Creating Realistic Power Distribution Networks using Interdependent Road Infrastructure

ABSTRACT

It is well known that physical interdependencies exist between networked civil infrastructures such as transportation and power system networks. In order to analyze complex nonlinear correlations between such networks, datasets pertaining to such real infrastructures are required. However, such data are not readily available due to their sensitive nature. This work proposes a methodology to generate realistic synthetic power distribution networks for a given geographical region. A network generated in this manner is not the actual distribution system, but is very similar to the real distribution network. The synthetic network connects high voltage substations to individual residential consumers through primary and secondary distribution networks. Here, the distribution network is generated by solving an optimization problem which minimizes the overall length of the network subject to structural and power flow constraints. This work also incorporates identification of long high voltage feeders originating from substations and connecting remotely situated customers in rural geographic locations. The proposed methodology is applied to Montgomery county in Virginia and creates synthetic distribution networks which are validated with respect to structural feasibility and ability to operate within acceptable voltage limits under average load demand scenarios.

ACM Reference Format:

. 2020. Creating Realistic Power Distribution Networks using Interdependent Road Infrastructure. In *San Diego '20: ACM Conference on Knowledge, Discovery and Data Mining, August 22–27, 2020, San Diego, CA*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

In today's world, human behavior, social networks and civil infrastructures are closely intertwined. Recent works show how decisions undertaken by human behavior impact the load on infrastructure which in turn affects subsequent human decision-making [**? ?**]. Recent advances in artificial intelligence and computational science have created new opportunities to analyze complex nonlinear network dynamics in interdependent networks [**?**]. In order to study coupled infrastructures, a central challenge is the lack of realistic data sets. Such datasets should have detailed representations of each network as well as the interactions across infrastructures and the interactions with the human population. Various works have designed different *domain specific* methodologies to create, validate, and apply realistic datasets in social networks [**? ?**], protein-protein interaction networks [**? ?**], epidemic networks [**?**], networked infrastructures such as gas networks [**?**], power grids [**? ? ?**], and interdependent networks [**? ?**].

With the increased penetration of distributed generation in traditional power distribution systems, the complexity of the network is continually increasing [**?**]. In light of such changes, the primary focus of power system research is either towards developing algorithms for resilient distribution grid planning through optimal hosting capacity of distributed generation [**?**], optimal expansion of distribution networks [**?**], or methodologies for reliable operation of the network through restoration capabilities [**?**], and volt-var control [**?**]. To this end, the proposed algorithms require a suitable distribution network to study their performance. Such networks should structurally resemble a real distribution system and should pose similar challenges to researchers as a real network would and thereby ensure a meaningful evaluation of algorithms for practical implementation.

However, due to security concerns and due to the proprietary nature of the data, utilities refrain from sharing real distribution system information publicly. Therefore, it is necessary to create synthetic distribution test systems so that realistic networks are readily available to researchers [**? ? ?**]. Over the past few years, several researchers have tried to address the problem of generating realistic power networks so that they are openly available to researchers for studying complex phenomena [**? ? ?**]. However, these works are limited to the power transmission network where the distribution system is considered to be a passive element and impact of individual consumer decisions is not prominent.

Problem This works proposes a methodology to generate realistic synthetic distribution networks for a geographical location using open source information which connects substations to individual residential buildings. We aim to generate a network which resembles a real distribution network structurally and at the same time satisfies the operational constraints (voltage and power flows limits) of an actual operating distribution system. The problem can be formally stated as: *Given a set of residences and electric substations with their respective geographical coordinates, construct a power distribution network connecting these points that resembles a real operating distribution system network.*

1.1 Related Work

Several methodologies have been proposed to generate random distribution network topologies based on statistical distributions. The statistical distributions of graph attributes (such as node degree, hops from source, etc.) have been analyzed for a real distribution network and fitted to specific theoretical distributions [**?**]. Random networks were thereafter generated from these distributions which created synthetic medium voltage feeder networks. A stochastic geometry based approach has also been proposed to place transformers in

San Diego '20, August 22–27, 2020, San Diego, CA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

synthetic power networks [**?**]. However, these works do not consider the various power system operation constraints such as node voltages and edge flows. Furthermore, the generated networks are not optimal in terms of network loss and cost of installation which is always the primary consideration of a distribution company. Other approaches consider a population-dependent aggregated load at each zipcode center and create a synthetic transmission network connecting them [**? ? ?**]. Though these works generate random realistic power system networks, they are mostly limited to high voltage and medium voltage level and do not consider residences connected along the network.

Kadavil et al. [**?**] proposed a bottom-up approach to build a synthetic distribution network from a substation and thereafter to populate it with consumer loads. In other work, a mixed integer second order conic program (MI-SOCP) based optimization problem is formulated to generate a synthetic medium voltage network for a geographical region [**?**]. However, the location of loads are considered to be the zip code centers and loads are aggregated for each zip code which results in an aggregated synthetic distribution network of the region. The generated synthetic networks do not replicate the distribution of actual consumers in a geographical region. Thus, similar networks would be generated for rural and urban locations. In this paper, we have considered actual residential locations and therefore, the created synthetic networks accurately represent a network particular to the geographical location. [**?**] proposes methodology to create synthetic feeder networks for a given geographic region based on openly available buildings data, unelectrified areas data and land usage data. However, this work aggregates the load based on land cover data and limits the network to transformers serving residences.

One of the important aspects to generate optimal realistic synthetic distribution networks is to maintain the radial configuration [**?**] and ensure that the proposed optimization framework does not produce isolated loops in the network. A real distribution network may be constructed with multiple loops and redundant paths in order to maintain reliability, however it is operated with a radial structure in order to facilitate protection coordination [**?**]. This aspect of maintaining a radial configuration has been revisited by multiple works on distribution network reconfiguration (DNR) [**? ? ?**] and distribution system planning (DSP) [**? ? ?**]. However, most of these works pertain to selecting edges in medium voltage networks with fixed number of known root nodes (substations or distributed energy resources). Furthermore, the non-substation (or non-root) nodes are either considered to have positive load demand [**? ? ?**] or handled by additional *single commodity flow* constraints [**? ?**]. In the proposed work, an optimal radial distribution network topology is generated for a given geographical location (rural or urban). The network is considered to have multiple unknown root nodes which are connected to the substation feeder via long distance high voltage feeder lines and are responsible to serve remotely located consumers.

1.2 Contribution

Our main contributions are: (i) We develop a *first principles* based methodology to create realistic synthetic distribution network using information from other infrastructures such as transportation networks, residential data etc., (ii) Our approach results in an optimal network by minimizing the overall length of distribution lines which is a principal consideration of power companies while planning distribution networks, (iii) Our method generates a distribution network which is particular to the geographical location of interest and hence provides a realistic representation of the actual network, (iv) The nodes and edges of the generated network are labeled with all necessary attributes required for power flow analysis and therefore, can be used as suitable networks to test distribution system planning and operation algorithms.

Our approach combines real world data such as building information and road network data with general electric engineering practice to construct realistic distribution systems. The economic aspect of minimizing the total length of the network has been considered while creating the synthetic networks. This, in turn reduces the total amount of investment required for constructing the network. We further include electric engineering aspects such as maintaining a tree structure (for protection coordination), placing local transformers (pole-top and pad-mounted) along the road network and avoiding branching in the secondary network (to reduce voltage sag at the leaf nodes). These aspects have been encoded as constraints for the proposed optimization framework to create the synthetic networks. The behavioral aspects of population has been considered to determine synthetic load demand profile at each individual residences. The inclusion of all these aspects has allowed us to validate power flow characteristics for the created synthetic networks.

The generated synthetic networks connect individual residential buildings with the substations through primary and secondary networks. We include the geographical aspect of the region for which the synthetic network is created through the actual location of residences and layout of the road network. A rural area is often characterized by remotely located isolated loads which are not so frequent in urban regions. Our method considers multiple high voltage feeders from each substation to connect these remotely located loads as well as nearby loads and thereby reduces voltage drop at the remote load points. Instead of considering connecting aggregated loads at zipcode centers, our method generates a detailed distribution network for a given geographical area. The key differences of our method with other related previous methods to generate synthetic distribution networks has been listed in Table 1.

The consideration of economic, engineering, behavioral and geographical aspects for inferring the distribution network has allowed us to choose certain parameters in the proposed optimization framework. These parameters can be modified depending on user's requirement or can be randomly selected to create ensembles of networks. Due to limitation of space, we have included ensembles of network by varying one parameter in this paper.

2 PRELIMINARIES

2.1 Distribution system

The distribution network consisting of overhead power lines, underground cables, pole top transformers is responsible to bring electrical power from high voltage (HV)(greater than 33kV) transmission system to the end residential consumers requiring a low voltage (LV) level (208-480V). This is normally done in a two-step procedure: (i) the high transmission level voltage is stepped down to medium voltage (MV) level at distribution substations and distributed to local transformers (pole-top/pad-mounted) through *primary distribution*

Table 1: Major contributions of proposed work

Aspect	Previous Works	Present work
Network type	Synthetic transmission networks with generators and aggregated loads [**? ? ? ?**]	Synthetic distribution networks are created which connect high voltage substations to individual residential consumers.
Realism of generated networks	Statistical distribution of network attributes to generate synthetic power grids [**?**]	Realistic distribution networks comprising of primary and secondary levels are generated for a given geographical location. The generated network resembles an optimal network designed by power distribution companies. It follows the usual structural and power flow constraints of a typical distribution system.
	Stochastic geometry based approach to place transformers in distribution networks [**?**]	
	Heuristic approach to synthesize distribution networks from substations and populate it with consumer loads [**?**]	
Radiality of network	Radiality is ensured by avoided isolated cycles or considering single commodity flow model [**? ? ? ?**].	Power balance constraint is proved to be sufficient condition to ensure radiality of generated network.
Network attributes	Small sized networks such as standard IEEE test systems are used [**? ?**].	Optimal radial network is identified for unknown number of root nodes (high voltage feeders) and for large sized networks with more than 20000 nodes.
	The number of root nodes (substations) are known beforehand in the problem definition.	

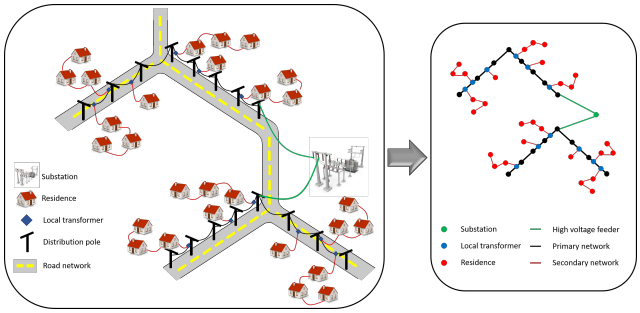

Figure 1: A schematic of synthetic distribution network. The substation feeds the distribution network through high voltage feeder lines. The primary network originates from the feeder lines and follows the road network as much as possible. The secondary network originates from transformers along primary network and connects individual residences. The single line diagram for the same network with its individual elements is shown on the right figure.

network, (ii) the voltage is further stepped down to LV at the local distribution transformers and distributed to individual customers through *secondary distribution network*.

Distribution systems (primary and secondary) are usually configured in a *radial* structure where power flow is unidirectional from source to consumers. Such radial structure ensures protection coordination among reclosures, breakers and downstream fuses in the distribution system feeders [**?**]. Fig. 1 shows a schematic of the structure of synthetic distribution network. The network connects substation to individual residential buildings through primary and secondary networks. The substation first feeds the primary network through high voltage feeder lines. It is assumed that the primary network follows the available road network to the maximum extent since distribution poles are generally placed along the road network. The secondary network originates from the pole top transformers which are placed along primary network. The secondary network connects individual residences through short chains. Branching in these chains is avoided in order to reduce voltage drop for the leaf nodes.

2.2 Available datasets

In this work, we try to generate the distribution network for a given geographical region (county/town/city) from different open source publicly available information. These data pertain to following sources: (i) transportation network data published by [**?**], (ii) Geographical location of HV substations from data sets published by [**?**] and (iii) Residential electric power demand information developed in the models by [**?**].

Roads The road network represented in the form of a graph $\mathcal{R} = (\mathcal{V}_\mathcal{R}, \mathcal{L}_\mathcal{R})$, where $\mathcal{V}_\mathcal{R}$ and $\mathcal{L}_\mathcal{R}$ are respectively the sets of nodes and links of the network. Each road link $l \in \mathcal{L}_\mathcal{R}$ is represented as an unordered pair of terminal nodes (u, v) with $u, v \in \mathcal{V}_\mathcal{R}$. Each road node has a spatial embedding in form of longitude and latitude. Therefore each node $v \in \mathcal{V}_\mathcal{R}$ can be represented in two dimensional space as $\mathbf{p_v} \in \mathbb{R}^2$. Similarly, a road link $l = (u, v)$ can be represented as a vector $\mathbf{p_u} - \mathbf{p_v}$.

Substations The set of substations $\mathsf{S} = \{s_1, s_2, \cdots, s_M\}$, where the area consists of M substations and their respective geographical location data. Each substation can be represented by a point in the 2-D space as $\mathbf{p_s} \in \mathbb{R}^2$.

Residences The set of residential buildings with geographical coordinates $\mathsf{H} = \{h_1, h_2, \cdots, h_N\}$, where the area consists of N home locations. Each residential building can be represented by a point in the 2-D space as $\mathbf{p_h} \in \mathbb{R}^2$.

Table 2: Datasets and related attributes used to generate synthetic distribution network

Dataset	Source	Attributes
Substation	Electric substation data published by US Department of Homeland Security [**?**]	• substation ID • longitude • latitude
Road network	GIS and electronic navigable maps published by NAVTEQ [**?**]	• node ID • node longitude • node latitude • link ID • link importance level
Residences	Synthetic population and electric load demand profiles generated by [**?**]	• residence ID • longitude • latitude • average hourly load demand

3 PROPOSED APPROACH

The goal of this work is to generate a realistic synthetic distribution network to connect the substations to all residential building locations. As discussed in Section 2, a typical distribution system consists of primary and secondary networks. The problem of creating a synthetic network which connects all residences to substations is computationally expensive due to the large number of variables. Therefore, the synthesis of such networks is considered to be a two-step bottom-up procedure: the first step constructs the secondary distribution network connecting the residential buildings to pole-top/pad-mounted transformers and the second step involves connecting these local transformers to distribution substations through feeders and laterals. In our work, the road network is used as a proxy for the primary distribution network. Therefore, the locations of local pole-top transformers are considered to be internal points on the road network links.

To summarize the overall procedure, the tasks would be the following: (i) Evaluate a mapping between sets of residential buildings and road network links such that each residence is mapped to the nearest road link. Additionally, identify probable locations of local distribution transformers along these road network links. (ii) Connect the local distribution transformers to mapped residences in a radial configuration to resemble a typical secondary distribution network. (iii) Identify subset of road network links which connects distribution substation transformers to the local pole-top/pad-mounted transformers in the form of a radial feeder network.

3.1 Mapping residences to road links

This section details the proposed methodology to identify subsets of residential buildings near each road network link. This information would be used in the successive steps to generate the secondary distribution network. The points along road network would serve as local distribution transformers delivering power to the residential buildings mapped to the link.

Algorithm 1 is used to compute the nearest road network link to a given point with associated spatial embedding. First, a bounding region of suitable size is evaluated for each road network link. This is done such that any point in the region is within a radius r from any internal point of the road network link l. The bounding region for link $l = (u, v) \in \mathcal{L}_\mathcal{R}$ is

$$\mathbf{B_l} = \left\{\mathbf{p} \middle| ||\mathbf{p} - \mathbf{p_l}||_2 \leq r, \forall \mathbf{p_l} = \theta \mathbf{p_u} + (1-\theta)\mathbf{p_v}, \theta \in [0,1]\right\} \quad (1)$$

Similarly, a bounding region is considered for a residential building $h \in \mathsf{H}$

$$\mathbf{B_h} = \left\{\mathbf{p} \middle| ||\mathbf{p} - \mathbf{p_h}||_2 \leq r\right\} \quad (2)$$

The intersections between the bounding region of the building and those for the links are stored and indexed in a *quad-tree* data structure [**?**]. This information is retrieved to identify the k selected links which are comparably nearer to the residential building than the others. This algorithm reduces the computational burden of evaluating the distance between all road links and residential buildings. Now, the geodesic distance of point $\mathbf{p_h}$ can be computed from the k links and the nearest link can be identified. The internal points along the road links which are mapped to at least one residence are assumed to be probable locations of local distribution transformers. Therefore, an additional objective is to identify points along the road network link where local distribution transformers may be placed. In this paper, we interpolate multiple points along the link which are a definite length apart from each other. This is depicted in Algorithm 2.

Algorithm 1 Find the nearest link in $\mathcal{L}_\mathcal{R}$ to a given point $\mathbf{p}$.

Require: Radius for bounding boxes r.

1: **for** each link $\mathbf{l} \in \mathcal{L}_\mathcal{R}$ **do**
2: evaluate bounding box $\mathbf{B_l}$ for each link $\mathbf{l}$ using Eq. 1.
3: Evaluate bounding box $\mathbf{B_p}$ for point $\mathbf{p}$ using Eq. 2.
4: Find the bounding boxes $\mathbf{B_{l_1}}, \mathbf{B_{l_2}}, \cdots, \mathbf{B_{l_k}}$ corresponding to the links $\mathbf{l_1}, \mathbf{l_2}, \cdots, \mathbf{l_k}$ which intersect with $\mathbf{B_p}$.
5: Find the link $\mathbf{l}^\star$ among k short-listed links, which is nearest to point $\mathbf{p}$.

Algorithm 2 Creating node sets for secondary network generation.

Require: Road network $\mathcal{R} = (\mathcal{V}_\mathcal{R}, \mathcal{L}_\mathcal{R})$, set of residential buildings H, minimum distance between local transformers d.

1: **for** each building $h \in \mathsf{H}$ **do**
2: find mapping $f : \mathsf{H} \rightarrow \mathcal{L}_\mathcal{R}$ using Algorithm 1 to generate the nearest link $e \in \mathcal{L}_\mathcal{R}$.
3: **for** each link $l = (u, v) \in \mathcal{L}_\mathcal{R}$ **do**
4: find the inverse mapping $f^{-1} : \mathcal{L}_\mathcal{R} \rightarrow \mathsf{H}$ which generates the set of buildings H_l associated with l.
5: Interpolate m points $\mathsf{T}_l = \{t_1, t_2, \cdots t_m\}$ on link l between road nodes u and v which are d distance apart from each other.

3.2 Creation of secondary network

This section details the generation of secondary distribution networks connecting set of residential buildings H_l mapped to a road network link $l \in \mathcal{L}_\mathcal{R}$ with set of local distribution transformer nodes

T_l located along l. Without loss of generality, we only consider non-trivial road links which are mapped to at least one residential building, i.e, $f^{-1}(l) = \mathsf{H}_l \neq \emptyset$. The generated network should be a *forest* of disconnected trees with each tree rooted at one of the transformer nodes $t \in \mathsf{T}_\mathsf{l}$. The forest should cover all residential buildings in such a way that the overall length of distribution lines is minimized and intersections/overlaps of road link (proxy for primary) and secondary networks are avoided.

Network initialization In the proposed methodology, for a given non-trivial road network link $l \in \mathcal{L}_\mathcal{R}$, a fully connected undirected graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ is constructed with node set $\mathcal{V}$ and edge set $\mathcal{E}$ that are incident on $\mathcal{V}$. The node set comprises of n_h residences mapped to the link l and n_t transformers along l (i.e.,$\mathcal{V} = \mathsf{H}_l \cup \mathsf{T}_l$). Any edge $e \in \mathcal{E}$ is defined by incident nodes (i, j) where $i, j \in \mathcal{V}$. Since $\mathcal{G}$ is fully connected, the edge set $\mathcal{E}$ consists of all pairs of entries (i, j) for $i, j \in \mathcal{V}$. The aim of this step is to select an optimal subgraph $\tilde{\mathcal{G}} := (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ which is a forest of disconnected trees.

Edge variables In order to identify which edges are required to be connected in the optimal topology, we introduce binary variables $\{x_e\}_{e\in\mathcal{E}} \in \{0, 1\}^{|\mathcal{E}|}$. Variable $x_e = 1$ indicates that the edge is present in the optimal topology and vice versa. Each edge $e = (i, j)$ is assigned a flow variable f_e directed from the source node i to destination node j. The binary variable and flows can be stacked in $|\mathcal{E}|$-length vectors $\mathbf{x}$ and $\mathbf{f}$ respectively.

Node variables Each node $i \in \mathcal{V}$ either consumes power (if it is a residential building) or it injects power into the network (if it is a pole-top/pad-mounted transformer). Since each residence is associated with an hourly load demand profile, the average hourly load can be computed and be denoted by $p_i > 0$ which denotes its power consumption. For transformer nodes, power flows into the network which requres $p_i < 0$. The nodal power consumption at all nodes can be stacked in $n_t + n_h$ length vector $\mathbf{p}$. Note that $\mathbf{p} = \begin{bmatrix}\mathbf{p_T}^T & \mathbf{p_H}^T\end{bmatrix}^T$ with $\mathbf{p_H}$ denoting the power consumption at the home nodes stacked in a n_h length vector.

Degree constraint Statistical surveys on distribution networks in [**?**] show that residences along the secondary network are mostly connected in series with at most 2 neighbors. This is ensured by (3) which limits degree of residence nodes to 2.

$$\sum_{e:(h,j)} x_e \leq 2, \quad \forall h \in \mathsf{H}_l \tag{3}$$

Power flow constraints Given the fully connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we define the $(n_h + n_t) \times |\mathcal{E}|$ branch-bus incidence matrix $\mathbf{A}_\mathcal{G}$ with the entries as

$$\mathsf{A}_{\mathcal{G}\,k,e} := \begin{cases} +1, & k = i \\ -1, & k = j \\ 0, & \text{otherwise} \end{cases} \qquad \forall e = (i, j) \in \mathcal{E} \tag{4}$$

Since the order of rows and columns in $\mathbf{A}_\mathcal{G}$ is arbitrary, we can partition the rows without loss of generality as $\mathbf{A}_\mathcal{G} = \begin{bmatrix}\mathbf{A_T}^T & \mathbf{A_H}^T\end{bmatrix}^T$ where the partitions are the rows corresponding to transformer and residence nodes respectively.

The *linearized distribution flow* (LDF) model has been extensively used for several grid optimization tasks in order to relate power consumption and flows to voltages. By ignoring network losses, the LDF model gives the relation between power consumption at the nodes and power flow along edges as (5a). Note that the optimal network is obtained from $\mathcal{G}$ after removing the edges for which $x_e = 0$. Therefore, (5a) is accompanied by (5b) which forces flows f_e to be zero for non-existing edges. The value of $\overline{f}$ in (5b) can be chosen to be the power flow limits in the edges.

$$\mathbf{A_H f} = \mathbf{p_H} \tag{5a}$$

$$-\overline{f} x_e \leq f_e \leq \overline{f} x_e, \quad \forall e \in \mathcal{E} \tag{5b}$$

Ensuring radial topology In our case, this condition is satisfied by the node power flow condition in (5a) if all the residential nodes consume power which is a reasonable assumption. This is an extension of the proposition in [**?**] which considers renewable generation at the nodes. However, contrary to the previous work, the following proposition is a special case where every non-root node has only positive load demand. Using the power balance constraints to ensure radial topology reduces the number of constraints when dealing with large sized networks. The proof of the following proposition is presented in the Appendix.

Proposition 1. *A graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with reduced branch-bus incidence matrix $\mathbf{A_H}$ and residential node power consumption vector $\mathbf{p_H}$, with strictly positive entries, is connected if and only if there exists a vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{E}|}$, such that (5a) is satisfied.*

Once the connectivity of the subgraph $\tilde{\mathcal{G}}$ is ensured the radiality requirement can be enforced by the following constraint

$$\sum_{e\in\mathcal{E}} x_e = n_h \tag{6}$$

Generating optimal network topology The edges $(u, v) \in \mathcal{E}$ for all $u, v \in \mathcal{V}$ are assigned weights $w(u, v)$

$$w(u, v) = \begin{cases} \infty, & \text{if } u, v \in \mathsf{T_e} \\ \text{dist}(u, v) + \lambda \mathsf{C}(u, v), & \text{otherwise} \end{cases} \tag{7}$$

where $\text{dist} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ denotes the geodesic distance between the residential locations u, v. The function C denotes if the edge crosses the nearest road link and is defined as

$$\mathsf{C}(u, v) = \begin{cases} 0, & \text{if } u, v \text{ are on same side of link} \\ 2, & \text{if } u, v \text{ are on opposite side of link} \\ 1, & \text{if } u \in \mathsf{V_R} \text{ or } v \in \mathsf{V_R} \end{cases}$$

λ is a weight factor to penalize multiple crossing of edges over the road links. It also penalizes multiple edges emerging from the root node. The weights are stacked in $|\mathcal{E}|$-length vector $\mathbf{w}$. Thereafter, a forest with trees rooted at the probable transformer nodes and spanning all the nodes of G is considered as the secondary network. This is obtained after solving the optimization problem.

$$\begin{aligned} &\min_{\mathbf{x}} \ \mathbf{w}^T \mathbf{x} \\ &\text{s.to. } (3), (5), (6) \end{aligned} \tag{8}$$

3.3 Creation of primary network

The previous optimization framework is used to create secondary distribution network for all links $l \in \mathcal{L}_\mathcal{R}$ for which $f^{-1}(l) \neq \emptyset$ holds true. Now, our goal is to connect the local transformer nodes (roots of tree components in secondary distribution network) to the substations with the road network as a proxy. The primary distribution

network should also maintain a radial configuration. Additionally, the substation might be located far away from remote local transformers (particularly in rural feeders). In such a case, a high voltage feeder line (higher than the primary distribution level) is used to deliver power over a long distance to the remotely located customers. Thereafter, a step down transformer with a voltage regulator is used to bring the voltage down to the primary voltage level.

Network initialization The first step to construct the primary network is to identify the set of possible edges which might be a part of the primary network. Note that the transformer nodes are internal points along the road network links. We follow Algorithm 3 to obtain the possible set of edges.

Algorithm 3 Initialization of primary distribution network.

Require: Road network $\mathcal{R} = (\mathcal{V}_\mathcal{R}, \mathcal{L}_\mathcal{R})$, set of road links $\mathcal{L}_\mathcal{T} = \{l_k | l_k \in \mathcal{L}_\mathcal{R}, f^{-1}(l_k) \neq \emptyset\}$.

1: Initialize network $\mathcal{G}_\mathcal{P}(\mathcal{V}_\mathcal{P}, \mathcal{E}_\mathcal{P}) \leftarrow \mathcal{R}$
2: **for** each road link $l_k = (u_k, v_k) \in \mathcal{L}_\mathcal{T}$ **do**
3: Solve optimization problem in Section 3.2.
4: Identify set of transformers $\mathsf{T_k} = \{t_1, t_2, \cdots, t_{m_k}\}$
5: Construct path $\mathcal{P}_k = (u_k, t_1, t_2, \cdots, t_{m_k}, v_k)$
6: Remove link $l_k = (u_k, v_k)$ from network, $\mathcal{E}_\mathcal{P} \leftarrow \mathcal{E}_\mathcal{P} \setminus \{l_k\}$
7: **for** each edge $p_i \in \mathcal{P}_k$ **do**
8: Add edge p_i to network, $\mathcal{E}_\mathcal{P} \leftarrow \mathcal{E}_\mathcal{P} \cup \{p_i\}$
9: Update nodes with transformer nodes $\mathcal{V} \leftarrow \mathcal{V} \cup \mathsf{T_k}$
10: Return the final network $\mathcal{G}_\mathcal{P}$

Clustering of nodes for each substation An optimization problem can be solved to identify the optimal set of edges which minimizes the overall length of the network and at the same time satisfies all structural constraints. However, the scale of such a problem can be as high as dealing with 15000 nodes when an entire county is considered.

The problem of creating primary network for a region with multiple substations can be solved for individual substation. To this end, the road network nodes and transformers are clustered so that each node (road network or transformer) is mapped to the nearest substation. The network $\mathcal{G}_\mathcal{P}(\mathcal{V}_\mathcal{P}, \mathcal{E}_\mathcal{P})$ is partitioned into M subgraphs $\{\mathcal{G}_{s_1}, \mathcal{G}_{s_2}, \cdots, \mathcal{G}_{s_M}\}$ corresponding to each of the substation. This is depicted in Fig 2 where each color represents a partition of road and transformer nodes. These partitions are known as Voronoi cells which are centered at the substation location. The partitioning is done based on the shortest path distance metric which ensures that each node is mapped to the nearest substation and each induced subgraph $\mathcal{G}_s(\mathcal{V}_s, \mathcal{E}_s)$ corresponding to substation s has a single connected component.

Our goal is to identify optimal primary network $\mathcal{P}_s(\tilde{\mathcal{V}}_s, \tilde{\mathcal{E}}_s)$ from the set of edges in $\mathcal{G}_s(\mathcal{V}_s, \mathcal{E}_s)$. Note that $\mathcal{V}_s$ consists of local transformer nodes as well as road network nodes ($\mathcal{V}_s = \mathcal{V}_{s_T} \cup \mathcal{V}_{s_R}$). The optimal primary network is a forest of trees covering all nodes in the set of transformer nodes $\mathcal{V}_{s_T}$. However, not all road network nodes need to be covered as they serve as *transfer* nodes which has no power consumption and are present to transfer power from preceding to succeeding node. Furthermore, the road network nodes can also act as root node connecting HV feeder lines from the substation with the primary network. That is, the nodes in set $\mathcal{V}_{s_R}$ may either

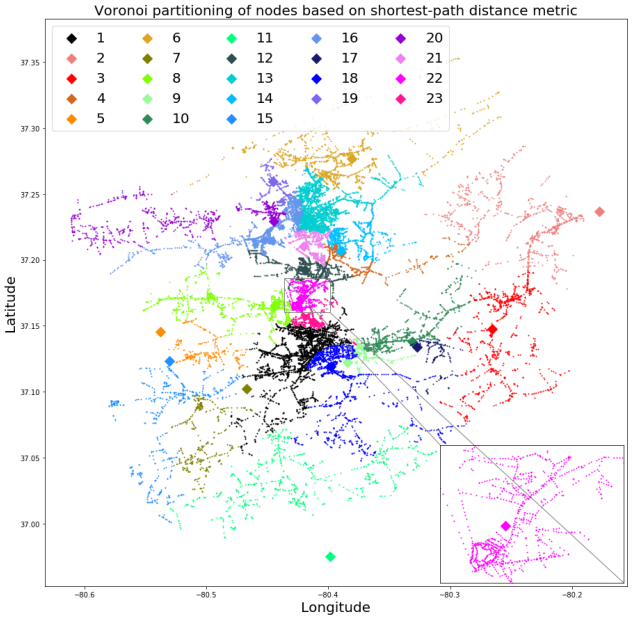

Figure 2: The nodes of initialized primary network are clustered based on the shortest-path distance to different substations. This enables us to create individual primary distribution network for each substation and hence reduces overall computation time of total network generation. The larger sized points are the substations and the small points are the nodes to be connected in the primary network. Each color represents a separate cluster of nodes corresponding to a substation and these nodes are required to be connected through the primary network.

be chosen to be included in $\mathcal{P}_s$ or not. If a road node is chosen it can either be a root node for a tree or can be a *transfer* node with a degree of at least 2.

Node variables $\mathcal{V}_s$ comprises of n_t transformer and n_r road nodes. Let p_i denote the power consumption at the i^{th} transformer node. This is obtained by summing up the total load demand of residences connected to the transformer. These power consumptions can be stacked in a n_t-length vector $\mathbf{p}$. Let v_i represent the voltage at the node i. The nodal voltages at all nodes can be stacked in $n_t + n_r$ length vectors $\mathbf{v}$.

We assign binary variables $\{y_r, z_r\}_{r \in \mathcal{V}_{s_R}} \in \{0, 1\}$. Variable $y_r = 1$ indicates that road network node r is part of the primary network and vice versa. Variable $z_r = 0$ indicates that road network node r is included and is a root node (connected to substation through a high voltage feeder line) in the primary network. $z_r = 1$ indicates that the road node is included but is not a root node. Furthermore, we need to enforce that if a road node is not chosen ($y_r = 0$), it is to be treated as a non-root node ($z_r = 1$).

Edge variables In order to identify which edges comprise of the primary network, each edge e is assigned a binary variable $\{x_e\}_{e \in \mathcal{E}_s} \in \{0, 1\}^{|\mathcal{E}_s|}$. Variable $x_e = 1$ indicates that the edge is part of the optimal primary network and vice versa. The binary variable for all

edges and edge power flows can be stacked in a $|\mathcal{E}_s|$-length vectors $\mathbf{x}$ and $\mathbf{f}$ respectively.

Connectivity constraint The first set of constraints are defined to ensures the structural feasibility of road nodes in the optimal primary network.

$$\sum_{e:(r,j)} x_e \leq |\mathcal{E}_s| y_r, \quad \forall r \in \mathcal{V}_{s_R} \tag{9a}$$

$$\sum_{e:(r,j)} x_e \geq y_r, \quad \forall r \in \mathcal{V}_{s_R} \tag{9b}$$

$$\sum_{e:(r,j)} x_e \geq 2(y_r + z_r - 1), \quad \forall r \in \mathcal{V}_{s_R} \tag{9c}$$

$$1 - z_r \leq y_r, \quad \forall r \in \mathcal{V}_{s_R} \tag{9d}$$

Let $e:(r,j)$ denote the set of all edges $e \in \mathcal{E}_s$ which are incident on the road node $r \in \mathcal{V}_{s_R}$. $\sum_{e:(r,j)} x_e$ is the degree of node r in graph $\mathcal{G}_s$. There are three possibilities for each road node r: (i) it is not included in the primary network, (ii) it is included and is not a root node and (iii) it is included and is a root node (connected to substation through a high voltage feeder line).

If r is not included ($y_r = 0$), (9d) ensures $z_r = 1$. Further, (9a), (9b)and (9c) ensures that there are no incident edges on r since we have $\sum_{e:(r,j)} x_e = 0$.

If r is included in primary network and is a root node ($y_r = 1, z_r = 0$), (9a), (9b)and (9c) ensures that the degree of the road node is positive.

Finally, if r is included and is not a root node ($y_r = 1, z_r = 1$), it has to be a transfer node with a minimum degree of 2 which is ensured by (9c) through $\sum_{e:(r,j)} x_e \geq 2$. The binary variables y_r, z_r can be stacked in n_r-length vectors $\mathbf{y}$ and $\mathbf{z}$ respectively.

Ensuring radiality constraint Since all the road network nodes need not be covered by the generated radial topology, we have the following equivalent of Eq. 6.

$$\sum_{e \in \mathcal{E}_s} x_e = |\mathcal{V}_s| - \sum_{r \in \mathcal{V}_{s_R}} (1 - y_r) - \sum_{r \in \mathcal{V}_{s_R}} (1 - z_r) \tag{10}$$

The total number of edges in the optimal primary network is equal to the difference between total number of nodes to be covered and number of root nodes. The last term in (10) denotes the number of root nodes and the first two terms together indicates the total number of nodes to be covered.

Power balance and flow constraints We can define the branch-bus incidence matrix $\mathbf{A}_{\mathcal{G}_s} \in \mathbb{R}^{|\mathcal{V}_s| \times |\mathcal{E}_s|}$ similar to Eq. 4. Since the order of rows and columns in $\mathbf{A}_{\mathcal{G}_s}$ is arbitrary, we can partition the rows without loss of generality as $\mathbf{A}_{\mathcal{G}_s} = \begin{bmatrix} \mathbf{A_{s_T}}^T & \mathbf{A_{s_R}}^T \end{bmatrix}^T$. Here, the partitions $\mathbf{A_{s_T}}$ and $\mathbf{A_{s_R}}$ are obtained by stacking the rows of $\mathbf{A}_{\mathcal{G}_s}$ corresponding to transformer nodes and road nodes respectively.

$$\mathbf{A_{s_T}} \mathbf{f} = \mathbf{p} \tag{11a}$$

$$-\overline{s}(1 - \mathbf{z}) \leq \mathbf{A_{s_R}} \mathbf{f} \leq \overline{s}(1 - \mathbf{z}) \tag{11b}$$

$$-\overline{f}\mathbf{x} \leq \mathbf{f} \leq \overline{f}\mathbf{x} \tag{11c}$$

(11a) ensures that a path exists between each transformer node and a root node following Proposition 1. If a road node is not a root node (with $z_r = 1$), (11b) enforces that the consumption at the node is 0. If a road node is a root node (with $z_r = 0$), the power consumption/injection at the node is limited by the substation power

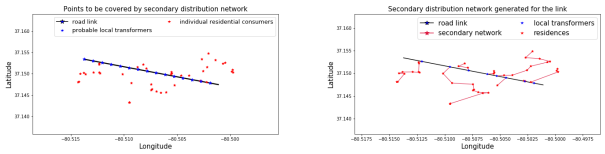

Figure 3: Secondary distribution network generated for a residences along a road link in Montgomery county of south-west Virginia, USA. The left figure shows the probable local transformers and residences to be connected. The right figure shows the generated optimal secondary distribution network.

capacity limits $[-\overline{s}, \overline{s}]$. (11c) ensures that if an edge $e \in \mathcal{E}_s$ is selected ($x_e = 1$), the power flow is limited by the line capacity limits $[-\overline{f}, \overline{f}]$.

Voltage constraints The long HV lines from the substation to the root nodes in the optimal primary distribution network end in voltage regulators which ensure that the root nodes have a voltage of 1pu. This is ensured by (12a). (12b) limits the voltage at all nodes within the acceptable limits of $[\underline{v}, \overline{v}]$.

$$(1 - v_r) \leq z_r \; \forall r \in \mathcal{V}_{s_R} \tag{12a}$$

$$\underline{v}\mathbf{1} \leq \mathbf{v} \leq \overline{v}\mathbf{1} \tag{12b}$$

$$-(1 - x_e)M \leq v_i - v_j - r_e f_e \leq (1 - x_e)M, \; \forall e \in \mathcal{E}_s \tag{12c}$$

If r_e denotes the resistance of the line $e:(i,j) \in \mathcal{E}_s$, the LDF model relates the squared voltage magnitude to power flows linearly as $v_i^2 - v_j^2 = 2r_e f_e$ where f_e is the entry from the vector $\mathbf{f}$ corresponding to edge e. The squared voltage can be approximated as $v_i^2 \approx 2v_i - 1$ which leads to the relation $v_i - v_j = r_e P_e$ (see [**?**]). Notice that this constraint is only activated for those edges $e \in \mathcal{E}_s$ for which $x_e = 1$. Therefore, we can enforce this constraint as (12c). If an edge is not selected ($x_e = 0$), (11c) ensures that $f_e = 0$. Therefore, M in (12c) can be selected to be $M = \overline{v} - \underline{v}$.

Generating optimal primary network Each edge $e = (u, v) \in \mathcal{E}_s$ is assigned a weight $w_e = w(u, v) = \text{dist}(u, v)$ which is the geodesic distance between the nodes. Additionally, for every road node $r \in \mathcal{V}_{s_R}$, we compute its geodesic distance from the substation s and is denoted by d_r. The optimal primary network topology is obtained by solving the optimization problem.

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \sum_{e \in \mathcal{E}_s} x_e w_e + \sum_{e \in \mathcal{V}_{s_R}} (1 - z_r) d_r$$
$$\text{s.to. } (9), (10), (11), (12) \tag{13}$$

4 RESULTS AND DISCUSSION

The proposed methodology is used to generate synthetic distribution networks connecting substations to individual residences in Montgomery county of south-west Virginia, USA. The details regarding the datasets used is shown in Table 3.

Table 3: Overview of datasets for Montgomery county

Dataset	Substations	Road network		Residences
		Nodes	Edges	
Size	23	8461	10114	37223

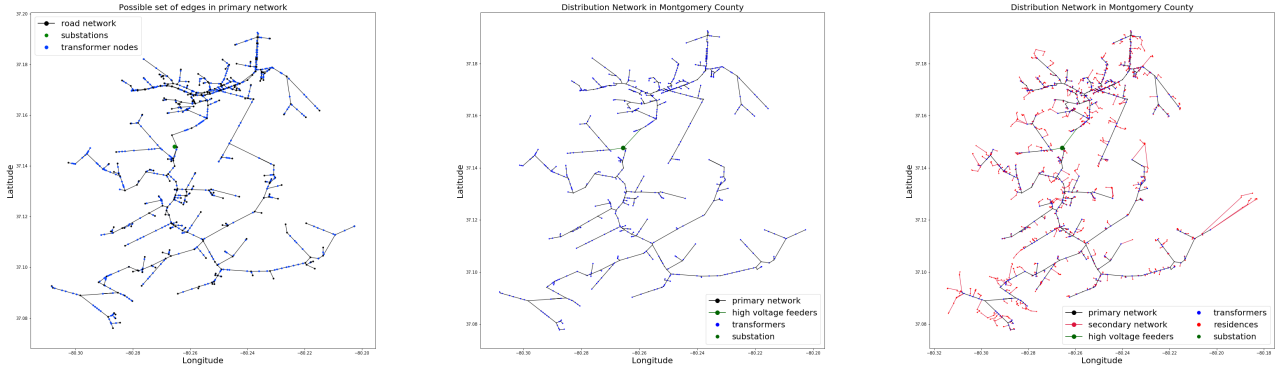

Figure 4: Primary distribution network generated for a substation in Montgomery county of south-west Virginia, USA. The left figure shows the set of possible road network edges from which the primary network is identified along with local transformers which are to be connected. The middle figure shows the optimal primary distribution with three feeders (green edges) originating from the substation. The right figure shows the entire generated synthetic network with primary and secondary sub-networks.

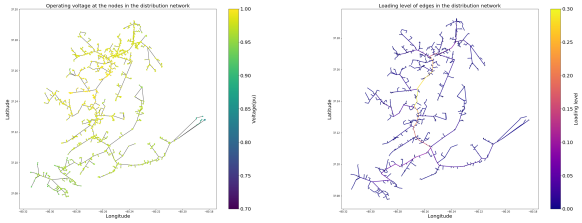

Figure 5: The left figure shows the operating voltage at different points in the network when individual customers are consuming average load demand. The right figure shows the loading level of every line in primary and secondary distribution networks.

Mapping of nodes. The first task is to map individual residential location to the nearest road network link. This is followed by the creation of secondary distribution network for each link. For example, Fig. 3 shows the creation of secondary distribution network for a set of residences mapped to a road link. The blue points along the road link indicate the probable locations of local transformers along the link and the red points are the residences which are required to be connected. The optimal secondary network is shown in the right figure which covers all the residences mapped to the road link and rooted at the local transformer nodes.

Creation of networks. The secondary network is generated for all road links mapped to at least one residence. The local transformers along the road links are identified as output of secondary network creation algorithm. The total number of identified local transformers is nearly 14000. The final task is to generate the primary distribution network which connects these local transformers to the substations. For this purpose, we divide the entire network into separate regions as shown in Fig. 2. The identified primary network originating from one substation is shown in Fig. 4. The first figure shows road network consisting of possible edges and the local transformers which are to be connected. The second figure shows the generated optimal primary network. Finally, the last figure shows the entire synthetic distribution network (primary and secondary) generated for the same substation.

Operational validation. The generated network is validated for the operational constraints in this section. A power flow case is run considering that all residential customers are consuming average hourly load. The operating voltage at each node is recorded and shown in the left figure of Fig 5 through a colored map. It is observed that the voltage profile is within acceptable limits of 0.95 to 1.0 pu. The voltage profile can be further improved by optimally placing capacitors along the network. The loading level of each line is also depicted in the right figure of Fig 5. The loading level of each line is observed to be significantly less than the rating of the line. This completes the operational validation of the network where we can conclude that when all residential customers consume average hourly load, the power distribution network operates in a *secure* state with voltages and line flows within acceptable operational limits.

Structural validation. In this section, we perform a structural validation of the generated synthetic distribution networks with respect to a sample of original power distribution system available for a small residential location in Montgomery county of south-west Virginia, USA. First, we ensure that the created network has a radial configuration by checking the number of cycles in it. Thereafter, we compare the degree distribution for the two networks in Fig. 6. The comparison is quantified through the Kolmogorov Smirnov test results shown in Table 4 for the synthetic distribution network generated for five substations.

5 CONCLUSION

This paper proposes a methodology to generate synthetic distribution networks for a particular geographical location based on available

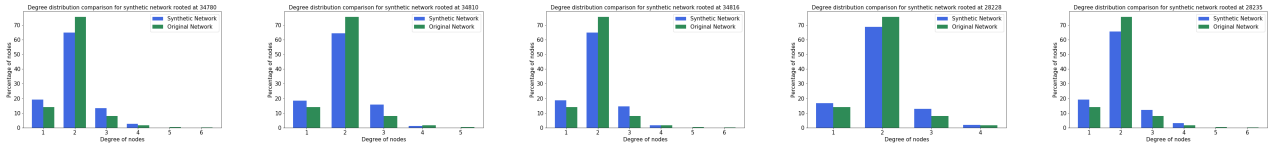

Figure 6: Comparison of degree distribution between the synthetic network and original network.

Table 4: Kolmogorov Smirnov test for comparing degree distribution of synthetic and original networks.

Substation ID	28228	28235	34780	34810	34816
D-Statistic	0.0335	0.0173	0.0532	0.0198	0.1011
p-value	0.9819	0.9999	0.6612	0.9999	0.0466

road network data. The generated network connects individual residential customers to substations while maintaining a radial configuration. Additionally, the network is created such that the overall length of overhead lines is minimized which is similar to planning methodologies undertaken by distribution companies. This ensures that generated synthetic distribution network is realistic and can be used to represent the network of the geographic location accurately. Furthermore, the consideration of several engineering and economic aspects in inferring the network has led us to identify certain parameters whose values can be modified to generate ensembles of synthetic networks.

A PROOF OF PROPOSISION 1

Statement of Proposition 1. A graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with reduced branch-bus incidence matrix $\mathbf{A_H}$ and residential node power consumption vector $\mathbf{p_H}$, with strictly positive entries, is connected if and only if there exists a vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{E}|}$, such that (5a) is satisfied.

PROOF. The proof follows along similar lines to proof of Proposition 1 in [**?**]. For the sake of completeness, this has been put in this paper. Proving by contradiction, suppose $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is not connected and there exists $\mathbf{f} \in \mathbb{R}^{|\mathcal{E}|}$ satisfying the proposed equality. If $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is not connected, then there exists a connected component $\mathcal{G}_\mathcal{S}(\mathcal{V}_\mathcal{S}, \mathcal{E}_\mathcal{S})$ which is a maximal connected subgraph with $\mathcal{V}_\mathcal{S} \subset \mathcal{V}$ and $\mathsf{T}_l \bigcap \mathcal{V}_\mathcal{S} = \emptyset$. Let $\mathbf{A}_\mathcal{S}$ denote the bus incidence matrix of $\tilde{\mathcal{G}}_\mathcal{S}$. By definition, it holds that $\mathbf{1}^T \mathbf{A}_\mathcal{S} = \mathbf{0}$.

Since graph $\mathcal{G}_\mathcal{S}(\mathcal{V}_\mathcal{S}, \mathcal{E}_\mathcal{S})$ is a maximal connected subgraph of $\mathcal{G}$, there exists no edge (i, j) with $i \in \mathcal{V}_\mathcal{S}$ and $j \in \mathcal{V}_{\overline{\mathcal{S}}}$, where $\mathcal{V}_{\overline{\mathcal{S}}} = \mathcal{V} \setminus \mathcal{V}_\mathcal{S}$. Since the order of rows and columns of $\mathbf{A_H}$ are arbitrary, we can partition without loss of generality as

$$\mathbf{A_H} = \begin{bmatrix} \mathbf{A}_{\overline{\mathcal{S}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_\mathcal{S} \end{bmatrix}$$

We can partition vectors $\mathbf{f}$ and $\mathbf{p_H}$ conformably to $\mathbf{A_H}$ to get the following equality

$$\begin{bmatrix} \mathbf{A}_{\overline{\mathcal{S}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_\mathcal{S} \end{bmatrix} \begin{bmatrix} \mathbf{f}_{\overline{\mathcal{S}}} \\ \mathbf{f}_\mathcal{S} \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{\overline{\mathcal{S}}} \\ \mathbf{p}_\mathcal{S} \end{bmatrix}$$

where $\mathbf{p}_\mathcal{S}, \mathbf{p}_{\overline{\mathcal{S}}}$ are respectively the power consumptions at residence nodes $i \in \mathcal{V}_\mathcal{S}$ and $i \in \mathcal{V}_{\overline{\mathcal{S}}}$ stacked together. From the second block, it implies that

$$\mathbf{A}_\mathcal{S} \mathbf{f}_\mathcal{S} = \mathbf{p}_\mathcal{S} \Rightarrow \mathbf{1}^T \mathbf{p}_\mathcal{S} = \mathbf{1}^T \mathbf{A}_\mathcal{S} \mathbf{f}_\mathcal{S} = \mathbf{0} \tag{14}$$

Since all entries of $\mathbf{p_H}$ are positive from the initial assumption, we have $\mathbf{1}^T \mathbf{p}_\mathcal{S} \neq 0$ which contradicts (14) and completes the proof. □

B RESULTS FOR ENSEMBLE NETWORKS

Ensemble of synthetic networks The proposed methodology uses several engineering and economic principles to generate realistic synthetic distribution networks. Such principles have been included through the different parameters in the optimization framework. In this section, we first list down the parameters which can be varied to create ensembles of synthetic distribution networks. These parameters are shown in Table 5.

Table 5: List of parameters to create ensembles of networks

Parameter	Line flow rating	Penalty factor	Substation rating	Line type/ parameters
Notation	$\overline{f}$	λ	$\overline{s}$	r

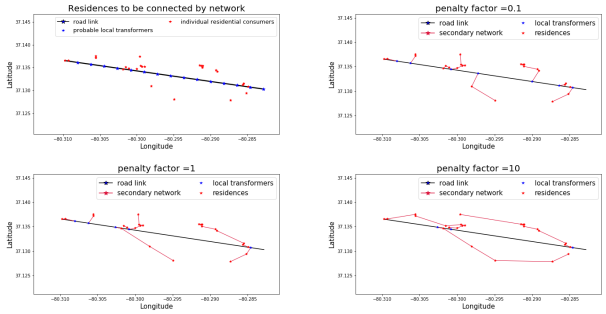

Figure 7: Ensemble network creation for the same set of residential buildings. The penalty factor in the objective function of optimization problem to create secondary network is varied in this case. The first figure shows the set of residences which are to be connected and probable locations of local transformers. The remaining three figures show generated ensembles of secondary network when the penalty factor is varied for three different values.

Now, we show how varying one such parameter creates ensembles of networks for the same set of residential building points. The penalty factor (λ) in the objective function (7) is varied for factors of $\lambda = 0.1, 1, 10$. The generated ensemble networks are shown in Fig. 7. As we increase the penalty factor, the optimization program tends to find networks with longer edges and lesser number of root nodes (because the objective function penalizes multiple root nodes).