Generation of Synthetic Spatially Embedded Power Grid Networks

Saleh Soltan, Gil Zussman
Department of Electrical Engineering
Columbia University, New York, NY
Email: {saleh,gil}@ee.columbia.edu

***Abstract*—The development of algorithms for enhancing the resilience and efficiency of the power grid requires evaluation with topologies of real transmission networks. However, due to security reasons, *such topologies and particularly the locations of the substations and 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 has several tunable parameters that allow generating grids similar to any given grid. 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 the generated grids have 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 grids.**

I. Introduction

The design of algorithms and methods for enhancing the power grid drew tremendous attention over the past decade [1]. These efforts focused on challenges stemming from renewable generation interconnection, Phasor Measurement Units (PMUs) placement, transmission expansion planning, and vulnerability analysis. 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 [2], the National Grid UK [3], the Polish grid [4], and an approximate model of the European interconnected system [5]. 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. Therefore, in this paper *we design the Geographical Network Learner and Generator (GNLG) 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.

To demonstrate the algorithm design and to evaluate its performance, we focus on the transmission networks of the North American and Mexican power grids using data that we obtained from the Platts Geographic Information System (GIS) [6]. We consider one of the two major interconnections – the Western Interconnection (WI) (see Fig. 1) and 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].

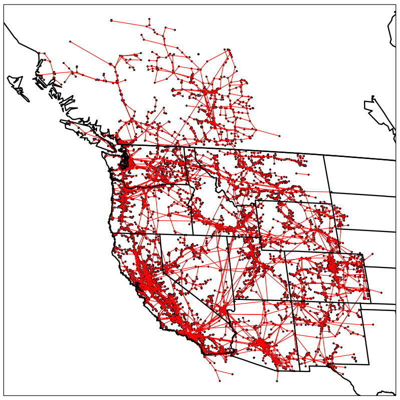


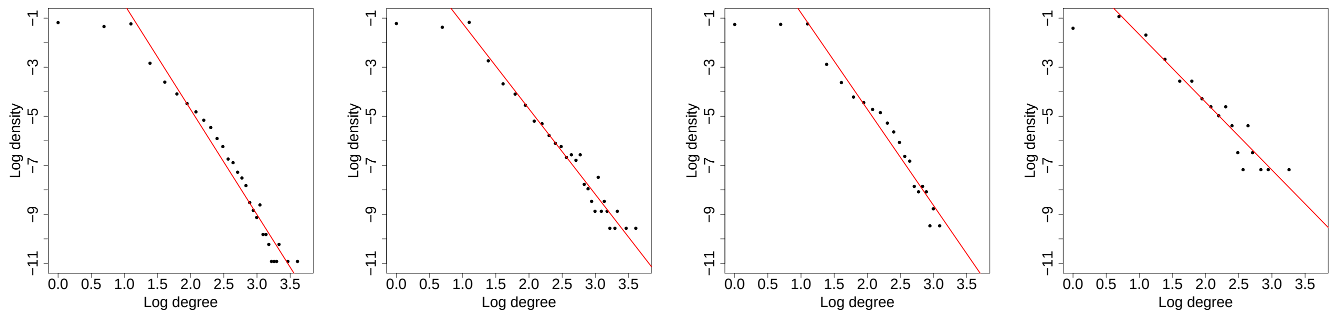
Fig. 1: The Western Interconnection (WI) power grid with 14,302 substations (nodes) and 18,769 lines (edges).

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. Due to space constraints more details on the structural properties of the networks and evaluation results appear in a technical report [7].

II. Related Work

The structural properties of various power grids (e.g., in North America, some European countries, and Iran) were studied in [8], [9], [10], [11], [12]. 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 a certain class of graphs (e.g. small-word and scale-free graphs)

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



(a) NA&M

(b) WI

(c) SERC

(d) FRCC

Fig. 2: 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

was suggested as a good representative of a power grid network, based on one or two structural properties. However, by comparing the WI with these models, [13] 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 [14], [15] 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 [16], 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.

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

B. Clustering coefficient

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: $C = \sum_{i \in V} C_i / n$.

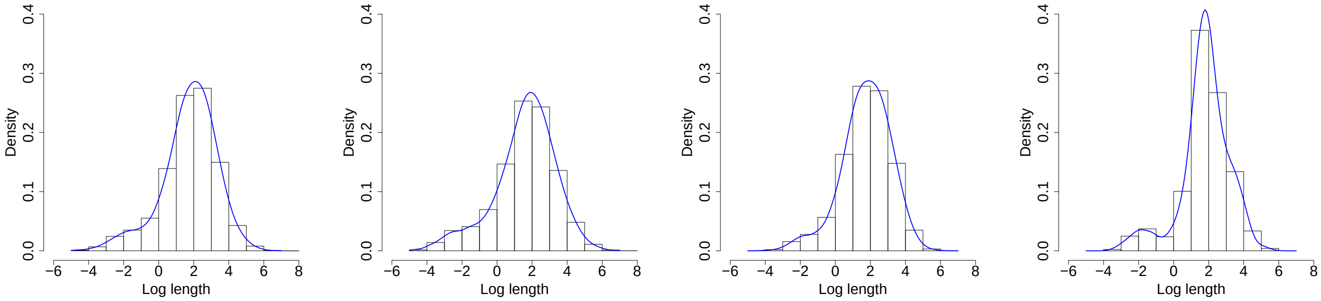
C. Degree distribution of the nodes

Fig. 2 shows the degree distribution of the nodes in the NA&M, WI, SERC, and FRCC grids in log-log scale. These figures may suggest that the tail of the degree distribution follows a power-law distribution in all the three networks. However, 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 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 them is defined as: $D_{KS} = \max_x |P(x) - Q(x)|$.

D. Length distribution of the lines

Fig. 3 shows the length distribution of the lines in the NA&M, WI, SERC, and FRCC grids. The line lengths in Figs. 3 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



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

Fig. 3: The distributions of the actual line lengths (in *km*) in the NA&M, WI, SERC, and FRCC grids. Nonparametric distribution fits to the log length distributions are shown in blue.

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 $G'$.
```

lengths in the real and the generated networks. For the detailed statistics of the length of the lines see [7].

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 between two probability distribution functions p and q is: $D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$.

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 first 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 (the TWST and Reinforcement procedures described in Subsection IV-B). In the following subsections, we describe the building blocks of the GNLG Algorithm.

A. Node positions

The node positions are correlated with the population and geographical properties (e.g., Fig. 1). Thus, the nodes can be clustered into groups based on their geographical proximity using mixture models and in particular Gaussian Mixture Models (GMM). Hence, the Spatially Distributed Nodes Generator (SDNG) Procedure uses the GMM for clustering the positions and uses the Bayesian Information Criterion (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.

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 $\pi$ 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 $\{\mathbf{p}'_i\}_{i=1}^n$.
```

Procedure 2: Tunable Weight Spanning Tree (TWST)

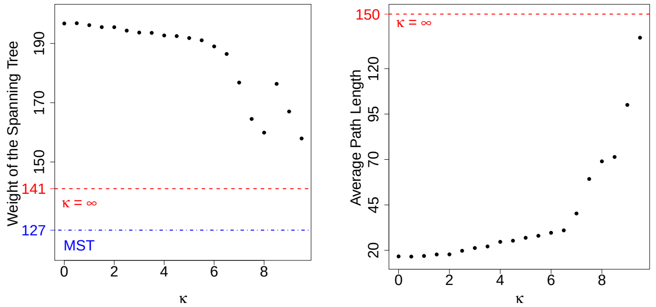
```
Input: $n, \{\mathbf{p}'_i\}_{i=1}^n$, and parameter $\kappa$.
1: $A = \{1, \ldots, n\}$, $\sigma$ is an empty array of size $n$.
2: for $i = 1 \ldots, n$ do
3:     Sample a node from $A$ such that the probability of sampling node $j$
       is $\frac{\|\mathbf{p}'_j - \bar{\mathbf{p}}'\|^{-\kappa}}{\sum_{a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

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.

1) Connectivity: In order for the power grid to operate, the substations (nodes) should be connected. We present the Tunable Weight Spanning Tree (TWST) Procedure (Procedure 2), which imitates the the gradual grid evolution. 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 $\sigma : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$. 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 $T = (V_T, E_T)$ whose weight ($\sum_{\{i,j\} \in E_T} \|\mathbf{p}'_i - \mathbf{p}'_j\|$) highly depends on the ordering of the nodes, and thereby on κ. Fig. 4(a) shows the relationship between the weight of the obtained tree and κ. Fig. 4(b) shows the relationship between κ and the average path length in



(a)

(b)

Fig. 4: (a) The weight of the spanning tree (in $10^3 km$) obtained by the TWST Procedure on the nodes generated by the SDNG Procedure 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 generated by the SDNG Procedure 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 3: Reinforcement

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

the obtained tree. Overall, Figs. 4(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.

2) Robustness: We present the Reinforcement Procedure (Procedure 3) whose objective is to increase the robustness of the generated 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. 2), (ii) it is inefficient and unsafe for the power grids to include very long lines (e.g., Figs. 3), and (iii) nodes in denser areas are more likely to have higher degrees. The last observation is demonstrated by Fig. 5 where as the degree increases, the ρ decreases[2] (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 [17]) which is also nearby (distance was not considered in [17]) (observations (i) and (ii)).

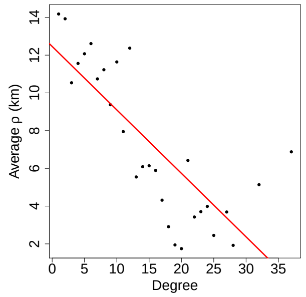


Fig. 5: The relationship between the degree of a node and its average ρ with $N = 10$, for the nodes in the WI (the red line is the linear regression fit to the data points).

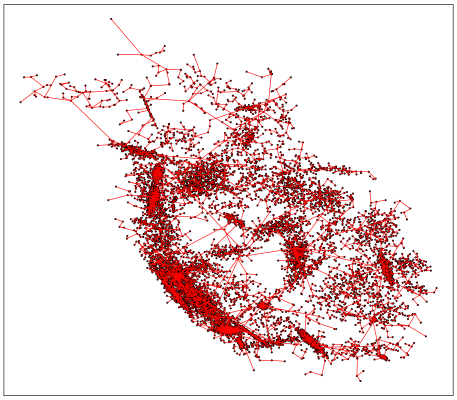


Fig. 6: A network with 14,302 nodes and 18,769 edges generated based on the WI grid using the GNLG Algorithm with $\kappa = 2.5$, $\alpha = 1, \beta = 3.2, \gamma = 2.5$, and $N = 10$.

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. 2, 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.

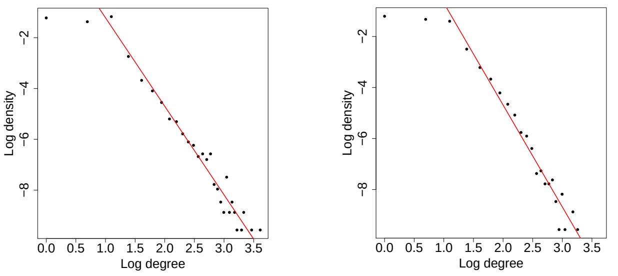
To connect the node sampled in the previous step to a high degree but nearby node, in the second step, the Reinforcement Procedure connects node i to node j sampled from all other nodes with probability $\propto \|\mathbf{p}'_i - \mathbf{p}'_j\|^{-\beta} d_j'^{\gamma}$. This implies that node i preferentially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

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. The details for the evaluation of the GNLG Algorithm on the SERC and FRCC can be found in [7].

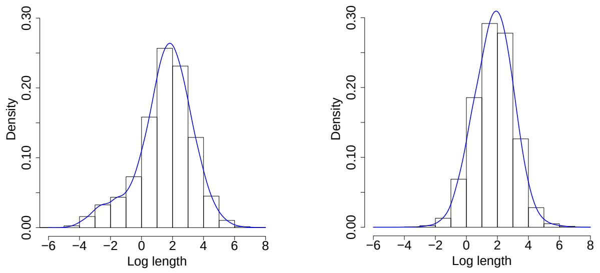
The network obtained by the GNLG Algorithm appears in Fig. 6 and visually resembles the WI. We empirically selected the following parameters values: $\kappa = 2.5, \alpha = 1, \beta = 3.2, \gamma = 2.5$, and $N = 10$. To study the structural similarity between the obtained network G'_{WI} and the G_{WI}, we evaluated G'_{WI}

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



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

Fig. 7: 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. 8: 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.

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. 7 and fig. 8 also show that the degree distribution of the nodes and length distribution of the lines in G_{WI} and G'_{WI} are very similar.

Table II summarizes the structural properties of the WI, SERC, and FRCC and corresponding Generated networks. The results indicate that the Algorithm can generate synthetic networks with similar structural properties to these grids.

VI. Conclusions

In this paper, we developed the Geographical Network Learner and Generator (GNLG) Algorithm for generating synthetic power grid networks with similar structural properties to a given network. 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.

In general, this is only a first step towards generation of comprehensive synthetic power grid network as described in a recent call by the U.S. department of energy [18] and there are clearly several future challenges. Specifically, 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. Moreover, we plan to compare the resiliency of the generated networks to real ones to line failures using DC power flow model. Generation of topologies where the are taken into account is also an interesting open problem.

TABLE II: Comparison between the structural properties of the WI (G_{WI}), SERC (G_{SERC}), and FRCC (G_{FRCC}) and the Generated WI (G'_{WI}), SERC (G'_{SERC}), and FRCC (G'_{FRCC}).

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_{SERC}	19.71	0.049	-3.93	0	0
G'_{SERC}	20.26	0.048	-4.12	0.047	0.081
G_{FRCC}	11.68	0.075	-2.76	0	0
G'_{FRCC}	10.81	0.045	-2.40	0.032	0.12

Acknowledgement

This work was supported in part by 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, and DTRA grant HDTRA1-13-1-0021. The work of G. Z. was also supported in part by the Blavatnik ICRC.

References

[1] 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.

[2] “IEEE benchmark systems,” available at http://www.ee.washington.edu/research/pstca/.

[3] “National Grid UK,” available at http://www2.nationalgrid.com/uk/services/land-and-development/planningauthority/.

[4] “Polish grid,” available at http://www.pserc.cornell.edu/matpower/.

[5] 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.

[6] Platts, “GIS Data,” http://www.platts.com/Products/gisdata.

[7] S. Soltan and G. Zussman, “Generation of synthetic spatially embedded power grid networks,” *arXiv:1508.04447*, 2015.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] M. Barthélemy, “Spatial networks,” *Physics Reports*, vol. 499, no. 1, pp. 1–101, 2011.

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

[18] Avanced Research Projects Agency Energy (ARPA-E), “Generating realistic information for the development of distribution and transmission algorithms (grid data),” 2015, available at https://arpa-e-foa.energy.gov/#FoaId986e4f22-e6af-4423-86ce-5632ce8dfafb.