrices.

In general, the constraint Eq. 16b assures that all the elements of an estimated path belong to the same feeder terminal junction. This is guaranteed when the number of elements in each estimated path is less than or equal to the total number of elements in the corresponding feeder terminal junction.

6.6.3. *Unique estimated path for customer constraint*

The constraint in Eq. 16c is given as follows:

$$\left(\hat{\mathbf{P}} \times \mathbf{H}_\mathtt{C}\right) = \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\ \end{array} \begin{pmatrix} 1 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \tag{28}$$

$$\left(\hat{\mathbf{P}} \times \mathbf{H}_{\mathtt{C}}\right)_{(k)} \leq 1, \quad \forall k \in \{1, ..., |\mathtt{C}|\} \tag{29}$$

Equation 28 is a matrix of dimensions $1 \times |\mathtt{C}|$ and in this academic example, all the values are 1 except for e_2 whose value is 0 since no path was identified for that customer.

6.6.4. *Elements assigned to a unique feeder terminal junction constraint*

The constraint in Eq. 16d is given as follows:

$$\sum_{k}^{|\mathtt{T}|} \mathbf{T}_{\mathtt{R}\,(k,\bullet)} = \begin{array}{c} \begin{matrix} e_7 & e_8 & e_9 & e_{10} & e_{11} & e_{12} \end{matrix} \\ \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{30}$$

$$\sum_{k}^{|\mathtt{T}|} \mathbf{T}_{\mathtt{R}\,(k,m)} \leq 1, \quad \forall m \in \{1, ..., |\mathtt{R}|\} \tag{31}$$

Equation 30 is a matrix of dimensions $1 \times |\mathtt{R}|$ and in this academic example, all the values are 1 except for e_{10} whose value is 0 since the element e_{10} is not used by any path, and therefore it is not assigned to any feeder terminal junction. A possible explanation for why e_{10} is not used is that it is a redundant element in the dataset, or that the optimization algorithm may have chosen that alternative paths provide a better solution.

6.7. Diagnostic function

The proposed solution, the matrices $\hat{\mathbf{P}}$ and $\mathbf{T}_{\mathtt{R}}$ are then evaluated using the $Diagnostic()$ function to detect any possible issues.

For example, the solution proposed does not find the paths for each customer, since it is not possible to find a path for customer e_2 given the data available to the DSO. This issue is reported to the DSO, who can perform some further analysis on the case and understand why no path was found.
A possible reason is an error in the data and in reality that the customer belongs to another feeder terminal junction or a missing line in the DSO data. In such a particular case, a possible solution could be to assign another feeder terminal junction to the customer, depending on the (k-)closest customer(s).

7. Belgian network

We apply our methodology to a real Belgian LV network characterized by incomplete GIS data and missing customer connections to the network. For this use case, we assume that the set of raw information, $\mathcal{I}$, the transformation functions, $\mathcal{F}$, and the well-defined information, $\mathcal{I}'$, are the same as those described in the academic example in Section 6.

For this real case, the set of elements known to the Belgian DSO is given by:

$$\mathcal{E} = \{e_1, ..., e_{1089}\}. \tag{32}$$

The set of attributes, $\mathcal{A}$, is given by:

$$\mathcal{A} = \{type, coordinate, junction\}. \tag{33}$$

The numbers of elements for the different types are given in Table 1 given below:

7.1. Hypothetical paths compatible with the well-defined information

The total number of hypothetical paths in the set $\mathcal{H}$ is generally determined using Eq. 9. Following the methodology $\mathfrak{M}()$, the hypothetical paths not compatible with the well-defined information are excluded, keeping only those compatible with each piece of information, obtaining the set $\mathcal{H}^{\mathcal{I}'}$.
For the real Belgian network considered, given the large number of elements in the network, it is impractical to explicitly represent the set $\mathcal{H}$ as a set of all the possible hypothetical paths.

Table 1
Distribution Network Element Counts

Subsets	Number of elements
$Subset(\mathcal{E}, customer)$	526
$Subset(\mathcal{E}, line)$	441
$Subset(\mathcal{E}, junction)$	96
$Subset(\mathcal{E}, transformer)$	26

For this reason, to efficiently identify the hypothetical paths that are compatible with the well-defined information, we develop a strategy to construct the set $\mathcal{H}^{\mathcal{I}'}$, or an approximation of it, without relying on the explicit enumeration of set $\mathcal{H}$. The strategy is to use an $A^*()$ pathfinding algorithm ([28]). The advantage of the $A^*()$ algorithm is to use a heuristic search approach to efficiently explore the possible hypothetical path space and to identify the hypothetical paths without exhaustively examining all possibilities. The implementation of the $A^*()$ function is available on the online repository (see Section 1.2). Below, we present the general concept.

The $A^*()$ function takes as input the set of network elements $\mathcal{E}$, the set of customers $\mathtt{C}$ for whom we want to identify hypothetical paths, the maximum number of hypothetical paths N and some specific conditions (for example the maximum connection distance D and the maximum length path L as specified by the well-defined information i'_3). For each customer, $c \in \mathtt{C}$, the $A^*()$ function explores its possible connections, using the function in Eq. 8, up to its feeder terminal junction, $c.junction$. The $A^*()$ function then outputs a set of hypothetical paths, $\mathcal{H}^*$, that is consistent with the well-defined information and that approximate the set $\mathcal{H}^{\mathcal{I}'}$.

After the execution of the $A^*()$ algorithm, considering a maximum number of paths for each customer of $N = 5$, the total number of hypothetical paths is $|\mathcal{H}^{\mathcal{I}'}| = 2348$. The value of N is selected as a balance between computational efficiency and the likelihood of identifying an optimal solution.

7.2. Matrix generation

The set $\mathcal{H}^{\mathcal{I}'}$ is decomposed into the three sub-matrices, $\mathbf{H}_{\mathtt{C}}$, $\mathbf{H}_{\mathtt{R}}$ and $\mathbf{H}_{\mathtt{T}}$, as explained in Section 5.5.1. These matrices are used as input parameters of the optimization problem.

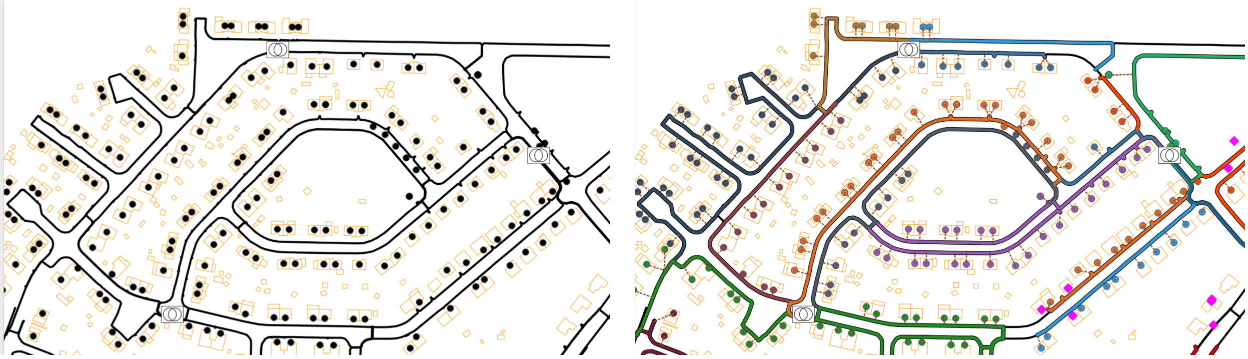
The matrices $\hat{\mathbf{P}}$ and $\mathbf{T}_{\mathtt{R}}$ are used as output decision variables of the optimization algorithm.

7.3. Optimization problem results

After executing the optimization algorithm, a solution is obtained in the form of matrices $\hat{\mathbf{P}}$ and $\mathbf{T}_{\mathtt{R}}$.
Displaying the matrices directly might be challenging for visualization and comprehension. Therefore, we summarize and present the results in a more comprehensible way.
Therefore, given the matrix, $\hat{\mathbf{P}}$ it is possible to represent graphically the connections among the elements of the estimated paths. In particular, each element of an estimated optimal path is assigned a specific color, with black reserved for unassigned elements.
The color that is assigned to each element depends on the matrix $\mathbf{T}_{\mathtt{R}}$. In particular, the same color is assigned to the elements that belong to the same feeder terminal junction.
Figure 5a shows a part of the Belgian network considered with three MV/LV transformers, some lines, and customers. In Fig. 5b it is possible to see the network after the identification of the paths: colors have been assigned to customers and lines belonging to the same feeder terminal junction. Dashed lines represent the connections between the elements, for example, customer-line connections.

7.4. Diagnostic function

Moreover, few purple squares are present. These are customers for which no path was identified and these are discovered with the $Diagnostic()$ function.
Possible reasons for failing to identify a path include incorrect or missing data for the feeder terminal junction associated with a customer, or the inability to assign a line to its corresponding feeder terminal junction.
These issues are reported to the DSO for further investigation.



(a) Before applying the methodology.

(b) After applying the methodology.

Figure 5: Part of the real network considered.

One of the main challenges in this TPI problem is the absence of a well-defined ground truth, as no explicit reference data is available. Consequently, quantitative criteria cannot be established for a rigorous evaluation of algorithmic performance. Instead, assessment must rely on qualitative evaluation through human interpretation. For example, in this case the DSO could access only the schematic images of customer paths and a human-in-the-loop process was employed to compare the identified paths against the schematics to assess their accuracy.

8. Conclusion

This paper introduces an optimization algorithm aimed at addressing the topological path identification (TPI) problem in power distribution networks. By leveraging information provided by distribution system operators (DSOs), the methodology constructs a set of hypothetical paths. This set is reduced by excluding the paths not compatible with the well-defined information. After that, an integer linear programming (ILP) algorithm identifies the best approximation of the real paths by selecting the paths that maximize customer connectivity to the correct medium-voltage/low-voltage (MV/LV) transformer. Our approach is based only on static data, without the requirements of data coming from advanced metering infrastructure (AMI) like smart meters. This makes the methodology more general and applicable to a wide range of distribution networks, including those in regions with limited technological infrastructure. Additionally, it minimizes the cost and complexity associated with collecting and integrating data from dynamic sources. The effectiveness of the proposed approach has been validated through demonstrations on both academic examples and real-world networks. This methodology is fundamental for constructing accurate digital twins of power distribution networks, ultimately aiding DSOs in network management and optimization.

Possible future work includes expanding the optimization algorithm to identify backup paths, using AMI sources such as smart meters and sensors to enhance the accuracy of the path identification, and improving the scalability and computational efficiency of the optimization algorithm, as proposed in [29].

References

[1] Frederik Geth, Marta Vanin, and Dirk Van Hertem. Data quality challenges in existing distribution network datasets, 06 2023.

[2] Maksymilian Kochański, Katarzyna Korczak, and Tadeusz Skoczkowski. Technology innovation system analysis of electricity smart metering in the european union. *Energies*, 13(4), 2020.

[3] Carlos Mateo Domingo, Tomas Gomez San Roman, Alvaro Sanchez-Miralles, Jesus Pascual Peco Gonzalez, and Antonio Candela Martinez. A reference network model for large-scale distribution planning with automatic street map generation. *IEEE Transactions on Power Systems*, 26(1):190–197, 2011.

[4] Andre Seack, Jan Kays, and Christian Rehtanz. Generating low voltage grids on the basis of public available map data. In *CIRED workshop 2014*.

[5] Alejandro Navarro-Espinosa and Luis F. Ochoa. Reconstruction of low voltage distribution networks: From GIS data to power flow models. 2015.

[6] Abdenago Guzmán, Jairo Quirós-Tortós, and Gustavo Valverde. Efficient connectivity identification of large-scale distribution network elements in GIS. In *2017 IEEE Manchester PowerTech*, page 6, 2017.

[7] Abdenago Guzmán, Andrés Argüello, Jairo Quirós-Tortós, and Gustavo Valverde. Processing and correction of secondary system models in geographic information systems. *IEEE Transactions on Industrial Informatics*, 15(6):3482–3491, 2019.

[8] Gao Zepu, Luo Yongjian, Xu Ziwei, Yu Yilan, and Zhang Lianmei. Knowledge graph-based method for identifying topological structure of low-voltage distribution network. *The Journal of Engineering*, 2020(12):1177–1184, 2020.

[9] X. Li, X. Feng, Z. Zeng, X. Xu, and Y. Zhang. Distribution feeder one-line diagrams automatic generation from geographic diagrams based on GIS. In *2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*, pages 2228–2232. IEEE, 2008.

[10] L. Wu, Y. Lin, and W. Pang. Distribution network topology modelling and automatic mapping based on CIM and GIS. In *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, page 5. IEEE, 2018.

[11] L. Shang, R. Hu, H. Ci, W. Zhang, and G. Ouyang. Automatic generation algorithm of distribution network topology map based on GIS drawing. In *IOP Conference Series: Earth and Environmental Science*, volume 384, 2019.

[12] Z. Yin, J. Luo, L. Wang, and R. Ye. Research on network topology analysis method of distribution management based on GIS. In *2021 International Conference on Intelligent Transportation, Big Data and Smart City (ICITBS)*, pages 145–148. IEEE, 2021.

[13] Mohammad Farajollahi, Alireza Shahsavari, and Hamed Mohsenian-Rad. Topology identification in distribution systems using line current sensors: An MILP approach. *IEEE Transactions on Smart Grid*, 11(2):1159–1170, March 2020.

[14] Daniele Marulli, Sébastien Mathieu, Amina Benzerga, Antonio Sutera, and Damien Ernst. Reconstruction of low-voltage networks with limited observability. In *2021 IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, page 5, 2021.

[15] Zahra Soltani and Mojdeh Khorsand. Real-time topology detection and state estimation in distribution systems using micro-PMU and smart meter data. 16(3):3554–3565.

[16] Changgang Wang, Jun An, and Gang Mu. Power system network topology identification based on knowledge graph and graph neural network. *Frontiers in Energy Research*, 2021.

[17] P. A. Parikh and T. D. Nielsen. Transforming traditional geographic information system to support smart distribution systems. In *2009 IEEE/PES Power Systems Conference and Exposition*, page 4, 2009.

[18] Jouni Peppanen, Santiago Grijalva, Matthew J. Reno, and Robert J. Broderick. Distribution system low-voltage circuit topology estimation using smart metering data. In *2016 IEEE/PES Transmission and Distribution Conference and Exposition*, page 5, 2016.

[19] Thomas Morrell, Venkatesh Venkataramanan, Anurag Srivastava, Anjan Bose, and Chen-Ching Liu. Modeling of electric distribution feeder using smart meter data. page 9, 04 2018.

[20] Shijie Cui, Peng Zeng, Chunhe Song, Zhongfeng Wang, and Guangye Li. Low-voltage distribution network topology identification based on constrained least square and graph theory. *Soft Computing*, 26(17):8509–8519, 2022.

[21] Haifeng Li, Wenzhao Liang, Yuansheng Liang, Zhikeng Li, and Gang Wang. Topology identification method for residential areas in low-voltage distribution networks based on unsupervised learning and graph theory. *Electric Power Systems Research*, 215, 2023.

[22] Xiao Yu, Jian Zhao, Haipeng Zhang, Xiaoyu Wang, and Xiaoyan Bian. Data-driven distributed grid topology identification using backtracking Jacobian matrix approach. *IEEE Transactions on Industrial Informatics*, 20(2):1711–1720, 2024.

[23] Abu Bakr Pengwah, Yasin Zabihinia Gerdroodbari, Reza Razzaghi, and Lachlan L. H. Andrew. Topology identification of distribution networks with partial smart meter coverage. *IEEE Transactions on Power Delivery*, 39(2):992–1001, 2024.

[24] Amina Benzerga, Daniele Maruli, Antonio Sutera, Alireza Bahmanyar, Sébastien Mathieu, and Damien Ernst. Low-voltage network topology and impedance identification using smart meter measurements. In *2021 IEEE Madrid PowerTech*, page 6. IEEE, 2021.

[25] Alexander Hoogsteyn, Marta Vanin, Arpan Koirala, and Dirk Van Hertem. Low voltage customer phase identification methods based on smart meter data. *ArXiv*, abs/2204.06372, 2022.

[26] Abu Bakr Pengwah, Yasin Zabihinia Gerdroodbari, Reza Razzaghi, and Lachlan L. H. Andrew. Topology identification of distribution networks with partial smart meter coverage. *IEEE Transactions on Power Delivery*, 39(2):992–1001, 2024.

[27] Maurizio Vassallo, Alireza Bahmanyar, Laurine Duchesne, Adrien Leerschool, Simon Gerard, Thomas Wehenkel, and Damien Ernst. A systematic procedure for topological path identification with raw data transformation in electrical distribution networks. In *Proceedings of International Conference on Energy, Electrical and Power Engineering*. CEEPE, 04 2024.

[28] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968.

[29] Thibaut Théate, Laurine Duchesne, Adrien Leerschool, Alireza Bahmanyar, Simon Gérard, Thomas Wehenkel, and Damien Ernst. Smart meters phase identification for topology verification: Practical challenges and insights from a case study. CIRED, Geneva, Switzerland, June 2025.