Generation of Synthetic Spatially Embedded Power Grid Networks

Saleh Soltan and Gil Zussman

***Abstract*—The development of algorithms for enhancing the resilience and efficiency of the power grid requires performance evaluation with real topologies of power transmission networks. However, due to security reasons, *such topologies and particularly the locations of the substations and the lines are usually not publicly available*. Therefore, we study the structural properties of the North American grids and present an algorithm for generating synthetic spatially embedded networks with similar properties to a given grid. The algorithm uses the Gaussian Mixture Model (GMM) for density estimation of the node positions and generates a set of nodes with similar spatial distribution to the nodes in a given network. Then, it uses two procedures, which are inspired by the historical evolution of the grids, to connect the nodes. The algorithm has several tunable parameters that allow generating grids similar to any given grid. Particularly, we apply it to the Western Interconnection (WI) and to grids that operate under the SERC Reliability Corporation (SERC) and the Florida Reliability Coordinating Council (FRCC), and show that it generates grids with similar structural and spatial properties to these grids. To the best of our knowledge, this is the first attempt to *consider the spatial distribution of the nodes and lines* and its importance in generating synthetic power grids.**

***Index Terms*—Power Grids, Structural Properties, Synthetic Networks, Spatial Networks, Data Mining.**

I. Introduction

The design of algorithms and methods for enhancing the power grid (namely, making it smarter) drew tremendous attention over the past decade [1], [2]. These efforts focused on challenges stemming from renewable generation interconnection [3], Phasor Measurement Units (PMUs) placement [4], [5], transmission expansion planning [6], and vulnerability analysis [7], [8], [9], [10]. The development of algorithms for coping with these challenges *requires performance evaluation with real grid topologies*. However, in order to avoid exposing vulnerabilities, *the topologies of the power transmission networks and particularly the locations of the substations and the lines are usually not publicly available* or are hard to obtain.

There are only very few and limited test cases and real-world power grid datasets that are publicly and freely available. These include the IEEE test cases [11], the National Grid UK [12], the Polish grid [13], and an approximate model of the European interconnected system [14]. To the best of our knowledge, among these, National Grid UK is the only publicly available dataset with geographical locations. Even if the data was available, it would be unwise to publish vulnerability results which are based on real topologies, due to the enormous cost of grid enhancements. On the other hand, it was recently shown that simple random graph models cannot be used to generate grids with appropriate structural and spatial characteristics [15] (for more details, see Section II). Therefore, in this paper *we design an algorithm for generating synthetic networks with similar structural and spatial properties to real power grids*. Such synthetic networks can be used for evaluation of various methods and techniques.

S. Soltan and G. Zussman are with the Department of Electrical Engineering, Columbia University, New York, NY, 10027.
E-mail: {saleh,gil}@ee.columbia.edu

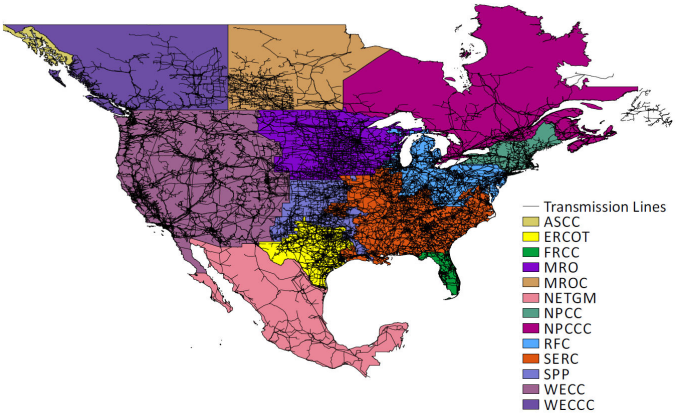


Fig. 1: The North American Electric Reliability Corporation (NERC) regional entities and the National Electricity Transmission Grid of Mexico (NETGM). Different reliability corporations/councils are marked with different colors.

To demonstrate the algorithm design and to evaluate its performance, we focus on the transmission networks of the North American and Mexican power grids (see NERC and NETGM in Fig. 1) using data that we obtained from the Platts Geographic Information System (GIS) [16]. We consider one of the two major interconnections – the Western Interconnection (WI) (see Fig. 2) which includes the Western Electricity Coordinating Council in the United States (WECC) and Canada (WECCC) (see Fig. 1 for their coverage areas). Moreover, we consider two regional entities that operate under the Eastern Interconnection (EI) which is the other major interconnection – the SERC Reliability Corporation (SERC), which is as large as the WI, and the Florida Reliability Coordinating Council (FRCC), which is much smaller than the WI. To the best of our knowledge, *this is the first time that the entire dataset of the North American and Mexican grids as well as those of SERC and FRCC are processed and analyzed*[1].

For the entire North American and Mexican grid as well as for WI, SERC, and FRCC, we consider four metrics that capture the networks' structural properties: average path length,

[1]Partial analysis of the WI dataset has been conducted before – see Section II.

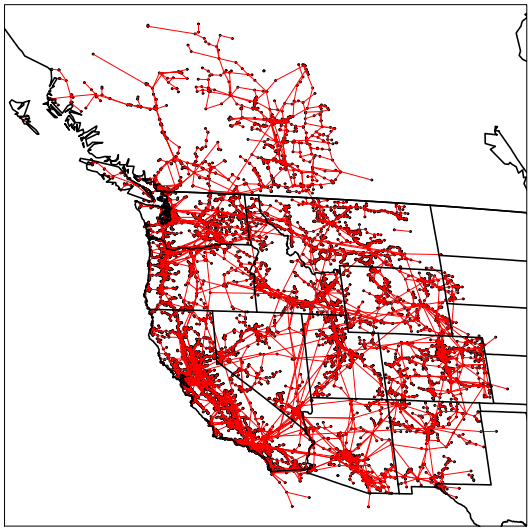


Fig. 2: The Western Interconnection (WI) power grid with 14,302 substations (nodes) and 18,769 lines (edges).

clustering coefficient, degree distribution of the nodes, and the length distribution of the lines. The first three metrics are very common [15], [17], [18], [19], [20], [21], [22]. However, to the best of our knowledge, *the length distributions of the lines have not been thoroughly studied before*. These distributions are particularly important, since the physical properties of a line (e.g., admittance and type) are directly correlated with its length [23], and hence, the distributions directly impact the performance of various algorithms.

Motivated by the results of the structural properties' analysis, we present the *Geographical Network Learner and Generator (GNLG) Algorithm for generating a network with similar properties to a given grid*. First, using Gaussian Mixture Model (GMM), the algorithm estimates the density of the node positions and uses the obtained parameters to generate a set of nodes with a similar spatial distribution to these nodes (the algorithm uses the Bayesian Information Criterion (BIC) to find the best number of clusters for the GMM). Then, the GNLG Algorithm uses two procedures, which are inspired by the historical evolution of power grids, to connect the generated nodes. Particularly, since the two main design considerations of the grid are connectivity and robustness, the algorithm obtains a spanning tree of the nodes to provide connectivity and then adds more edges to the network graph to increase its robustness. The addition of edges is tuned to create a synthetic network with properties that are similar to those of a given network.

To evaluate the performance of the GNLG Algorithm, we use it to generate networks similar to the WI, SERC, and FRCC. We show that by adapting a number of tunable parameters, the GNLG Algorithm can generate synthetic networks with similar structural and spatial properties to these power grid networks. Overall, we believe that by adapting the algorithm's tunable parameters, it is possible to generate synthetic networks similar to any given power grid network.

This paper is organized as follows. Section II reviews related work. Section III describes the dataset and the metrics, and presents the metrics for the different grids. Section IV describes the GNLG Algorithm and Section V numerically evaluates its performance. We conclude and discuss future research directions in Section VI.

II. Related Work

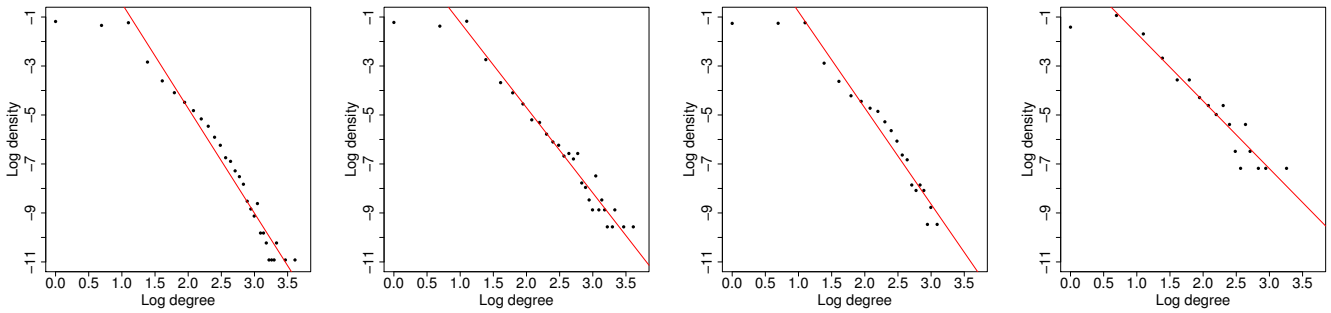
The structural properties of various power grids (e.g., in North America, some European countries, and Iran) were studied in [17], [21], [24], [25], [26], [27]. Most of these studies considered one or two properties (e.g., average degree, degree distribution, average path length, and clustering coefficient) and computed it in a given power grid. In some cases (e.g., [15], [17], [18], [19], [20], [21], [22]) a certain class of graphs was suggested as a good representative of a power grid network, based on one or two structural properties. For example, Watts and Strogatz [17] suggested the small-world graph as a good representative, based on the shortest path lengths between nodes and the clustering coefficient of the nodes. Barabási and Albert [18] showed that scale-free graphs are better representatives based on the degree distribution. However, by comparing the WI with these models, Cotilla-Sanchez, et al. [15] showed that none of them can represent the WI properly.

More detailed models that are specifically tailored to the power grid characteristics were proposed in [28], [29] but they did not consider the nodes' *spatial distribution* and the length distribution of the lines. The spatial distribution of the nodes is correlated with the length of the lines, and as mentioned above, it is important to consider line lengths when designing a method for synthetic power grid generation. While there are several models for generating spatial networks [30], [31], [32], most of them were not designed to generate networks with properties similar to power grid networks. To the best of our knowledge, this paper is the first to consider the spatial distribution of the nodes in power grids and its importance in generating synthetic networks with similar structural properties.

III. Preliminaries and Structural Properties

In this section, we study the structural properties of the entire North American and Mexican grid (denoted by NA&M) as well as of the WI, SERC, and FRCC grids. We obtained the data from the Platts GIS [16] and conducted longitude-latitude to planar (x, y) coordinate transformation, using the great-circle distance method. Since the files containing substations and files containing lines are not always consistent, we extracted the coordinates of the substations from the end point coordinates of the lines. We then used the geographical coordinates of the substations and the lines to construct the graphs with nodes and edges that represent substations and lines, respectively. We used the map of reliability coorporations/councils boundaries to divide the graph into regional entities (as in Fig. 1). To the best of our knowledge, beside [7], [8] where an approximation of the WI graph was extracted from the Platts GIS dataset for simulations, it is the first time that this dataset is processed and analyzed.

In addition to the number of the nodes and edges, we use four metrics for classifying the structural properties of these



(a) NA&M (b) WI (c) SERC (d) FRCC

Fig. 3: The degree distribution of the nodes in the NA&M, WI, SERC, and FRCC grids (in log-log scale). Linear regression lines with slopes $\zeta = -4.28$, $\zeta = -3.48$, $\zeta = -3.93$, and $\zeta = -2.76$, respectively, are fitted to the tail distribution of the degrees.

TABLE I: Summary of the structural properties of the NA&M, WI, SERC, and FRCC grids.

Network	NA&M	WI	SERC	FRCC
Number of Nodes (n)	55,231	14,302	12,946	1,312
Number of Edges (m)	70,088	18,769	16,658	1,780
Average Path Length (L)	26.66	17.33	19.71	11.68
Clustering Coefficient (C)	0.049	0.049	0.049	0.075
Degree Distribution (ζ)	-4.28	-3.48	-3.93	-2.76

networks: *average path length, clustering coefficient, degree distribution of the nodes, and length distribution of the lines.* Table I includes these metrics for the NA&M, WI, SERC, and FRCC grids.

Notation. We denote the WI, SERC, and FRCC power grid transmission networks by graphs G_{WI}, G_{SERC}, and G_{FRCC}, respectively. For each network, n and m denote the number of the nodes and edges. d_i denotes the degree of node i and $\mathbf{p}_i \in \mathbb{R}^2$ denotes its position. We define ρ as the average Euclidean distance of a node from its N nearest neighbors. We use the prime symbol ($'$) to denote the values for a generated network (e.g., G'_{WI} denotes the generated network). All the logarithms in this paper are natural logarithms. All the geographical distances in this paper are Euclidean distances (i.e., $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ is the distance between nodes i and j).

A. Average path length

The average path length, denoted by L, is one of the common metrics used for classifying graphs. It is defined as the number of edges in the shortest path between two nodes, averaged over all pairs of vertices:

$$L = \frac{1}{n(n-1)} \sum_{\substack{i \neq j \\ i,j \in V}} \text{dist}(i,j),$$

where $\text{dist}(i,j)$ is the number of edges in the shortest path between nodes i, j. As can be seen in Table I, the average path length in all the four networks is in $O(\log(n))$ which is very small and suggests that these networks have the small-world property.

B. Clustering coefficient

An important metric is the clustering coefficient, denoted by C and defined as follows. For each node i, with degree d_i at most $d_i(d_i-1)/2$ edges can exist between its neighbors $N(i)$. Let C_i denotes the fraction of these allowable edges that actually exist:

$$C_i = \frac{|\{\{r,s\}|r,s \in N(i), \{r,s\} \in E\}|}{d_i(d_i-1)/2}.$$

Then, averaging C_i over all the nodes: $C = \sum_{i \in V} C_i/n$. As can be seen in Table I, the clustering coefficient for all the four networks is very small.

C. Degree distribution of the nodes

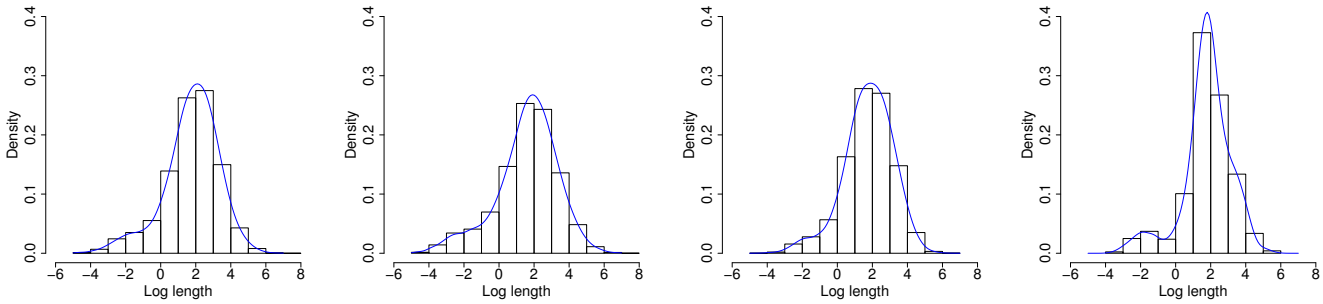
The degree distribution of the nodes is another important metric for classifying graphs (e.g., scale-free networks). Fig. 3 shows the degree distribution of the nodes in the NA&M, WI, SERC, and FRCC grids in log-log scale. The degree one nodes in these networks usually correspond to power plants or small towns. These figures may suggest that the tail of the degree distribution follows a power-law distribution in all the three networks. However, following [33] and since these networks are finite, we do not have enough statistical evidence to support the power-law hypothesis. Therefore, we only use the slope (ζ) of the fitted linear regression line to the tail distribution for comparison purposes.

In Section V, we use the Kolmogrov-Smirnov (KS) statistic [34] to compare the degree distribution of the nodes in a given network and a generated network. If $P(x)$ and $Q(x)$ are two Cumulative Distribution Functions (CDFs), the KS statistic between these two is defined as follows:

$$D_{KS} = \max_x |P(x) - Q(x)|.$$

D. Length distribution of the lines

As mentioned above, the length distribution of the lines is one of the important parameters that needs to be sustained in synthetic power grid generation. Fig. 4 shows the length distribution of the lines in the NA&M, WI, SERC, and FRCC grids. The length distribution of the lines in the NA&M grid



(a) NA&M (b) WI (c) SERC (d) FRCC

Fig. 4: The distributions of the actual line lengths (in *km*) in the NA&M, WI, SERC, and FRCC grids (the lengths' statistics appear in Table II). Nonparametric distribution fits to the log length distributions are shown in blue.

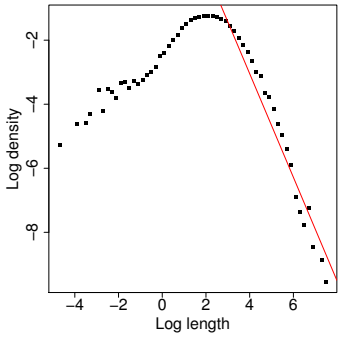


Fig. 5: The distribution of the actual line lengths (in *km*) in the NA&M grid in log-log scale. A linear regression line with slope -1.61 is fitted to the tail distribution of the lengths.

TABLE II: Statistics of the actual line lengths in the NA&M, WI, SERC, and FRCC grids and of the corresponding straight lines (Euclidean distances) between substations in those grids (in km). The statistics of the straight lines are shown in the grey cells.

Network	NA&M	WI	SERC	FRCC
Mean	15.46	16.63	13.29	12.82
	14.30	15.78	11.39	9.95
Standard Deviation	32.55	43.91	22.29	20.14
	30.68	40.78	17.90	15.6
Maximum	1,714.82	1,714.82	795.44	282.74
	1,380.35	1,380.35	409.92	226.25

in log-log scale is shown in Fig. 5.[2] The lengths' statistics appear in Table II.

The line lengths in Figs. 4 and 5 are the *actual lengths* of the power lines (these lines are not necessarily straight lines between two substation). To enable the comparison between the length distributions of the lines in the real and generated networks, in Section V we use the *point-to-point Euclidean distances* to represent the line lengths in the real and the generated networks. Table II includes the statistics regarding both the actual line lengths and the lengths of the straight lines between the substations, in order to demonstrate the differences between the metrics.

In Section V, we use Kullback-Leibler (KL) divergence to measure the similarity between the length distribution of the lines in a given network and a generated network. The KL-divergence is a non-symmetric measure of the difference between two probability distribution functions p and q. Specifically, the KL-divergence of q from p, denoted $D_{KL}(p\|q)$, is a measure of the information lost when q is used to approximate p:

$$D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx.$$

To estimate the KL-divergence between distributions, we use the `FNN` library in `R` which utilizes the method introduced in [35] for estimating the KL-divergence between two distributions using their samples.

[2]As can be seen in Figs. 4 and 5, there are some very short lines ($\approx 30m$) in the considered networks. We checked the dataset to verify the credibility of these lines and did not find any issues (these lines are categorized as *below 230kV* lines).

Algorithm 1: Geographical Network Learner and Generator (GNLG)

Input: G, $\{\mathbf{p}_i\}_{i=1}^{n}$, and parameters $\kappa, \alpha, \beta, \gamma > 0$ and $N \in \mathbb{N}$.
1: Generate a set of nodes with similar spatial distribution to the nodes in G using the SDNG Procedure (Subsection IV-A).
2: Connect the generated nodes using the TWST Procedure (Subsection IV-B).
3: Add more edges to the generated graph using the Reinforcement Procedure (Subsection IV-B).
4: **return** the generated graph G'.

IV. Generating a Synthetic Network

In this section, we introduce the Geographical Network Learner and Generator (GNLG) Algorithm (Algorithm 1) for generating a synthetic network similar to a given network. The algorithm uses the Gaussian Mixture Model (GMM) for density estimation of the node positions and generates a set of nodes with similar spatial distribution to the nodes in a given network (the SDNG Procedure described in Subsection IV-A). Then, it connects the nodes using two procedures whose design principles are inspired by historical evolution of the grids (the TWST and Reinforcement procedures described in Subsection IV-B). The GNLG Algorithm can be applied to any network, where the important part is tuning the parameters to a given network. In the following subsections, we describe the building blocks of the GNLG Algorithm and use the WI to demonstrate the algorithm design and operation. Then, in Section V, we evaluate the algorithm using the WI, SERC, and FRCC grids.

Procedure 1: Spatially Distributed Nodes Generator (SDNG)

Input: G, $\{\mathbf{p}_i\}_{i=1}^n$.
1: Fit a GMM model to $\{\mathbf{p}_i\}_{i=1}^n$ to cluster them into c clusters that maximizes the BIC.
2: For all $i = 1, \ldots, n$ sample z_i from the categorical probability distribution π obtained from GMM.
3: For all i sample $\mathbf{p}'_i$ from the probability distribution $\mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$ obtained from GMM.
4: **return** $\{\mathbf{p}'_i\}_{i=1}^n$.

A. Node positions

We now introduce the Spatially Distributed Nodes Generator (SDNG) Procedure (Procedure 1) for generating a set of nodes with similar spatial distribution to the nodes in a given network. The node positions are correlated with the population and geographical properties (e.g., Fig. 2). Thus, the nodes can be clustered into groups based on their geographical proximity. Mixture models and in particular Gaussian Mixture Models (GMM) are commonly used for clustering and density estimation [36]. Hence, the SDNG Procedure uses the GMM for clustering the positions and uses BIC to find the best number of clusters (c). It obtains the mean and covariance matrix (μ_j, Σ_j) of the points in clusters $j = 1, \ldots, c$ along with the categorical probability of the clusters $\pi = (\pi_1, \ldots, \pi_c)$. Then, it uses these parameters to generate n nodes with similar spatial distribution as the nodes in a given network.

For implementing the SDNG Procedure, we used the `mclust` library in `R` [37] to apply GMM to our dataset. This library uses the Expectation Maximization (EM) algorithm to fit a GMM and provides the Bayesian Information Criterion (BIC) for the selected number of clusters. Clustering the nodes in the WI into 55 clusters results in the maximum BIC. Hence, the SDNG Algorithm clusters WI into $c = 55$ clusters. As can be seen in Fig. 7, the distribution of the generated nodes appears very similar to the distribution of the nodes in the WI.

Notice that for a given network, step 1 in the Procedure should be executed only once. Then, having the fitted GMM parameters, the procedure can be used to generate several instances of nodes with similar spatial distribution to the nodes in the given network. Hence, once the parameters are available, synthetic grids can be generated with no need to access the real grid data.

B. Connections between the nodes

We introduce two procedures (steps 2 and 3 in the GNLG Algorithm) for connecting the generated nodes. Their design is inspired by the historical evolution of power grids. The two main design consideration of the grid are (i) connectivity and (ii) robustness. Therefore, we first present the Tunable Weight Spanning Tree (TWST) Procedure for finding a spanning tree and to ensure connectivity. We then describe the Reinforcement Procedure for adding more edges and ensuring the network robustness as well as for tuning the structural properties of the synthetic network to resemble those of a given network.

1) Connectivity: In order for the power grid to operate, the substations (nodes) should be connected. Due to construction costs, in the real world new substations are usually connected to the nearest substation in the existing grid. Since the power grids have evolved gradually and locally, they do not necessarily contain the Minimum weight Spanning Tree (MST) of the nodes in the plane (the weight of a spanning tree $T = (V_T, E_T)$ is the sum of the edge lengths in T: $W_T = \sum_{\{i,j\} \in E_T} \|\mathbf{p}'_i - \mathbf{p}'_j\|$). Hence, we do not focus on finding the MST. Instead, we present the TWST Procedure (Procedure 2), which imitates the the gradual grid evolution. It is a low complexity procedure for finding a spanning tree with a tunable weight.

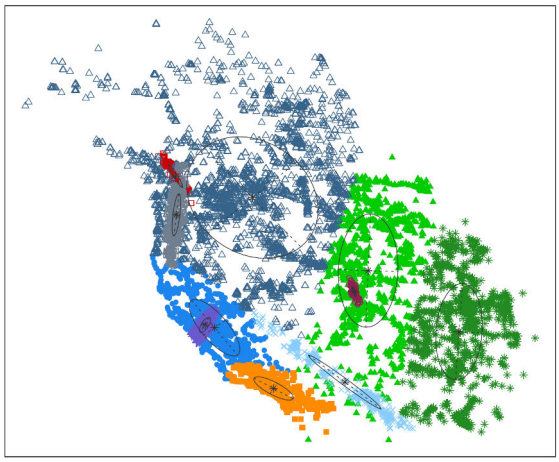


Fig. 6: An example of clustering the nodes in the WI into 10 clusters using GMM.

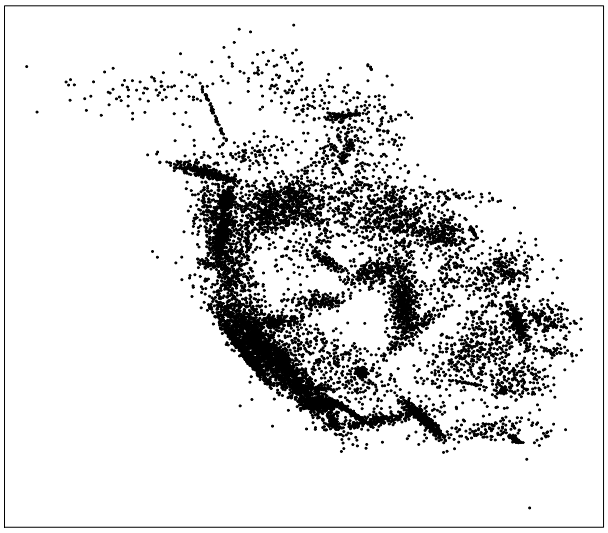
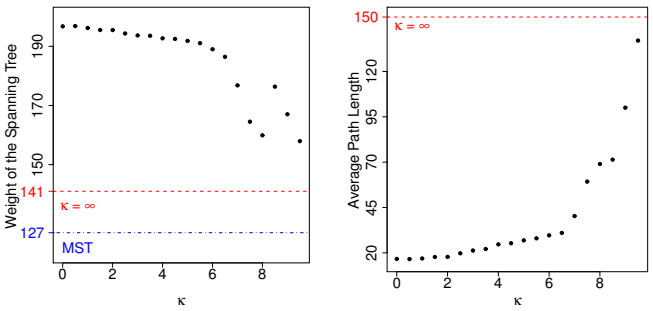


Fig. 7: A set of nodes, that were generated using the SDNG Procedure, with a similar spatial distribution to the nodes in the WI.

The procedure uses the average node location, denoted by: $\bar{\mathbf{p}}' = \sum_i \mathbf{p}'_i / n$. It first orders the nodes in n rounds (see step 2) to obtain a permutation of indices $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. At round i, it samples a node j from the nodes that were not already sampled with probability proportional to $\|\mathbf{p}'_j - \bar{\mathbf{p}}'\|^{-\kappa}$, where κ is a parameter. It then sets $\sigma(i) \leftarrow j$. In step 5 it connects each node $\sigma(i)$ to its nearest neighbor $\sigma(j^*)$ such that $j^* < i$.

The procedure results in a tree whose weight highly depends on the ordering of the nodes, and thereby on κ. Moreover, there



(a) (b)

Fig. 8: (a) The weight of the spanning tree (in $10^3 km$) obtained by the TWST Procedure on the nodes shown in Fig. 7 vs. κ. Each point is the average over 10 generated trees. The blue dash-dot line shows the weight of the MST and the red dashed line shows the weight of the obtained spanning tree for $\kappa = \infty$. (b) The average path length in the spanning tree obtained by the TWST Procedure on the nodes shown in Fig. 7 vs. κ. Each point is the average over 10 generated trees. The average path length in a specific MST (an MST may not be unique) is 520. The red dashed line shows the average path length in the obtained spanning tree for $\kappa = \infty$.

Procedure 2: Tunable Weight Spanning Tree (TWST)

Input: $n, \{\mathbf{p}'_i\}_{i=1}^n$, and parameter κ.
1: $A = \{1, \ldots, n\}$, σ is an empty array of size n.
2: **for** $i = 1 \ldots, n$ **do**
3: Sample a node from A such that the probability of sampling node j is $\frac{\|\mathbf{p}'_j - \bar{\mathbf{p}}'\|^{-\kappa}}{\sum_{a \in A} \|\mathbf{p}'_a - \bar{\mathbf{p}}'\|^{-\kappa}}$.
4: $\sigma(i) \leftarrow j$, $A \leftarrow A \backslash \{j\}$.
5: **for** $i = 2, \ldots, n$ **do**
6: Connect node $\sigma(i)$ to node $\sigma(j^*)$ such that $j^* = \text{argmin}_{j<i} \|\mathbf{p}'_{\sigma(i)} - \mathbf{p}'_{\sigma(j)}\|$.

is a specific ordering of the nodes such that the procedure provides the MST (the nodes should be ordered according to their appearance in Prim's Algorithm [38] for finding the MST). Specifically, κ determines the difference between the obtained spanning tree and the MST. Fig. 8(a) shows the relationship between the weight of the obtained tree and κ. When $\kappa = 0$, the nodes are ordered randomly and the weight of the obtained spanning tree significantly differs from the MST's weight. However, As κ increases the weight of the spanning tree decreases. When κ is very large, the nodes are ordered based on their distance from the average location, and therefore, the obtained spanning tree's weight is close to the MST's (shown by the blue dash-dot line).

Fig. 8(b) shows the relationship between κ and the average path length in the obtained tree. As κ increases, the average path length increases. For large κ, this increase is more significant. Moreover, the average path length in an MST (520) is significantly larger than in trees obtained by the TWST Procedure. Overall, Figs. 8(a),(b) suggest that selecting a relatively small κ results in a spanning tree with smaller average path length than the MST and with a reasonable total weight. We show in Section V that for generating a network similar to the WI, $\kappa = 2.5$ is a relatively good choice.

2) Robustness: We present the Reinforcement Procedure whose objective is to increase the robustness of the generated

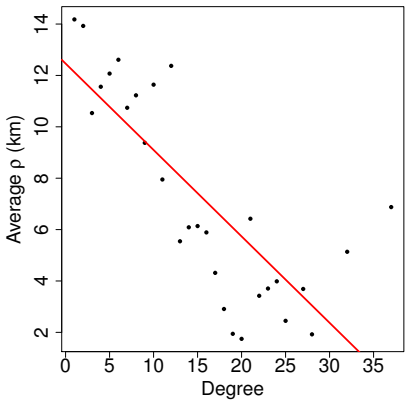


Fig. 9: The relationship between the degree of a node and its average ρ with $N = 10$, for the nodes in the WI (the red line is the linear regression fit to the data points).

Procedure 3: Reinforcement

Input: $n, m, \{\mathbf{p}'_i\}_{i=1}^n$, and parameters $\alpha, \beta, \gamma, \eta > 0$, $N \in \mathbb{N}$.
1: For each node i, compute ρ_i (the average distance of node i from its N nearest neighbors).
2: **for** *count* $= 1$ to $m - n + 1$ **do**
3: **if** *large network*: From all nodes with degree less than 3, sample node i with probability $\propto \rho_i'^{-\alpha}$.
4: **if** *small network*: Sample node i with probability $\propto d_i'^{-\eta} \rho_i'^{-\alpha}$.
5: Connect node i to node j sampled from all other nodes with probability $\propto \|\mathbf{p}'_i - \mathbf{p}'_j\|^{-\beta} d_j'^{\gamma}$.

network and adjust its properties (e.g., L and C) to resemble those of a given network. The procedure is based on three observations: (i) the degree distributions of power grids are very similar to those of scale-free networks, but grids have less degree 1 and 2 nodes and do not have very high degree nodes (e.g., Fig. 3), (ii) it is inefficient and unsafe for the power grids to include very long lines (e.g., Figs. 4 and 5), and (iii) nodes in denser areas are more likely to have higher degrees. The last observation is demonstrated by Fig. 9 where as the degree increases, the ρ decreases[3] (i.e., the density around a node increases).

The Reinforcement Procedure aims to create a network whose properties are similar to those observed above. Hence, it repeats the following steps $m-n+1$ times: (1) selects a low degree node in a dense area (observations (i) and (iii)), and (2) connects it to a high degree node (as in the preferential attachment model [18]) which is also nearby (distance was not considered in [18]) (observations (i) and (ii)).

To select a low degree node in a dense area, the Reinforcement Procedure samples a node i with probability $\propto d_i^{-\eta} \rho_i^{-\alpha}$. However, as can be seen in Fig. 3, the distribution of the degree 1 and 2 nodes is almost equal in the WI and SERC grids. Hence, for large networks, the procedure only considers degree 1 and 2 nodes and select a node among them with probability $\propto \rho_i^{-\alpha}$. α and η are the tunable parameters.

To connect the node sampled in the previous step to a high degree but nearby node, in the second step, the Reinforcement Procedure connects node i to node j sampled from all other nodes with probability $\propto \|\mathbf{p}'_i - \mathbf{p}'_j\|^{-\beta} d_j'^{\gamma}$. This implies that

[3]Recall that ρ is the average Euclidean distance of a node from its N nearest neighbors.

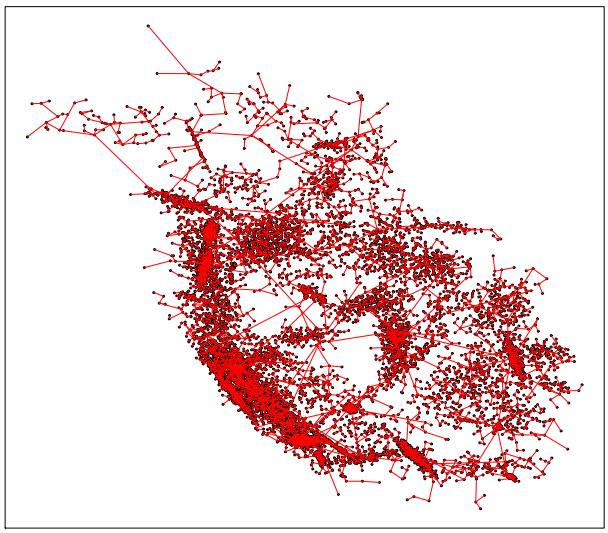


Fig. 10: A network with 14,302 nodes and 18,769 edges generated based on the WI grid using the GNLG Algorithm with $\kappa = 2.5$, $\alpha = 1, \beta = 3.2, \gamma = 2.5$, and $N = 10$.

node i preferentially connects to a high-degree node, unless the high-degree node is too far in which case it is desirable to connect to a low-degree but nearby node. This is very similar to the model introduced in [31], [32]. However, here we only use these probabilities for sampling and do not use them for connecting every pair of nodes.

We note that β determines the length distribution of the new lines and γ determines the likelihood of the existence of high degree nodes. If β is large compared to γ, then new edges connect nearby nodes, thereby resulting in a large clustering coefficient and a large average path length. If γ is large compared to β, then new edges connect nodes to high degree nodes regardless of their distance, thereby resulting in very high degree nodes and long edges. Hence, there should be a balance between the β and γ values. We show in Section V that for generating a network similar to the WI, $\beta = 3.2$ and $\gamma = 2.5$ are relatively good choices.

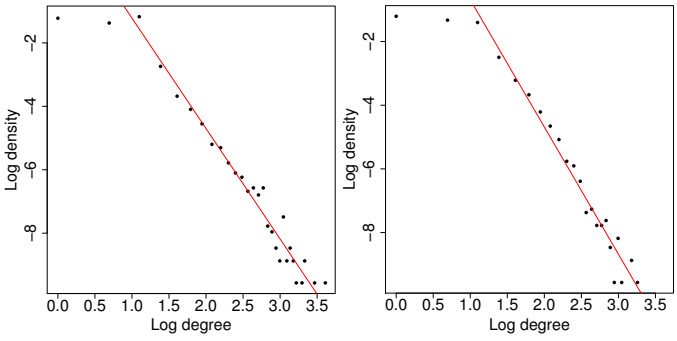
V. Evaluation

In this section, we use the GNLG Algorithm to generate networks similar to the WI, SERC, and FRCC grids. We evaluate the structural properties of the obtained networks and show that they have similar properties to the real networks.

A. WI

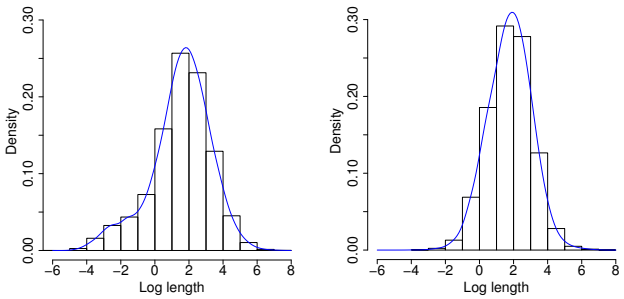
As mentioned in Section IV-B, the parameters $\kappa, \alpha, \beta, \gamma, N$ can be used to tune the structural properties of the obtained network. Therefore, we conducted several numerical experiments in which the parameters were adapted and the structural properties were evaluated. We observed empirically that the following parameters values provide a network with similar properties to the WI: $\kappa = 2.5$, $\alpha = 1, \beta = 3.2, \gamma = 2.5$, and $N = 10$. Moreover, as mentioned in Section IV-A, BIC was used to determine the number of clusters ($c = 55$).

The nodes generated by the SDNG Procedure were shown in Fig. 7. The network obtained by the GNLG Algorithm appears in Fig. 10 and visually resembles the WI. To study the



(a) G_{WI} (b) G'_{WI}

Fig. 11: The degree distribution of the nodes in G_{WI} and G'_{WI} (in log-log scale). Linear regression lines with slopes $\zeta = -3.48$ and $\zeta = -3.99$ are fitted to the distributions of the nodes with degree greater that 2 in G_{WI} and G'_{WI}, respectively. The KS statistic between the degree distributions is 0.047.



(a) G_{WI} (b) G'_{WI}

Fig. 12: The length (in *km*) distribution of the point-to-point lines in G_{WI} and G'_{WI} and nonparametric distribution fit (shown in blue). The KL-divergence between the length distributions in G_{WI} and G'_{WI} is 0.14.

structural similarity between the obtained network G'_{WI} and the G_{WI}, we evaluated G'_{WI} based on the metrics described in Section III. The clustering coefficient and the average path length of G'_{WI} are $C' = 0.052$ and $L' = 17.40$, respectively, and are very close to those of G_{WI} ($C = 0.049$ and $L = 17.33$).

Fig. 11 shows the degree distribution of the nodes in G'_{WI}. As can be seen, the slope of the fitted regression line to the tail of the distribution is -3.99 which is similar to that of G_{WI} (-3.4). Moreover, the KS statistic between the cumulative degree distributions in G_{WI} and G'_{WI} is 0.047, indicating the similarity between the degree distributions. Fig. 12 shows the length distribution of the lines in G'_{WI}. Since the GNLG Algorithm uses straight lines to connect the nodes, we compare the length distribution of the lines in G'_{WI} with the length distribution of the straight point-to-point lines in G_{WI}. The KL-divergence between the length distributions of the lines in G_{WI} and G'_{WI} is $D_{KL} = 0.14$, indicating that distributions are similar.

Table III summarizes the structural properties of the G_{WI} and five instances generated by the GNLG Algorithm. The results indicate that the Algorithm can generate synthetic networks with similar structural properties to the WI grid.

TABLE III: Comparison between the structural properties of WI (G_{WI}) and the Generated WI (G'_{WI}). Five instances of G'_{WI} are shown to illustrate that the metric values are similar. All networks have 14,302 nodes and 18,769 edges.

Networks	L	C	ζ	D_{KS}	D_{KL}
G_{WI}	17.33	0.049	-3.48	0	0
G'_{WI}	17.40	0.052	-3.99	0.047	0.14
$G'_{WI}(2)$	18.36	0.052	-3.65	0.050	0.15
$G'_{WI}(3)$	18.36	0.049	-3.99	0.047	0.12
$G'_{WI}(4)$	19.06	0.052	-3.61	0.049	0.14
$G'_{WI}(5)$	17.79	0.051	-3.50	0.049	0.14

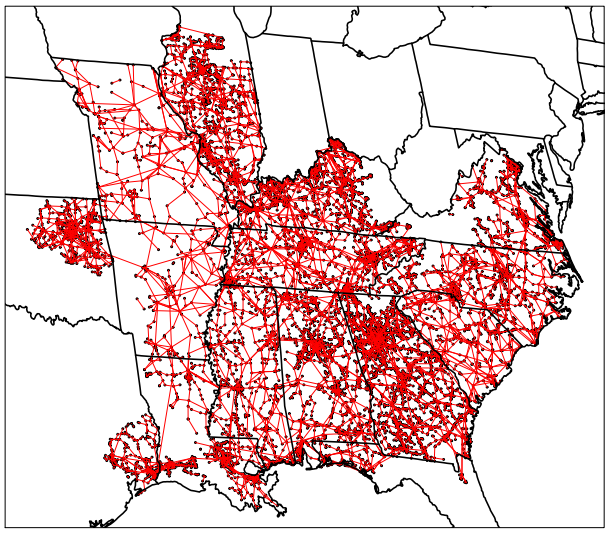


Fig. 13: A part of the Eastern Interconnection (EI) with 12,946 substations (nodes) and 16,658 lines (edges) that operates under the SERC.

B. SERC

We apply the GNLG Algorithm to part of the EI that operates under the SERC (see Fig. 13) that has 13,602 substations (nodes) and 17,767 lines (edges) . Fig. 14 shows the obtained network using the GNLG Algorithm with $\kappa = 3$, $\alpha = 0.5, \beta = 3.2, \gamma = 2.5$, and $N = 5$ that are selected empirically following several numerical experiments. In the SDNG Procedure, SERC has been clustered into $c = 50$ clusters based on the BIC.

The comparison between the degree distribution of the nodes and the length distributions of the lines in G_{SERC} and G'_{SERC} are shown in Figs. 15 and 16. Table IV, summarizes the structural properties of G_{SERC} and five instances generated by the GNLG Algorithm. As with the WI, it can be seen that the Algorithm can generate synthetic networks with similar structural properties to the SERC grid.

C. FRCC

Finally, we apply the GNLG Algorithm to a smaller part of the EI with 1,312 substations (nodes) and 1,780 lines (edges) that operates under the FRCC (see Fig. 17). As can be seen in Fig. 18, the degree distribution of the nodes in G_{FRCC} is different from the degree distribution of the nodes in G_{WI} and G_{SERC}. In G_{FRCC}, only the density of the nodes with degree 1 is not on the fitted regression line. This suggests that in the Reinforcement Procedure, the step for small networks

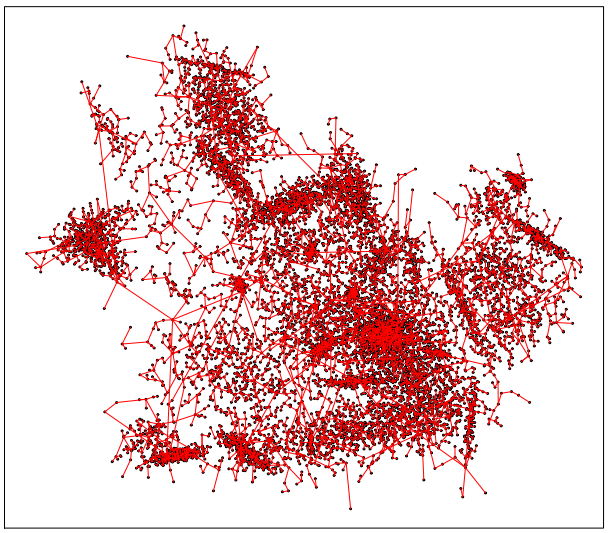
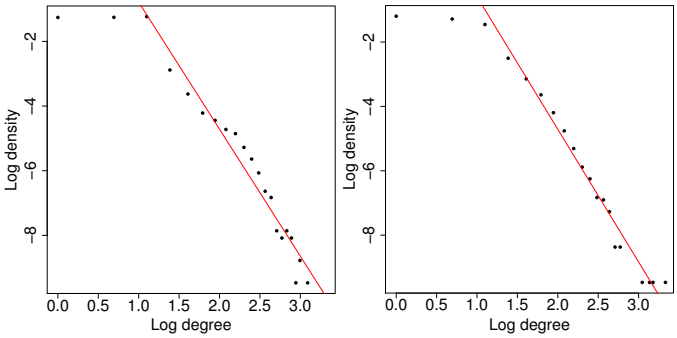


Fig. 14: A network with 12,946 nodes and 16,658 edges generated based on the SERC grid using the GNLG Algorithm with $\kappa = 3$, $\alpha = 0.5, \beta = 3.2, \gamma = 2.5$, and $N = 5$.



(a) G_{SERC} (b) G'_{SERC}

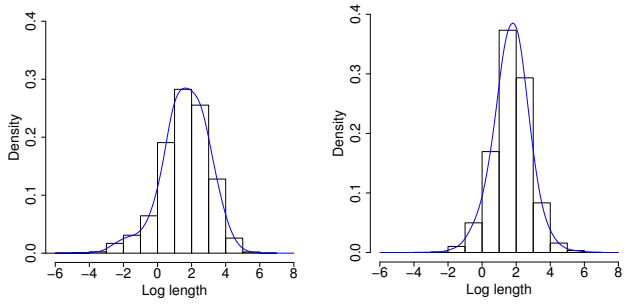
Fig. 15: The degree distribution of the nodes in G_{SERC} and G'_{SERC} (in log-log scale). Linear regression lines with slopes $\zeta = -3.93$ and $\zeta = -4.12$ are fitted to the distribution of the nodes with degree greater that 2 in G_{SERC} and G'_{SERC}, respectively. The KS statistic between the degree distributions is 0.047.

should be used and nodes should be sampled with probability $\propto d_i'^{-\eta}\rho_i'^{-\alpha}$. Here, we use $\eta = 2$.

Fig. 17 shows the obtained network using the GNLG Algorithm with $\kappa = 1.8, \alpha = 0.5, \beta = 2.5, \gamma = 2.8$, and $N = 5$ that were selected empirically. Nodes in the FRCC has been clustered into $c = 15$ clusters. The comparison between the degree distributions of the nodes and length distributions of the lines between G_{FRCC} and in G'_{FRCC} are shown in Figs. 18 and 19. Table V, summarizes the structural properties of the FRCC and five instances generated by the GNLG Algorithm. The results suggest that the GNLG algorithm can generate smaller networks as well.

VI. Conclusions

In this paper, we developed the GNLG Algorithm for generating synthetic power grid networks with similar structural properties to a given network. We applied the algorithm to the WI and two parts of the EI (SERC and FRCC) and showed

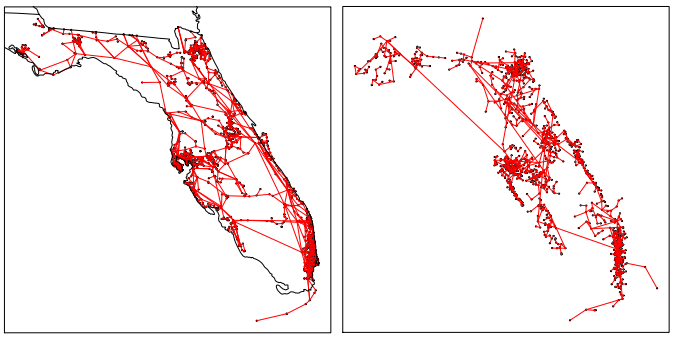


(a) G_{SERC} (b) G'_{SERC}

Fig. 16: The length (in *km*) distribution of the point-to-point lines in G_{SERC} and G'_{SERC} and nonparametric distribution fit (shown in blue). The KL-divergence between the length distribution of the lines in G_{SERC} and G'_{SERC} is 0.081.

TABLE IV: Comparison between the structural properties of the SERC (G_{SERC}) and the Generated SERC (G'_{SERC}). Five instances are shown to illustrate that the metric values are similar. All networks have 12,946 nodes and 16,658 edges.

Networks	L	C	ζ	D_{KS}	D_{KL}
G_{SERC}	19.71	0.049	-3.93	0	0
G'_{SERC}	20.26	0.048	-4.12	0.047	0.081
$G'_{SERC}(2)$	19.43	0.045	-4.25	0.044	0.077
$G'_{SERC}(3)$	17.56	0.048	-4.72	0.044	0.084
$G'_{SERC}(4)$	17.95	0.047	-4.46	0.048	0.083
$G'_{SERC}(5)$	19.87	0.049	-4.5	0.046	0.080

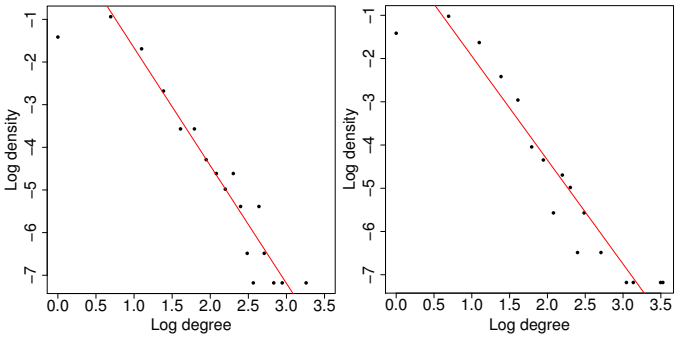


(a) G_{FRCC} (b) G'_{FRCC}

Fig. 17: (a) Part of the Eastern Interconnection (EI) with 1,312 substations (nodes) and 1,780 lines (edges) that operates under the FRCC. (b) A network with the same number of nodes and edges that is generated using the GNLG Algorithm with $\kappa = 1.8, \alpha = 0.5, \beta = 2.5, \gamma = 2.8,$ and $N = 5$.

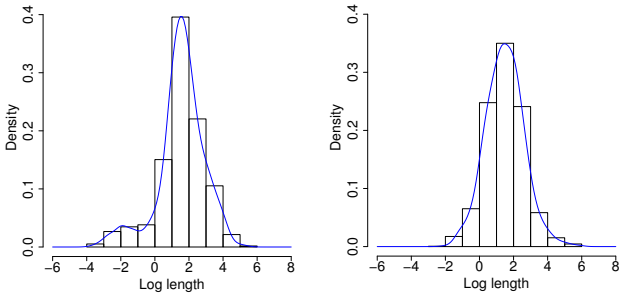
that it can generate networks with similar structural properties to these networks. In a broader perspective, the algorithm can be used for anonymizing network data that cannot be published otherwise, thereby enabling research in power grid vulnerability and resilience.

This is only a first step towards generation of synthetic power grid networks and there are clearly several future research directions. Specifically, for a given network, step 1 of the GNLG Algorithm and tuning the parameters need to be done only once. Then, the algorithm can be used to generate several networks similar to a given network. Hence, we plan to provide a web application that would allow obtaining synthetic networks similar to a given reliability regions in the Northern American power grid with specific set of parameters (e.g.,



(a) G_{FRCC} (b) G'_{FRCC}

Fig. 18: The degree distribution of the nodes in G_{FRCC} and G'_{FRCC} (in log-log scale). Linear regression lines with slopes $\zeta = -2.76$ and $\zeta = -2.40$ are fitted to the distribution of the nodes with degree greater that 1 in G_{FRCC} and G'_{FRCC}, respectively. The KS statistic between the degree distributions is 0.032.



(a) G_{FRCC} (b) G'_{FRCC}

Fig. 19: The length (in *km*) distribution of the point-to-point lines in G_{FRCC} and G'_{FRCC} and nonparametric distribution fit (shown in blue). The KL-divergence between the length distributions in G_{FRCC} and G'_{FRCC} is 0.12.

TABLE V: Comparison between the structural properties of the FRCC (G_{FRCC}) and the Generated FRCC (G'_{FRCC}). Five instances are shown to illustrate that the metric values are similar. All networks have 1,312 nodes and 1,780 edges.

Networks	L	C	ζ	D_{KS}	D_{KL}
G_{FRCC}	11.68	0.075	-2.76	0	0
G'_{FRCC}	10.81	0.045	-2.40	0.032	0.12
$G'_{FRCC}(2)$	11.86	0.057	-2.70	0.025	0.12
$G'_{FRCC}(3)$	11.13	0.053	-2.78	0.022	0.10
$G'_{FRCC}(4)$	11.27	0.051	-2.86	0.025	0.13
$G'_{FRCC}(5)$	11.66	0.057	-2.36	0.015	0.12

currently it takes less than 3.5 minutes for our server to generate a synthetic network similar to the WI). Moreover, we plan to improve the algorithm and to focus on locations of power generators and demand nodes as well as on generation and demand values. Generation of topologies where the line voltages are taken into account is also an interesting open problem. Finally, we believe that the approach can be extended for generating various types of spatially distributed networks.

ACKNOWLEDGEMENT

This work was supported in part by DTRA grant HDTRA1-13-1-0021, CIAN NSF ERC under grant EEC-0812072, the People Programme (Marie Curie Actions) of the European

Unions Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no. [PIIF-GA-2013-629740].11.

REFERENCES

[1] M. Amin and J. Stringer, "The electric power grid: Today and tomorrow," *MRS bulletin*, vol. 33, no. 4, pp. 399–407, 2008.

[2] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid - the new and improved power grid: A survey," *IEEE Commun. Surveys Tuts.*, vol. 14, no. 4, pp. 944–980, 2012.

[3] D. Bienstock, M. Chertkov, and S. Harnett, "Chance-constrained optimal power flow: risk-aware network control under uncertainty," *SIAM Review*, vol. 56, no. 3, pp. 461–495, 2014.

[4] S. Soltan, M. Yannakakis, and G. Zussman, "Joint cyber and physical attacks on power grids: Graph theoretical approaches for information recovery," in *Proc. ACM SIGMETRICS'15*, June 2015.

[5] Y. Zhao, A. Goldsmith, and H. V. Poor, "On PMU location selection for line outage detection in wide-area transmission networks," in *Proc. IEEE PES'12*, July 2012.

[6] G. Latorre, R. D. Cruz, J. M. Areiza, and A. Villegas, "Classification of publications and models on transmission expansion planning," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 938–946, 2003.

[7] S. Soltan, D. Mazauric, and G. Zussman, "Cascading failures in power grids – analysis and algorithms," in *Proc. ACM e-Energy'14*, June 2014.

[8] A. Bernstein, D. Bienstock, D. Hay, M. Uzunoglu, and G. Zussman, "Power grid vulnerability to geographically correlated failures - analysis and control implications," in *Proc. IEEE INFOCOM'14*, Apr. 2014.

[9] A. Asztalos, S. Sreenivasan, B. K. Szymanski, and G. Korniss, "Cascading failures in spatially-embedded random networks," *PloS one*, vol. 9, no. 1, p. e84563, 2014.

[10] M. Chertkov, F. Pan, and M. G. Stepanov, "Predicting failures in power grids: The case of static overloads," *IEEE Trans. Smart Grid*, vol. 2, no. 1, pp. 162–172, 2011.

[11] "IEEE benchmark systems," available at http://www.ee.washington.edu/research/pstca/.

[12] "National Grid UK," available at http://www2.nationalgrid.com/uk/\services/land-and-development/planningauthority/.

[13] "Polish grid," available at http://www.pserc.cornell.edu/matpower/.

[14] Q. Zhou and J. W. Bialek, "Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 782–788, 2005.

[15] E. Cotilla-Sanchez, P. D. Hines, C. Barrows, and S. Blumsack, "Comparing the topological and electrical structure of the North American electric power infrastructure," *IEEE Syst. J.*, vol. 6, no. 4, pp. 616–626, 2012.

[16] Platts, "GIS Data," http://www.platts.com/Products/gisdata.

[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[19] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *PNAS*, vol. 97, no. 21, pp. 11 149–11 152, 2000.

[20] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Phys. Rev. E*, vol. 69, no. 2, p. 025103, 2004.

[21] P. Crucitti, V. Latora, and M. Marchiori, "A topological analysis of the Italian electric power grid," *Phys. A*, vol. 338, no. 1, pp. 92–97, 2004.

[22] D. P. Chassin and C. Posse, "Evaluating North American electric grid reliability using the Barabási–Albert network model," *Phys. A*, vol. 355, no. 2, pp. 667–677, 2005.

[23] J. D. Glover, M. Sarma, and T. Overbye, *Power System Analysis & Design, 4th Edition*. Cengage Learning, 2011.

[24] M. Rosas-Casals, S. Valverde, and R. V. Solé, "Topological vulnerability of the European power grid under errors and attacks," *Int. J. Bifurcat. Chaos*, vol. 17, no. 07, pp. 2465–2475, 2007.

[25] R. V. Solé, M. Rosas-Casals, B. Corominas-Murtra, and S. Valverde, "Robustness of the European power grids under intentional attack," *Phys. Rev. E*, vol. 77, no. 2, p. 026102, 2008.

[26] M. A. S. Monfared, M. Jalili, and Z. Alipour, "Topology and vulnerability of the Iranian power grid," *Phys. A*, vol. 406, pp. 24–33, 2014.

[27] M. M. Danziger, L. M. Shekhtman, Y. Berezin, and S. Havlin, "Two distinct transitions in spatially embedded multiplex networks," *arXiv:1505.01688*, 2015.

[28] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 28–39, 2010.

[29] P. Schultz, J. Heitzig, and J. Kurths, "A random growth model for power grids and other spatially embedded infrastructure networks," *Eur. Phys. J. Spec. Top.*, vol. 223, no. 12, pp. 2593–2610, 2014.

[30] M. Barthelemy, "Spatial networks," *arXiv:1010.0302v2*, 2010.

[31] S. S. Manna and P. Sen, "Modulated scale-free network in Euclidean space," *Phys. Rev. E*, vol. 66, no. 6, p. 066114, 2002.

[32] R. Xulvi-Brunet and I. M. Sokolov, "Evolving networks with disadvantaged long-range connections," *Phys. Rev. E*, vol. 66, no. 2, p. 026118, 2002.

[33] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

[34] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[35] S. Boltz, E. Debreuve, and M. Barlaud, "High-dimensional statistical measure for region-of-interest tracking," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1266–1283, 2009.

[36] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Am. Statist. Assoc.*, vol. 97, no. 458, pp. 611–631, 2002.

[37] C. Fraley, A. E. Raftery, T. B. Murphy, and L. Scrucca, "mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation." Department of Statistics, University of Washington, Tech. Rep. 597, 2012.

[38] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.