Ensembles of realistic power distribution networks

Rounak Meyur[a,1], Anil Vullikanti[a,b], Samarth Swarup[a], Henning S. Mortveit[a], Virgilio Centeno[c], Arun Phadke[c], H. Vincent Poor[d,1], and Madhav V. Marathe[a,b,1]

Contributed by H. Vincent Poor; received April 5, 2022; accepted September 12, 2022; reviewed by Peter Sauer and Anna Scaglione

The power grid is going through significant changes with the introduction of renewable energy sources and the incorporation of smart grid technologies. These rapid advancements necessitate new models and analyses to keep up with the various emergent phenomena they induce. A major prerequisite of such work is the acquisition of well-constructed and accurate network datasets for the power grid infrastructure. In this paper, we propose a robust, scalable framework to synthesize power distribution networks that resemble their physical counterparts for a given region. We use openly available information about interdependent road and building infrastructures to construct the networks. In contrast to prior work based on network statistics, we incorporate engineering and economic constraints to create the networks. Additionally, we provide a framework to create ensembles of power distribution networks to generate multiple possible instances of the network for a given region. The comprehensive dataset consists of nodes with attributes, such as geocoordinates; type of node (residence, transformer, or substation); and edges with attributes, such as geometry, type of line (feeder lines, primary or secondary), and line parameters. For validation, we provide detailed comparisons of the generated networks with actual distribution networks. The generated datasets represent realistic test systems (as compared with standard test cases published by Institute of Electrical and Electronics Engineers (IEEE)) that can be used by network scientists to analyze complex events in power grids and to perform detailed sensitivity and statistical analyses over ensembles of networks.

synthetic networks | digital twin | power distribution networks | mixed integer programming | ensemble of networks

Significance

The availability of power distribution network datasets is essential to researchers when testing different control algorithms prior to their actual implementation and deployment on distribution systems. The proprietary nature of actual distribution system data effectively prevents them from being publicly available for use. Most educators and researchers rely on standard test systems published by Institute of Electrical and Electronics Engineers (IEEE), which neither are representative of actual distribution networks nor provide challenges in terms of the size of actual systems. We develop a framework that employs freely available data to construct ensembles of synthetic distribution networks resembling their physical counterparts with realistic scale and complexity. These synthetic networks are geographically embedded and include realistic residential demand profiles, and thus, they can be used to study smart grid applications.

A reliable power grid constitutes the backbone of a nation's economy, providing vital support to various sectors of society and other civil infrastructures. Power distribution networks are created in a bottom-up fashion connecting small clusters of residential loads to distribution substations, thereby electrifying the entire community. These bear a structural resemblance to other common networked infrastructures, such as transportation, communication, water, and gas networks, and are often interdependent in their operations (1). One may use these resemblances and interdependencies to infer one network from available data about these other networks.

Over the past decade, power engineers have aimed to enhance the resilience of power systems through incorporation of distributed energy resources (DERs) by deploying advanced metering and monitoring infrastructures (2) and by performing system vulnerability and criticality assessments, thus reinforcing cybersecurity (3). Furthermore, spatiotemporally variable consumer load demands, such as electric vehicles (EVs), along with an evolving trend toward a distributed operation of the power grid have posed new challenges to system planners and operators (4). Network scientists have emphasized the importance of realistic power network models for accurate analysis as opposed to stylized statistical models (5–7). In order to address these challenges, there is a pressing need for openly available data containing realistic grid topologies along with available geographic information. For example, in the context of power grid expansion planning, the current grid information in conjunction with geographical knowledge of wind maps and solar trajectories can aid in optimized power grid expansion while introducing DERs in the grid (8). Similarly, for system vulnerability analysis, a geographic correlation of grid information with cyclone/hurricane paths can help us identify critical sections in the network and raise preparedness levels for natural disasters (9). Further, a detailed knowledge about individual residential load usage and consumer behavior can help address policy-level questions. Examples of such problems include identifying the impact of EV adoption and DER penetration on the current power grid infrastructure as the society moves toward net-zero emission (10, 11).

Simulation-based frameworks capable of performing spatiotemporally resolved simulations can be utilized to analyze the impact of such evolving trends and analyze system vulnerability. Such assessments are useful to system planners aiming to make decisions

Author contributions: R.M., A.V., S.S., H.S.M., A.P., H.V.P., and M.V.M. designed research; R.M., A.V., S.S., H.S.M., V.C., A.P., H.V.P., and M.V.M. performed research; R.M., A.V., V.C., and M.V.M. contributed new reagents/analytic tools; R.M., A.V., H.V.P., and M.V.M. analyzed data; and R.M., A.V., S.S., H.S.M., V.C., A.P., H.V.P., and M.V.M. wrote the paper.

Reviewers: P.S., University of Illinois at Urbana–Champaign; and A.S., Cornell Tech.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: rm5nz@virginia.edu, poor@princeton.edu, or marathe@virginia.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2205772119/-/DCSupplemental.

Published October 10, 2022.

about infrastructure development and to operators while handling emergency system conditions. A common drawback of this simulation-based approach is that it requires detailed information regarding the power network and associated components, such as locations and capacities of generation, load demands, and line parameters (12–15). Furthermore, since the majority of grid infrastructure advancements are being done at the low-voltage (LV) distribution level, a high-resolution analysis of the power distribution systems is important. This necessitates a comprehensive knowledge of customer energy use profiles, customer behavior, and most importantly, the distribution network topology that connects them. Most such data are, at best, partially available but more typically, are not available at all due to their proprietary nature (16). The lack of such openly available detailed real-world data has been identified as a significant hurdle for conducting research in smart grid technology (17). In recent years, there has been an increased interest in generating synthetic power network data to address this issue. The synthetic data are not the real-world data; rather, they are generated by mathematical models operating on openly available information and are designed to ensure the generated data are similar to the real-world data, thus allowing them to be used as a proxy for the actual data. Some examples of synthetic power grid data include synthetic transmission networks (18–21), synthetic distribution networks (22–25), and synthetic residential customer energy usage data (26–28).

In this work, we focus on constructing a modular framework for generating synthetic power distribution networks: that is, networks connecting individual residential customers to the distribution substations. We present a first-principles approach, where we generate an optimal synthetic distribution network connecting all residences in a given geographic region to the high-voltage (HV) substations through medium-voltage (MV) and LV networks. We use the example of Montgomery County of southwest Virginia (United States) to create the synthetic power distribution networks, consider all residences and HV substations within the state boundary, and connect them through the synthetic distribution network.

In this context, there are two critical questions. 1) Are the created networks the only feasible networks connecting the residences and substations? 2) How similar are the created synthetic networks and the actual power distribution networks?

To tackle the first problem, we present a methodology for generating an ensemble of feasible synthetic power distribution networks for a given region. In the literature related to modeling real-world networks, statistical physics has been used to learn significant structural patterns from an ensemble of networks (29) and thereby, to help in network reconstruction from incomplete data. In recent years, statistical aspects of the power networks have drawn the attention of the scientific community for similar reasons. A dataset spanning 70 years for the electric power grid of Hungary has been studied (30) for small-world and scale-free properties. Due to a lack of real-world power distribution data, ensembles of distribution networks, which have significant resemblance to actual networks, can suffice for a detailed statistical analysis.

The generated synthetic networks require detailed validation before they can be used as a substitute for actual networks for various applications. To this end, the networks need to be compared against actual distribution networks in terms of their structural properties as well as their power engineering attributes (31). Hence, the second problem deals with the comparison of synthetic and actual networks using suitable metrics meaningful to power distribution networks, which follow a particular structure. In this work, we have compared the created synthetic networks for the town of Blacksburg (in Montgomery County of southwest Virginia, United States) against actual networks obtained from a power company operating in the same region. In addition to comparing standard graph attributes, such as degree and hop distributions, we compute the difference in geometries between the actual and synthetic networks and provide a measure of deviation.

Our contributions include 1) a holistic modular framework to create synthetic power distribution networks that satisfy structural and power engineering constraints along with an accurate representation of residential load demand profiles, 2) a method to create an ensemble of networks by generating multiple feasible networks for a given region, and 3) an open dataset consisting of ensembles of distribution networks for Montgomery County of southwest Virginia (United States). This dataset is unique in terms of both size and details. The geographically embedded networks, along with the detailed residential customer usage data, become suitable tools for system-wide planning studies and for addressing policy-level questions.

Related Work

In recent years, a substantial amount of work has gone into creating synthetic HV transmission networks (18, 19) or combinations of transmission and distribution networks (32, 33). The primary focus of these papers is to model the transmission grid with a high level of resemblance to the actual grid.

For distribution networks, Schweitzer et al. (24) were one of the first to analyze real power distribution networks, learn statistical distributions of network attributes from an extensive dataset of actual distribution networks in the Netherlands, and create synthetic networks that preserve these attributes. The reference network model (RNM) framework (22, 23, 34) and some of its variants (35, 36) have proposed heuristics to generate synthetic distribution networks that satisfy structural and power engineering constraints. These heuristics include clustering load groups to identify feeders (21, 23), identifying substation locations from cluster centroids (22, 23), and constructing networks using a minimum spanning tree algorithm (22, 34, 35). Most of these papers do not use individual residence locations and have populated the created synthetic networks with random residences or aggregated loads to zip code centers. Some of these works used a top-down approach, where feeder networks are generated and followed by populating with random loads (21, 35). In ref. 37, the authors use generative adversarial networks (GANs) to create synthetic power networks. However, the approach is inherently data intensive and requires a large number of samples for training, making it practically challenging to use. Here, we propose a rigorous mathematical framework that provides optimality guarantees on the quality of networks created. The ensemble of synthetic networks created can potentially be used to train GANs.

The RNM (22, 23) is an important heuristic-based planning tool for efficient investment options in distribution grid planning. It uses OpenStreetMap (OSM) and relevant geographic data to create distribution networks in a given region. The comparison of networks generated by the RNM with actual power distribution networks shows that the real and synthetic networks are quite similar (31). However, the methods used to compare such networks were somewhat ad hoc. In contrast, in ref. 24, the authors performed a statistical fit of the distributions of network attributes and performed a numerical comparison, yielding a more rigorous approach to measuring network similarity. The RNM framework uses four independent layers (namely logical, topological, electrical, and continuity of supply) to assign constraints while creating the networks. Although this approach is natural, the set of constraints is not mathematically well defined to the extent that it can be reproduced. For instance, the framework uses a set of heuristics to satisfy the constraints, which does not always guarantee a feasible solution. Furthermore, several steps in the heuristic-based method involve user-defined parameters, which lead to multiple possible networks for different choices. Furthermore, these papers do not consider the creation of ensembles of networks. We present a summary of previous results in *SI Appendix*, Table S1.

Recent works (26, 38) have provided detailed synthetic residential demand models along with household geographic footprints. We create synthetic distribution networks connecting substations to these individual residence locations. Our methods differ from the earlier works in the following ways. 1) Instead of populating with imaginary demand profiles, we have used behavior-based consumer load modeling, which results in an accurate representation of household load demand profiles. 2) Unlike other heuristic-based approaches, the structural and power engineering constraints are mathematically well defined, which makes the framework reproducible for creating other networks with similar constraints. 3) We use actual substation locations obtained from ref. 39, and the optimal feeder locations are identified as an output of our optimization framework. 4) We propose a method to create an ensemble of realistic power networks.

A wide variety of graph comparison methods have been studied in the literature. Tantardini et al. (40) analyze multiple graph comparison methods, which include comparing whole graphs as well as small portions of the graph known as motifs. Several methods for assessing structural similarities of graphs have also been studied (41, 42). However, none of the comparison methods consider the node and edge geometries of the graphs. Edit distance, or evaluating the minimum number of edit operations to reach from one network to the other, has been widely used to compare networks having structural properties (43–45). Among these works, Riba et al. (45) have used Hausdorff distance between nodes in the network to compare network geometries. Morer et al. (46) include edge geometry–based comparison and propose an "efficiency" metric to measure the distance of a network (where edges have nonstraight-line geometries between nodes) from its most optimal version (where each edge has a straight-line geometry).

Methods

Datasets Used. We use open-source, publicly available information regarding several infrastructures to generate the synthetic distribution networks (more details are presented in *SI Appendix*, Table S3): 1) road network data from OSM (47), 2) geographic locations of HV (greater than 33 kV) substations from datasets published by Energy Information Administration (EIA) (39), and 3) residential electric power demand information developed in earlier work from our research group (26). We also obtained actual power distribution networks for the town of Blacksburg to validate the synthetic networks.

Approach. *Algorithm 1* summarizes the steps we use in the paper. The synthetic distribution networks are constructed in two steps using a bottom-up approach. First, we identify local pole-top transformers along the road network and connect the residential buildings to them to create the LV (208 to 480 V) secondary network (*Step 1*). Thereafter, we use the road network as a proxy to construct the MV (6 to 11 kV) primary network connecting the local transformers placed along roads to the substations (*Step 2*). To construct the ensemble of synthetic networks, we propose a Markov chain starting from the already created network to a variant network, which is also a feasible distribution network (*Step 3*). Finally, we add attributes to nodes and edges in each network (*Step 4*) to create an ensemble of synthetic power distribution networks. Several aspects of the first two and last steps are similar to the approach taken in earlier papers. The difference is in the specifics of problem formulation and the resulting algorithmic approach. The third step that creates the ensemble of networks has largely not been explored in the context of distribution networks. Fig. 1 shows the proposed framework for constructing and validating ensembles of realistic power distribution networks.

```
Algorithm 1  Create ensemble of synthetic networks
Input Set of residences $\mathcal{H}$, set of substations $\mathcal{S}$, road network $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$,
      required ensemble size $N$
Step 1: Construct LV secondary network.
    a: Map residences to nearest road network link.
    b: Connect residences to local transformers along road link.
Step 2: Construct MV primary network.
    a: Map local transformers to nearest substation.
    b: Use road network as proxy to connect transformers to substation.
Step 3: Construct an ensemble of networks.
    a: Construct Markov Chain $\mathcal{M}$ to create a variant from an existing network.
    b: Run $\mathcal{M}$ to create $N$ variant networks.
Step 4: Add additional attributes to nodes and edges of each network in the
        ensemble as follows.
    a: Assign one of the three phases (A,B,C) to each residence.
    b: Assign a distribution line type to each edge.
Output Ensemble of $N$ attributed networks.
```

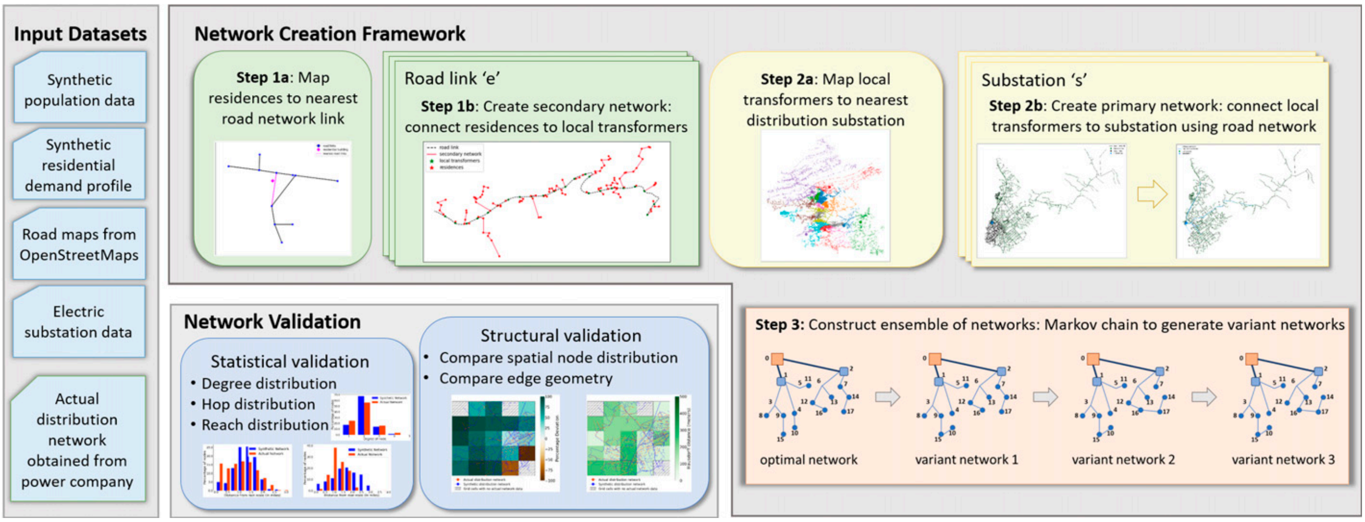
Input Datasets
Synthetic population data
Synthetic residential demand profile
Road maps from OpenStreetMaps
Electric substation data
Actual distribution network obtained from power company
Network Creation Framework
Step 1a: Map residences to nearest road network link
Road link 'e'
Step 1b: Create secondary network: connect residences to local transformers
road link
secondary network
local transformers
residences
Step 2a: Map local transformers to nearest distribution substation
Substation 's'
Step 2b: Create primary network: connect local transformers to substation using road network
Network Validation
Statistical validation
• Degree distribution
• Hop distribution
• Reach distribution
Structural validation
• Compare spatial node distribution
• Compare edge geometry
Step 3: Construct ensemble of networks: Markov chain to generate variant networks
optimal network
variant network 1
variant network 2
variant network 3

Fig. 1. Proposed framework for constructing ensembles of realistic power distribution networks. The framework uses the input datasets and constructs an ensemble of networks using the steps detailed in *Algorithm*. The created networks are validated against actual power distribution networks.

***Step 1.* Constructing Secondary Networks.** We extract residence and road network data for the geographic region. Let $\mathcal{H}$ be the set of residences and $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$ be the road network graph. We evaluate a many-to-one mapping $\mathcal{F}_M\colon \mathcal{H} \to \mathcal{E}_R$ such that each residence $h \in \mathcal{H}$ is mapped to the nearest road network link $e \in \mathcal{E}_R$. The inverse mapping $\mathcal{F}_M^{-1}$ defined by $\mathcal{F}_M^{-1}(e) = \{h \in \mathcal{H}; \mathcal{F}_M(h) = e\}$ provides the set of residences assigned to each road link $e \in \mathcal{E}_R$.

The secondary network creation problem (denoted by $\mathcal{P}_{\text{sec}}$) is defined for each road link $e \in \mathcal{E}_R$. We provide more details for $\mathcal{P}_{\text{sec}}$ in *SI Appendix*. The objective is to identify local transformers $\mathcal{V}_T(e)$ along the link and connect them to the assigned residences $\mathcal{F}_M^{-1}(e)$, thereby constructing the secondary distribution network $\mathcal{G}_S(e)$ with node set $\mathcal{V}_S(e) = \mathcal{V}_T(e) \cup \mathcal{F}_M^{-1}(e)$ and edges $\mathcal{E}_S(e)$. We impose structural constraints to connect residences in chains, ensuring the tree network structure so that the created networks mimic their physical counterpart.

Problem 1 ($\mathcal{P}_{\text{sec}}$ Construction). *Given a road link $e \in \mathcal{E}_R$ with a set of residences $\mathcal{F}_M^{-1}(e)$ assigned to it, construct an optimal forest of trees, $\mathcal{G}_S(e)$, rooted at points (local transformers) along the link and connecting the residences.*

The problem $\mathcal{P}_{\text{sec}}$ is modeled as a mixed integer linear program (MILP), which usually requires exponential computation time. We use different heuristics to reduce the number of binary variables, which in turn, reduce the overall time complexity. A formal problem statement for the problem has been provided in *SI Appendix* along with our approach to solve it. The secondary network creation process can be executed simultaneously for different road links $e \in \mathcal{E}_R$ in the geographic region. In our framework, we execute the task sequentially for all edges in a county, with the entire sequence performed simultaneously for different counties. The secondary network generated for the region is

$$\mathcal{G}_S = \bigcup_{e \in \mathcal{E}_R} \mathcal{G}_S(e) = \bigcup_{e \in \mathcal{E}_R} \mathcal{P}_{\text{sec}}\left(e, \mathcal{F}_M^{-1}(e)\right).$$

***Step 2.* Constructing Primary Networks.** The secondary network results in local transformer nodes $\mathcal{V}_T = \bigcup_{e \in \mathcal{E}_R} \mathcal{V}_T(e)$ along the road network links. The goal of the primary network construction is to connect these transformers to the set of substation nodes $\mathcal{S}$ using the road network as proxy. First, we define a many-to-one mapping $\mathcal{F}_V\colon \mathcal{V}_T \to \mathcal{S}$ based on a Voronoi partitioning. The details of this mapping are provided in *SI Appendix*. We are interested in the inverse mapping $\mathcal{F}_V^{-1}(s) = \{t \in \mathcal{V}_T; \mathcal{F}_V(t) = s\}$, which assigns a group of transformers to each substation node.

The primary network creation problem (denoted by $\mathcal{P}_{\text{prim}}$) is defined for each substation node $s \in \mathcal{S}$, and the goal is to create a minimum-length primary network $\mathcal{G}_P(s)$ connecting substation node s to the mapped transformers $\mathcal{F}_V^{-1}(s)$ using road network $\mathcal{G}_R$ as proxy, such that the following set of structural and operational constraints is valid. 1) The network should be a tree rooted at the substation, 2) all transformer nodes are to be connected, and 3) all nodes should have acceptable voltages (based on American National Standards Institute [ANSI] standards between 0.95 and 1.05 per unit [pu]) when the residential customers are consuming average hourly loads.

Problem 2 ($\mathcal{P}_{\text{prim}}$ Construction). *Given a substation $s \in \mathcal{S}$ with an assigned set of local transformer nodes $\mathcal{F}_V^{-1}(s)$, construct a tree network $\mathcal{G}_P(s)$ using the road network $\mathcal{G}_R$ as a proxy that connects all local transformers while ensuring acceptable node voltages by power engineering standards.*

We formulate an MILP to solve the problem $\mathcal{P}_{\text{prim}}$, which might take exponential computation time. A formal problem statement for the same has been provided in *SI Appendix*. We do not use any heuristic to reduce the computational complexity, which is determined by the size of the underlying road network (used as the proxy). On many occasions, we terminate the optimization program reaching an optimal solution in order to reduce the running time. This has resulted in the constructed network being a near-optimal solution but with an acceptable optimality gap of 0 to 5%. In our framework, we execute the task of primary network creation simultaneously for all the substations in the geographic region. The created primary network $\mathcal{G}_P$ for the entire region is

$$\mathcal{G}_P = \bigcup_{s \in \mathcal{S}} \mathcal{G}_P(s) = \bigcup_{s \in \mathcal{S}} \mathcal{P}_{\text{prim}}\left(s, \mathcal{F}_V^{-1}(s)\right).$$

***Step 3.* Constructing Ensembles of Networks.** In this section, we address the problem of creating multiple realizations of the distribution network that connects the residences to substations. Albeit that the modification of user-defined parameters in $\mathcal{P}_{\text{sec}}$ and $\mathcal{P}_{\text{prim}}$ can produce different realizations of synthetic networks, the procedure is computationally expensive since optimization problems of similar order need to be solved. We propose a methodology that uses the already created (near)-optimal primary network for a region and creates an ensemble of synthetic networks by reconnecting the transformer nodes in a different manner from the (near)-optimal primary network while maintaining the structural and power engineering operational constraints. Thereafter, we connect the residences in the same way as in the optimal secondary network. Thus, we construct an ensemble of networks where each network is a combination of a variant primary network and the optimal secondary network (solution of $\mathcal{P}_{\text{sec}}$). The variant primary networks are "feasible" (but not necessarily "optimal") solutions of $\mathcal{P}_{\text{prim}}$.

Problem 3. *Given a near-optimal primary network $\mathcal{G}_P^0 := (\mathcal{V}_0, \mathcal{E}_0)$ constructed using the underlying road network graph $\mathcal{G}_R := (\mathcal{V}_R, \mathcal{E}_R)$, construct N variants of the primary network $\mathcal{G}_P^1, \cdots, \mathcal{G}_P^N$ by identifying respective edge sets $\mathcal{E}_1, \cdots, \mathcal{E}_N \subseteq \mathcal{E}_R$ such that the networks are feasible solutions of $\mathcal{P}_{\text{prim}}$.*

We consider the ensemble of networks generation problem for each substation s and the mapped transformer nodes $\mathcal{F}_V^{-1}(s)$. Let $\mathcal{F}_{\text{feas}}$ denote the set of feasible solutions of $\mathcal{P}_{\text{prim}}\left(s, \mathcal{F}_V^{-1}(s)\right)$. From here on, we omit the dependency on s in our notation. We design a Markov chain $\mathcal{M}$ to create variant networks with each state denoting a feasible realization of the network $\mathcal{G}_P^t \in \mathcal{F}_{\text{feas}}$. The steps involved in transitioning from the primary network $\mathcal{G}_P^t := (\mathcal{V}_t, \mathcal{E}_t)$ to $\mathcal{G}_P^{t+1} := (\mathcal{V}_{t+1}, \mathcal{E}_{t+1})$ are described below.

Let $\mathcal{F}_{\text{rstr}}(e) = \{\mathcal{G} := (\mathcal{V}, \mathcal{E}) \in \mathcal{F}_{\text{feas}} : e \notin \mathcal{E}\}$. If $\mathcal{F}_{\text{rstr}}(e) \neq \emptyset$, we select a random edge $e \in \mathcal{E}_t$ to be deleted with probability $1/|\mathcal{E}_t|$ and then, pick $\mathcal{G}_P^{t+1} := (\mathcal{V}_{t+1}, \mathcal{E}_{t+1}) \in \mathcal{F}_{\text{rstr}}(e)$ uniformly at random; else, $\mathcal{G}_P^{t+1} = \mathcal{G}_P^t$. The ensemble of synthetic power distribution networks for the region is

$$\mathcal{E} := \mathcal{G}_S \bigcup \{\mathcal{G}_P^t : t = 1, \ldots, N\}.$$

***Step 4.* Postprocessing of Networks.** The final step of our framework involves the addition of attributes and labels to nodes and edges to each network in the ensemble. We compute the required distribution line ratings for each edge in the network. We assign a suitable type of distribution line from the catalog of distribution lines (48) and add edge attributes accordingly. We include positive sequence impedance (resistance and reactance) for each edge in the network. Additionally, we use a framework to assign one of the three phases (A, B, or C) to each residence in the network. The phase assignment ensures that the three phases are balanced at each substation feeder. We add the assigned phase as a node attribute to the network. The details of this step have been provided in *SI Appendix*. The node and edge attributes are listed in *SI Appendix*, Table S8.

Although we have assigned one of the three phases to each residence in the network, we do not consider three phase circuits with different transformer configurations (wye and delta). Therefore, we limit the created synthetic networks with only positive sequence impedance. To this end, such networks can be useful in performing studies involving balanced loads across three phases.

Results: Synthetic Network Attributes

Degree, Hop, and Reach Distribution. In this section, we compare the statistical attributes of the synthetic distribution networks created for rural and urban areas. The degree of a node in a network denotes the number of edges connected to it. The degree distribution gives an idea about the connectivity within the network. The "hop" of a node from the substation (root) node is defined as the number of edges between them. Hence, the "hop

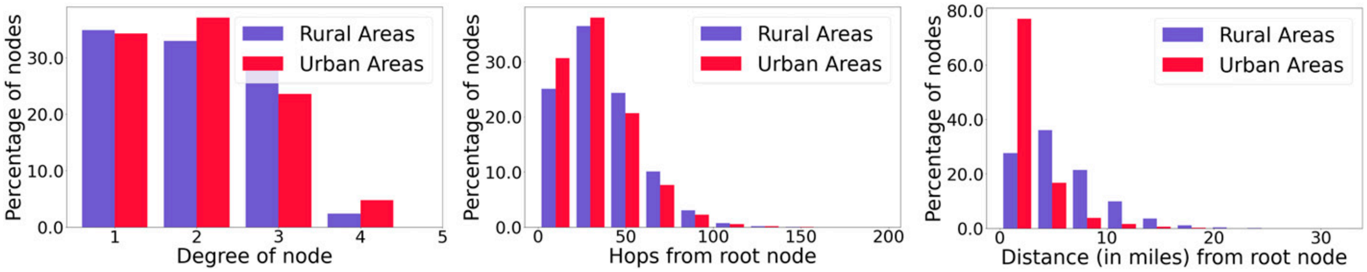

Fig. 2. Plots showing degree distribution (*Left*), hop distribution (*Center*), and reach distribution (*Right*) in rural and urban areas. Colors depict network attributes of urban vs. rural areas. The degree and hop distribution are similar for both rural and urban regions. The reach distribution of urban networks peaks at small value since the distribution network nodes are more closely placed to the substation than rural areas.

distribution" provides an idea about the radial layout of nodes around the root substation node. Finally, we define "reach" of a node as the length of the network (in miles) connecting it to the substation. The associated "reach distribution" of a network becomes a relevant statistic in the context of networks with associated geographic attributes since it provides a distance metric to the hop distribution.

Fig. 2 shows a comparison of degree, hop, and reach distributions in urban and rural distribution networks. We observe that the degree and hop distributions are fairly similar. However, the reach distribution differs for rural and urban areas. In the case of urban areas, we notice that a majority of nodes are located very close to the substation, whereas rural areas are often characterized by long-length network edges. This observation is also consistent with the distribution of residences in urban and rural regions, where rural regions have more widely spread out residences than urban areas.

Network Motifs. Network motifs are interesting subgraphs that build up the entire network. Network motifs have been used as a metric to understand network resilience in earlier work (49). We focus our attention to small-size subgraphs with at most four nodes. Since the created distribution networks are tree graphs, we are interested in two types of network motifs: 1) four-node path and 2) four-node star. Fig. 3 shows the number of four-node motifs in the synthetic distribution networks. The two colors show the results for urban and rural networks separately. The star motifs are higher for urban networks as compared with rural networks of similar size. This can be explained from the observation in degree distribution, where we notice that urban networks have higher fractions of nodes with degree 4. A single node with degree 4 results in $\binom{4}{3} = 4$ counts of four-node star motifs.

Features in Ensemble of Networks. We create ensembles of distribution networks for Montgomery County in southwest Virginia. The entire network within Montgomery County is composed of 19 subnetworks (each fed by a different substation). We create an ensemble of 20 networks for each subnetwork and study the variation in network attributes over the ensembles. We plot the variation in degree, hop, and reach distributions in Fig. 4. The error bar shows the extent of variation in the ensemble. Fig. 5 shows variation in four-node path and star motif counts for the networks in each ensemble. The bar plots show the motif counts for each ensemble of networks, and the error bars (on top of each bar) depict the variation over the ensemble. We observe that the variation of network features over each ensemble is not significant. This shows that the networks are fairly close to each other, and each of them can be considered as a digital twin of the actual network. Thus, our framework is capable of creating an ensemble of synthetic distribution networks, which are statistically equivalent to each other. In order to create statistically different networks, the Markov chain in *Step 3* needs to be altered—deleting multiple random edges instead of one.

In general, an "ensemble" of networks consists of multiple structurally different networks that connect the same set of residences to the substation. Each synthetic network in the ensemble is a feasible network (has a tree structure and satisfies power engineering constraints) but is not the optimal-length network.

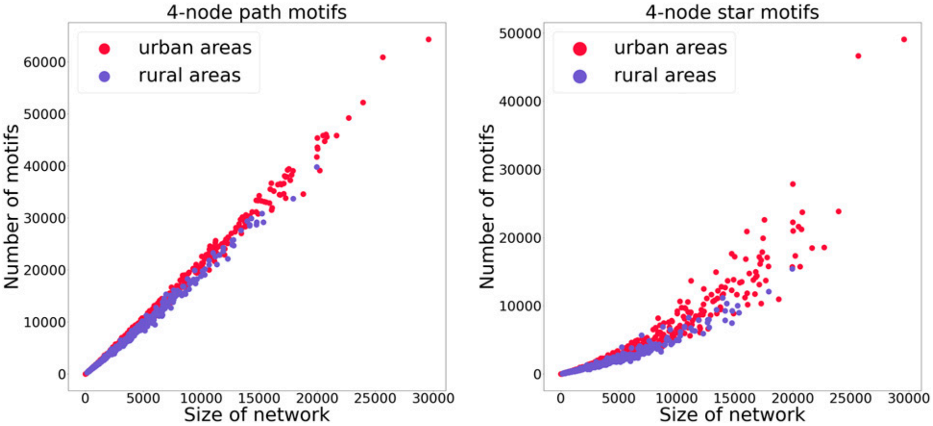

Fig. 3. Plots showing the number of four-node paths (*Left*) and four-node star motifs (*Right*) as a function of network size (measured as the number of nodes in the network). Colors depict motif numbers in urban vs. rural areas. Urban distribution networks have a larger number of star motifs than rural networks. In contrast, the path motif count does not differ significantly across rural and urban areas. Urban networks are often larger than rural networks as measured by the number of nodes due to the larger population size.

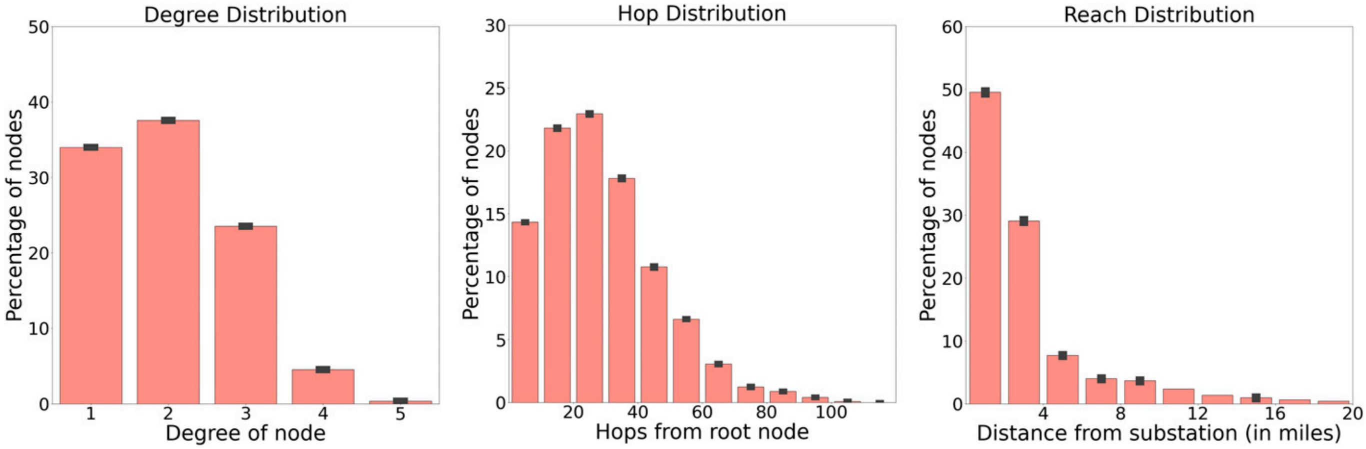

Fig. 4. Plots showing variation in degree distribution (*Left*), hop distribution (*Center*), and reach distribution (*Right*) for the ensemble of distribution networks created for Montgomery County of southwest Virginia. The error bars in the bar plots show the variation over the networks in the ensemble.

Therefore, we can consider it as a single random realization of the actual network. This allows us to perform analysis on an ensemble of networks instead of a single network and thereby, capture the deviation arising due to the different network structure in the ensemble.

Validation

In the earlier section, we presented the proposed framework to create synthetic distribution networks for a geographic region. The aim is to create networks that resemble their actual physical counterparts. We obtained real-world power distribution networks for the town of Blacksburg in southwest Virginia from a distribution company to validate the created networks. This network has been incrementally built over a long period of time with a close dependency on the population growth in the region. In contrast, our proposed framework uses the current population information with no consideration of any historical data. The created synthetic networks are optimal in terms of economic and engineering perspectives. Therefore, it is expected that there would be structural differences between the networks. Furthermore, the comparison methods need to be relevant in the context of distribution networks with associated geographic attributes.

This section compares the generated synthetic networks with the actual distribution network based on various operational, statistical, and structural attributes. The operational validation ensures that we observe similar node voltages and edge flows in both networks. This makes the networks suitable to be used by the scientific community to aid in their research. The methods to compare statistical attributes help us compare the overall connectivity properties of the networks. The comparison of structural attributes involving node and edge geometries enables us to validate the created synthetic networks on a much higher resolution. The results of the comparison show that the created networks bear a significant amount of resemblance to the actual networks.

Operational Validation. We compare voltages at the residences when they are connected to the actual and synthetic networks in Fig. 6, *Left*. We term this validation as operational validation, where the basic idea is that if we substitute the actual network with the synthetic network, we should see minimal voltage differences at the residences connected to either network. Here, the black dashed line denotes the identity line (exact same voltages), and green lines signify $\pm 0.4\%$ deviation from the identity line. We observe that the majority of residence voltages in the synthetic network remain within this $\pm 0.4\%$ regulation. We also compare the edge flows in the two networks through the histogram in Fig. 6, *Right*, which also bear a significant resemblance. We performed statistical fit of the flow distributions, and the Kullback–Leibler (KL) divergence is 0.15.

Statistical Validation. The created networks are expected to have similar graph attributes to the actual network. We

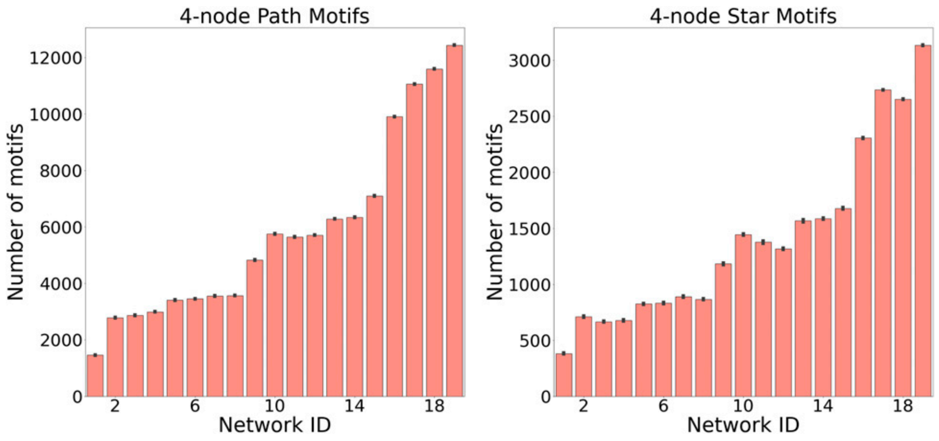

Fig. 5. Plots showing variation in the number of four-node path motifs (*Left*) and the number of four-node star motifs (*Right*) for the ensembles of distribution networks created for Montgomery County of southwest Virginia. Results are shown for 19 ensembles of varying size in the county fed by different substations. Each ensemble consists of 20 networks. The error bars in the bar plots show the variation over the networks in each ensemble.

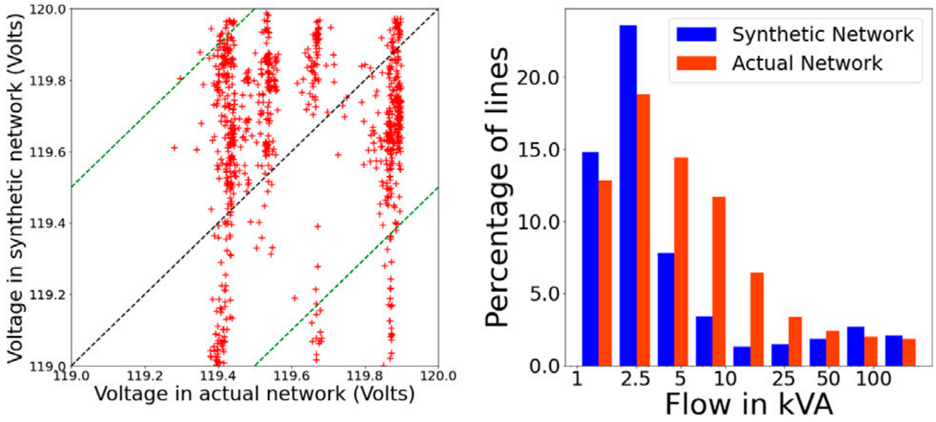

Fig. 6. Plots comparing the residential node voltages (*Left*) and edge power flows (*Right*) for actual and synthetic networks. The majority of residence voltages in the synthetic network are within ±0.4% voltage regulation of the voltages in the actual network. The edge flows in both networks follow similar distributions, with a computed KL divergence of 0.15.

focus on basic graph attributes, such as degree and hop distributions, and also, the newly defined reach distribution. Fig. 7 compares the synthetic and actual networks for the town of Blacksburg in southwest Virginia in terms of these statistical attributes. We use the KL divergence to compare each pair of distributions. KL-divergence values for various structural measures are as follows: 1) degree distributions: 0.0208; 2) hop distribution: 0.0323; and 3) reach distribution: 0.0096. The small KL-divergence values indicate that the real and synthetic networks are structurally very similar.

Structural Validation. One of the important aspects of our work is that the created synthetic networks have a geographic attribute associated with them. Therefore, we need to include network comparison methods that incorporate the geographic embedding while measuring the deviation. In this paper, we use a metric for geometry comparison (i.e., how the edge geometries in the networks deviate from each other). Due to the unavailability of actual network information for the entire region, we propose an effective way to compare the structural attributes of the networks. We divide the entire geographic region into multiple rectangular grid cells and perform comparison in each cell separately. In this way, we can omit the cells for which network data are missing.

We compare the Hausdorff distance between the edge geometries of the two networks in each rectangular grid cell. Let $\mathcal{P}_{\text{act}}$ and $\mathcal{P}_{\text{syn}}$ represent the set of points representing the geometries of the actual and synthetic networks. We define the Hausdorff distance between networks $\mathcal{G}_{\text{act}}$ and $\mathcal{G}_{\text{syn}}$ for a rectangular grid cell as

$$\mathsf{D}_{\mathsf{H}}^{\mathsf{CELL}}\left(\mathcal{G}_{\text{act}}, \mathcal{G}_{\text{syn}}\right) := \max_{x \in \mathcal{P}_{\text{act}}} \min_{y \in \mathcal{P}_{\text{syn}}} \mathsf{dist}(x, y).$$

The above metric of geometry comparison allows us to measure a degree of proximity for edge geometries that are nonoverlapping yet close to each other. Fig. 8 shows the edge geometry comparison between actual and synthetic networks for uniform rectangular grid partitions of two different resolutions. Note that network geometries in certain regions show a significant deviation when compared with low resolution, while comparing with a higher grid resolution shows a small deviation. This shows that the networks are fairly close to each other.

Case Study: Impact of Photovoltaic Penetration

We now present a representative study where we analyze the impact of photovoltaic (PV) penetration on the system node voltages. We compare the PV penetration in multiple levels of the network (MV primary network or LV secondary network). We consider the following two cases: 1) PV penetration in the LV network, where PV generators are installed on residence rooftops, and 2) PV penetration in the MV network, where a single PV generator is installed at a location in the MV network. In the first case, we randomly identify a group of residences (for example, 50% of all residences) and assign PV generation to them. The penetration level indicates the rating of the PV generation installed on these residences. For the latter case, a single-node PV penetration represents a "solar farm," which is connected to the distribution grid, and the penetration level indicates the PV generation rating as a fraction of the total load.

We perform the comparison on two different synthetic feeders: urban and rural. An urban distribution network is characterized with shorter lines as compared with rural networks where remote nodes are connected by long lines. Fig. 9 compares the impact of

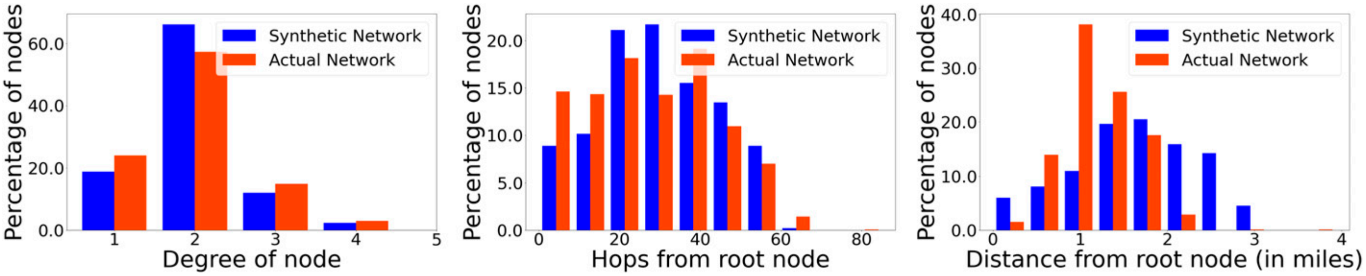

Fig. 7. Plots comparing the degree distribution (*Left*), hop distribution (*Center*), and reach distribution (*Right*) of actual and synthetic distribution networks for the town of Blacksburg in southwest Virginia. The degree and hop distributions are fairly close to each other, which signifies their resemblance. The reach distribution differs between the networks because of the difference in the way each of them is created.

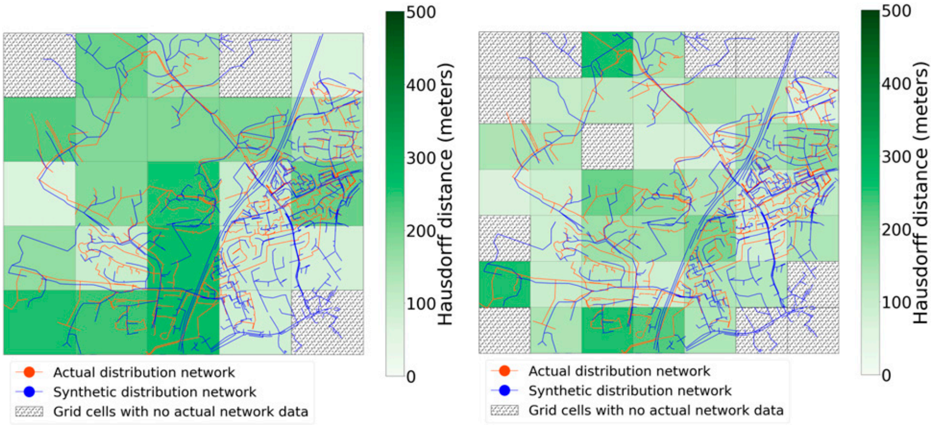
500
400
300
200
100
0
Hausdorff distance (meters)
Actual distribution network
Synthetic distribution network
Grid cells with no actual network data
500
400
300
200
100
0
Hausdorff distance (meters)
Actual distribution network
Synthetic distribution network
Grid cells with no actual network data

Fig. 8. Plots showing Hausdorff distance–based geometry comparison of actual and synthetic networks for the town of Blacksburg in southwest Virginia. The geometry comparison is performed for grid cells with two different resolutions: low resolution of 5 × 5 grid cells (*Left*) and high resolution of 7 × 7 grid cells (*Right*). The color in each grid cell denotes the magnitude of deviation in meters. Grid cells with no available actual network data are shaded with black dots.

LV-level and MV-level PV penetration for two networks. Here, we focus on the percentage of nodes that face overvoltage issues due to different levels of PV penetration. We observe that for either case, the percentage of nodes with overvoltage increases with higher penetration level. Further, we see that LV-level PV generation is less likely to cause overvoltage issues as compared with a single-node MV-level PV integration. Additionally, in the case of rural feeders, the percentage of nodes experiencing severe overvoltage (around 1.05 pu, which is the extreme limit of acceptable overvoltage) is higher as compared with urban feeder networks. Therefore, an optimal placement of PV generators is required for the rural feeders so that they do not suffer from overvoltage issues.

Discussion and Limitations

Although the synthetic power distribution network dataset produced by our framework is comprehensive, it is not without its limitations. In this work, we generate networks with only positive sequence parameters. The ensemble of synthetic networks can be used as a tool for performing planning studies or addressing system-wide policy-level questions. We can also perform short circuit analysis with symmetrical three-phase faults.

However, distribution systems are networks of mixed phase order and mixed network configuration. They are usually three phases in the primary network, and the secondary network consists of mixed single- and three-phase circuits. We have provided a framework in *SI Appendix* to partition the residences into three phases and thereby, create a three-phase network. A complete three-phase network requires the inclusion of zero sequence line parameters and transformer configurations (wye–wye, delta–wye, wye–delta, and delta–delta). Therefore, in their current version, these networks might not be suitable to be used for performing dynamic stability analysis or studying detailed transient responses to power grid contingencies.

Further, shunt compensation is used in the primaries for maintaining voltage level within engineering standards. These are composed of capacitor banks, which elevate voltage level along the network. Hence, they can be optimally placed in the network to avoid severe undervoltage issues at remotely located residences. We can consider critical sections of the network and design necessary shunt compensation to maintain a high degree of reliability of the network. Additionally, the proposed framework creates a network to connect only the residential buildings in a geographic region. In order to connect heavy load centers, networked secondaries with pad-mounted transformers are used in some large

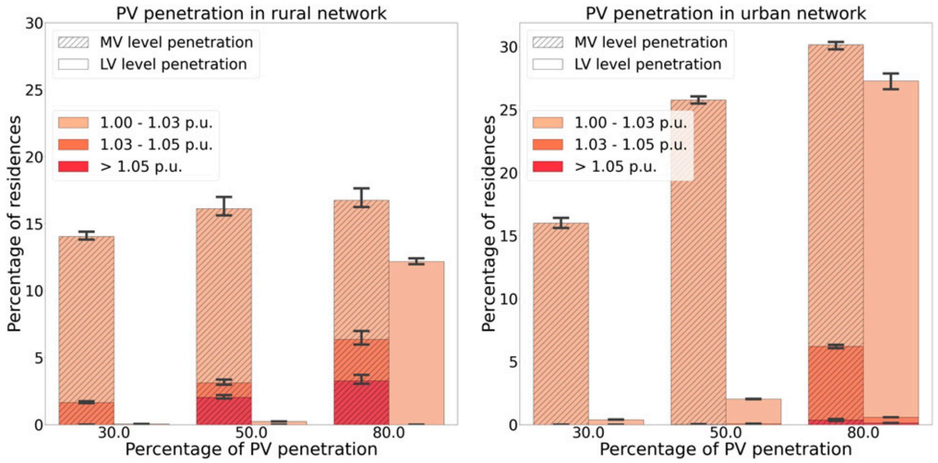

Fig. 9. Plots showing the impact of PV penetration in rural and urban networks. Colors depict the percentages of nodes with various levels of overvoltages. Shaded and nonshaded bars denote MV- and LV-level penetration, respectively. LV-level penetration is less likely to cause severe overvoltages as compared with MV-level penetration. PV penetration in rural networks is more likely to cause overvoltage issues (greater than 1.05 pu).

urban areas. These additions can be made to our existing synthetic networks and would be a direction for future research.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work is partially supported by University of Virginia Strategic Investment Fund Award SIF160, 3Cavaliers Grant 166914-WMS5F-LC00295-30015, NSF Grants EAGER 1927791, CINES 1916805, and ECCS 2039716, and a grant from the C3.ai Digital Transformation Institute.

Author affiliations: [a]Biocomplexity Institute, University of Virginia, Charlottesville, VA 22911; [b]Department of Computer Science, University of Virginia, Charlottesville, VA 22911; [c]Electrical and Computer Engineering Department, Virginia Tech, Blacksburg, VA 24060; and [d]Department of Electrical Engineering, Princeton University, Princeton, NJ 08544

1. G. Byeon, P. Hentenryck, R. Bent, H. Nagarajan, Communication-constrained expansion planning for resilient distribution systems. *INFORMS J. Comput.* **32**, 968–985 (2020).
2. J. Richler, Tell me something I don't know. *Nat. Energy* **5**, 492–492 (2020).
3. I. Onyeji, M. Bazilian, C. Bronk, Cyber security and critical energy infrastructure. *Electr. J.* **27**, 52–60 (2014).
4. D. Quiroga, E. Sauma, D. Pozo, Power system expansion planning under global and local emission mitigation policies. *Appl. Energy* **239**, 1250–1264 (2019).
5. C. D. Brummitt, P. D. H. Hines, I. Dobson, C. Moore, R. M. D'Souza, Transdisciplinary electric power grid science. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 12159–12159 (2013).
6. Z. Wang, A. Scaglione, R. J. Thomas, Generating statistically correct random topologies for testing smart grid communication and control networks. *IEEE Trans. Smart Grid* **1**, 28–39 (2010).
7. S. Soltan, G. Zussman, "Generation of synthetic spatially embedded power grid networks" in *2016 IEEE Power and Energy Society General Meeting (PESGM)* (IEEE, Boston, MA, 2016), pp. 1–5.
8. S. You, S. W. Hadley, M. Shankar, Y. Liu, Co-optimizing generation and transmission expansion with wind power in large-scale power grids–implementation in the US eastern interconnection. *Electr. Power Syst. Res.* **133**, 209–218 (2016).
9. A. Bernstein, D. Bienstock, D. Hay, M. Uzunoglu, G. Zussman, "Power grid vulnerability to geographically correlated failures–analysis and control implications" in *IEEE INFOCOM 2014 IEEE Conference on Computer Communications* (IEEE, Toronto, ON, Canada, 2014), pp. 2634–2642.
10. N. D. Popovich, D. Rajagopal, E. Tasar, A. Phadke, Economic, environmental and grid-resilience benefits of converting diesel trains to battery-electric. *Nat. Energy* **6**, 1017–1025 (2021).
11. C. Gaete-Morales, H. Kramer, W. P. Schill, A. Zerrahn, An open tool for creating battery-electric vehicle time series from empirical data, emobpy. *Sci. Data* **8**, 152 (2021).
12. P. Hines, E. Cotilla-Sanchez, S. Blumsack, Do topological models provide good information about electricity infrastructure vulnerability? *Chaos* **20**, 033122 (2010).
13. X. Fan *et al.*, "Model validation study for Central American regional electrical interconnected system" in *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)* (IEEE, Washington, DC, 2021), pp. 1–5.
14. I. C. Decker *et al.*, "System wide model validation of the Brazilian interconnected power system" in *IEEE PES General Meeting* (IEEE, Minneapolis, MN, 2010), pp. 1–8.
15. S. Biswas, E. Bernabeu, D. Picarelli, "Proactive islanding of the power grid to mitigate high-impact low-frequency events" in *2020 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)* (IEEE, Washington, DC, 2020), pp. 1–5.
16. F. Postigo *et al.*, A review of power distribution test feeders in the United States and the need for synthetic representative networks. *Energies* **10**, 1896 (2017).
17. National Academies of Sciences, Engineering, and Medicine, *Analytic Research Foundations for the Next-Generation Electric Grid* (The National Academies Press, Washington, DC, 2016).
18. K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models" in *2016 IEEE Power and Energy Conference at Illinois (PECI)* (IEEE, Urbana, IL, 2016), pp. 1–6.
19. A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, T. J. Overbye, Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies. *IEEE Trans. Power Syst.* **32**, 1502–1510 (2017).
20. A. Trpovski, D. Recalde, T. Hamacher, "Synthetic distribution grid generation using power system planning: Case study of Singapore" in *2018 53rd International Universities Power Engineering Conference (UPEC)* (IEEE, Glasgow, UK, 2018), pp. 1–6.
21. R. Kadavil, T. M. Hansen, S. Suryanarayanan, "An algorithmic approach for creating diverse stochastic feeder datasets for power systems co-simulations" in *2016 IEEE PES General Meeting (PESGM)* (IEEE, Boston, MA, 2016), pp. 1–5.
22. C. Domingo, T. Roman, A. Sanchez-Miralles, J. P. Gonzalez, A. Martinez, A reference network model for large-scale distribution planning with automatic street map generation. *IEEE Trans. Power Syst.* **26**, 190–197 (2011).
23. L. Gonzalez-Sotres, C. Domingo, A. Sanchez-Miralles, M. Miro, Large-scale MV/LV transformer substation planning considering network costs and flexible area decomposition. *IEEE Trans. Power Deliv.* **28**, 2245–2253 (2013).
24. E. Schweitzer, A. Scaglione, A. Monti, G. A. Pagani, Automated generation algorithm for synthetic medium voltage radial distribution systems. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **7**, 271–284 (2017).
25. R. Meyur *et al.*, "Creating realistic power distribution networks using interdependent road infrastructure" in *IEEE Big Data Conference* (IEEE, Atlanta, GA, 2020) pp. 1226–1235.
26. S. Thorve *et al.*, "Simulating residential energy demand in urban and rural areas" in *2018 Winter Simulation Conference (WSC)* (IEEE, Gothenburg, Sweden, 2018), pp. 548–559.
27. K. Tong, A. S. Nagpure, A. Ramaswami, All urban areas' energy use data across 640 districts in India for the year 2011. *Sci. Data* **8**, 104 (2021).
28. C. Klemenjak, C. Kovatsch, M. Herold, W. Elmenreich, A synthetic energy dataset for non-intrusive load monitoring in households. *Sci. Data* **7**, 108 (2020).
29. G. Cimini *et al.*, The statistical physics of real-world networks. *Nat. Rev. Phys* **1**, 58–71 (2019).
30. B. Hartmann, V. Sugár, Searching for small-world and scale-free behaviour in long-term historical data of a real-world power grid. *Sci. Rep.* **11**, 6575 (2021).
31. V. Krishnan *et al.*, Validation of synthetic U.S. electric power distribution system data sets. *IEEE Trans. Smart Grid* **11**, 4477–4489 (2020).
32. R. Atat, M. Ismail, M. F. Shaaban, E. Serpedin, T. Overbye, Stochastic geometry-based model for dynamic allocation of metering equipment in spatio-temporal expanding power grids. *IEEE Trans. Smart Grid* **11**, 1–12 (2019).
33. H. Li *et al.*, Building highly detailed synthetic electric grid data sets for combined transmission and distribution systems. *IEEE Open Access J. Power Energy* **7**, 478–488 (2020).
34. C. Mateo *et al.*, Building large-scale U.S. synthetic electric distribution system models. *IEEE Trans. Smart Grid* **11**, 5301–5313 (2020).
35. S. S. Saha, E. Schweitzer, A. Scaglione, N. Johnson, "A framework for generating synthetic distribution feeders using openstreetmap" in *2019 North American Power Symposium (NAPS)* (IEEE, Wichita, KS, 2019), pp. 1–6.
36. A. Bidel, T. Schelo, T. Hamacher, "Synthetic distribution grid generation based on high resolution spatial data" in *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC/ICPS Europe)* (IEEE, Bari, Italy, 2021), pp. 1–6.
37. M. Liang, Y. Meng, J. Wang, D. Lubkeman, N. Lu, Feedergan: Synthetic feeder generation via deep graph adversarial nets. *IEEE Trans. Smart Grid* **12**, 1163–1173 (2021).
38. R. Subbiah, A. Pal, E. K. Nordberg, A. Marathe, M. V. Marathe, Energy demand model for residential sector: A first principles approach. *IEEE Trans. Sustain. Energy* **8**, 1215–1224 (2017).
39. US Department of Homeland Security, *Electric substations* (2019). https://hifld-geoplatform.opendata.arcgis.com/datasets/electric-substations. Accessed 20 April 2021.
40. M. Tantardini, F. Ieva, L. Tajoli, C. Piccardi, Comparing methods for comparing networks. *Sci. Rep.* **9**, 17557 (2019).
41. Y. Bai *et al.*, Graph edit distance computation via graph neural networks. arXiv [Preprint] (2018). https://arxiv.org/abs/1808.05689v3 (Accessed 26 September 2022).
42. S. Ok, "A graph similarity for deep learning" in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin, Eds. (Curran Associates, Inc., Vancouver, BC, Canada, 2020), vol. 33, pp. 1–12.
43. H. Xu, An algorithm for comparing similarity between two trees. arXiv [Preprint] (2015). https://arxiv.org/abs/1508.03381 (Accessed 26 September 2022).
44. B. Paaßen, Revisiting the tree edit distance and its backtracing: A tutorial. arXiv [Preprint] (2022). https://arxiv.org/abs/1805.06869 (Accessed 26 September 2022).
45. P. Riba, A. Fischer, J. Lladós, A. Fornes, Learning graph edit distance by graph neural networks. *Pattern Recognit.* **120**, 108132, (2021).
46. I. Morer, A. Cardillo, A. Díaz-Guilera, L. Prignano, S. Lozano, Comparing spatial networks: A one-size-fits-all efficiency-driven approach. *Phys. Rev. E* **101**, 042301 (2020).
47. OpenStreetMap, (2021). https://www.openstreetmap.org/. Accessed 20 April 2021.
48. W. H. Kersting, *Distribution System Modeling and Analysis* (CRC Press, Abingdon, United Kingdom, 2012).
49. A. K. Dey, Y. R. Gel, H. V. Poor, What network motifs tell us about resilience and reliability of complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 19368–19373 (2019).