A Learning-based Method for Generating Synthetic Power Grids

Saleh Soltan, *Member, IEEE,* Alexander Loh, *Student Member, IEEE,* and Gil Zussman, *Senior Member, IEEE*

***Abstract*—Analysis and improvement of power grids resilience and efficiency requires the topologies and geographical coordinates of the real transmission networks. However, due to security reasons, such topologies and particularly the locations of the substations and lines are usually not publicly available. In this work, we thoroughly study the structural properties of the U.S. Western Interconnection grid (WI) and based on the results present the Network Imitating Method Based on LEarning (NIMBLE) for generating synthetic spatially embedded networks with similar properties to a given grid. We apply NIMBLE to the WI and show that it can generate networks with similar structural and spatial properties as well the same level of robustness to failures to the WI, without revealing the real locations of the lines and substations. To the best of our knowledge, this is the first attempt to consider the spatial distributions of the buses (nodes) and lines and their importance in generating synthetic grids. Moreover, this is the first time that the power flows and vulnerability against failures are considered in evaluating a synthetic power grid.**

***Index Terms*—Power Grids, Topology, Robustness, Data Mining, Complex Networks.**

I. Introduction

Enhancing power grids performance and resilience (namely, making it smarter) has been one of the greatest challenges in science and engineering over the past decade [2], [3]. This challenge spans numerous aspects of the power systems such as incorporating renewable resources [4], efficient and robust grid monitoring [5], transmission expansion planning [6], grid vulnerability analysis and control [7]–[10], cyber security, and energy market design. Addressing most of these challenges requires real grid topologies with real geographical coordinates. For example, incorporating renewable energy sources in the grid requires the approximate locations of the grid lines and buses to be matched with the wind maps. Similarly, to study the vulnerability of the grid to natural disasters one needs to match the grid map with the paths of hurricanes or water flood maps. However, in order to avoid exposing vulnerabilities, 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 limited test cases and real-world power grid data sets 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]. Therefore, there is a growing interest in generating synthetic power grids [16]–[19].

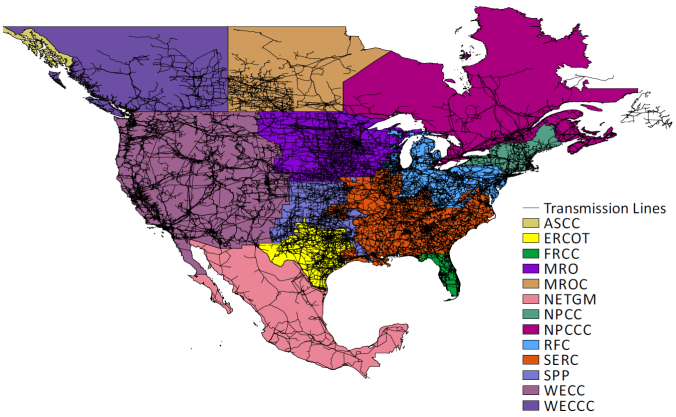


Fig. 1: The North American Electric Reliability Corporation (NERC) regional entities and the National Electricity Transmission Grid of Mexico (NETGM). Different reliability corporations/councils are marked with different colors.

Motivated by this need, we thoroughly study the structural properties of real power grids and introduce the Network Imitating Method Based on LEarning (NIMBLE) for generating synthetic networks with similar structural and spatial properties. We focus on the transmission networks of the North American and Mexican power grids (see Fig. 1) using data that we obtained from the Platts Geographic Information System (GIS) [20] (Similar techniques could be applied to industry-grade data sets, if they become publicly available). In particular, we consider the Western Interconnection (WI), one of the two major interconnections in North America, which includes the Western Electricity Coordinating Council in the United States (WECC) and Canada (WECCC).

We evaluate the networks' structural properties under five metrics: *average path length (L), clustering coefficient (C), degree distribution of the nodes, number of line intersections ($\mathcal{X}$), and the length distribution of the lines*. For each node i with degree d_i, at most $d_i(d_i-1)/2$ lines can exist between its neighbors. The clustering coefficient is the fraction of these allowable lines that actually exist, averaging over all the nodes.

S. Soltan is with the Electrical Engineering Department at Princeton University, Princeton, NJ. This work was done while Saleh Soltan was with Columbia University. A. Loh, and G. Zussman are with the Department of Electrical Engineering at Columbia University, New York, NY, 10027.
E-mails: ssoltan@princeton.edu, al3475@columbia.edu, gil@ee.columbia.edu
A partial and preliminary version appeared in Proc. IEEE PES-GM'16 [1].

The average path length is defined as the number of lines in the shortest path between two nodes, averaged over all pairs of vertices. The first three metrics are very common [15], [21]–[26]. The importance of the number of line intersections in generating synthetic power grids was first discussed by Birchfield, et al. [17]. However, to the best of our knowledge, the length distribution of the lines has not been thoroughly studied before in power grids. The line lengths are particularly important in power grids, since the physical properties of a line (e.g., admittance and type) are directly correlated with its length [27], and hence, directly impact the grid's structural properties.

To compare the robustness of the WI and the generated network to failures, we simulate cascading failures initiated by double line failures as well as circular area failures (as failures caused by natural disasters), and compute the *yield (the ratio between the demand supplied at the end of a cascade and the original demand), number of failed lines, and number of connected components* at the end of the cascade in these networks. Cascading failures in networks and power grids and their robustness have been widely studied [25], [26], [28]–[37]. In this paper, we follow one of the previously suggested cascade models due to line overloads in power grids with a *deterministic outage rule*: namely, a line fails when the magnitude of the flow on that line exceeds its capacity [7], [8], [31]. We show that the generated networks have similar structural and spatial properties as well the same level of robustness to failures to the WI.

The main contributions of our paper are: (i) introducing a novel method to generate synthetic power grid networks–similar to any given grid network–that preserve the topological and operational properties of the actual grid without revealing any of its critical information, (ii) providing a rigorous analysis of the topological and robustness properties of the WI, (iii) applying our method to the WI and demonstrating the accuracy of the generated networks in representing WI's topological and robustness properties. Our synthetically generated version of the WI is publicly available at our website [38] and also on the Data Repository for Power system Open models With Evolving Resources (DR POWER) project [39].

Our work allows the grid operators to provide a synthetic version of their grid data to the researchers without revealing any critical information, and at the same time allows the researchers to get access to many test cases. Such test cases are critical for designing new methods for improving power grids' robustness against failures and attacks.

II. Related work

The structural properties of various power grids (e.g., in North America, some European countries, and Iran) were studied before [21], [25], [40]–[43]. 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 was suggested as a good representative of a power grid network, based on one or two structural properties [15], [21]–[26]. For example, Watts and Strogatz

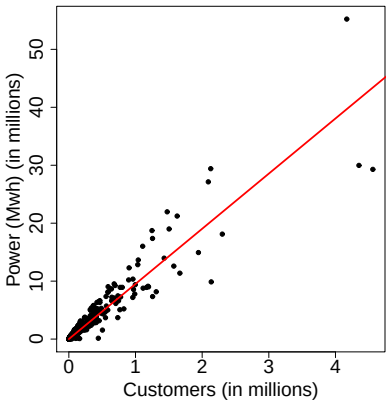


Fig. 2: The average residential power usage versus the number of customers in the US. The slope of the regression line is 9.51.

[21] 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 [22] 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 by Wang, et al. [44] and also by Schultz, et al. [45] but they did not consider the *spatial distribution* of the nodes 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 [46]–[48], most of them were not designed to generate networks with properties similar to power grid networks'. In very recent novel papers [16], [17], a synthetic network based on the locations of the cities and the power plants in Texas was generated. Despite using the geographical locations, no comparisons to the real grid in Texas (neither topological nor performance wise) were provided by the authors. The generated networks, however, was shown to have similar topological (e.g., degree distribution) properties to the WI. In a follow up work [18], more engineering properties (e.g., transformer and generator parameters) of power grids were studied in detail.

The novel methods in [16], [17] consider more engineering details than our work but do not provide a general framework to generate multiple synthetic power grids. This paper is the first to consider the spatial distribution of the nodes (buses) in power grids and to provide a general framework for generating multiple synthetic networks with realistic structural properties. Moreover, this is the first time that *the power flows and vulnerability against cascading failures* are considered in evaluating a synthetic power grid. In our future work, we plan to incorporate more engineering details [18] to improve the quality of our generated power grids to the level of industrial grade grids.

III. Preliminaries

A. Data

We obtained the topological network of the WI from the Platts GIS [20] and conducted longitude-latitude to planar (x, y) coordinate transformation, using the great-circle distance method. We extracted the coordinates of the buses/substations from the endpoint coordinates of the lines. We then used the geographical coordinates of the substations and the lines to construct the graph 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 (See Fig. 1).

The GIS does not provide the reactance values of the lines. The reactance of a line depends on its geometrical properties and there is a linear relation between the line's length and reactance: the longer the line is, the larger its reactance. Thus, we assumed that all lines have the same physical properties (other than length) and used the length to determine the reactance. Moreover, since the power flow solution is scale invariant of the reactance values, we simply use the length of each line as its reactance. We consider the same for the lines in the generated networks.

To estimate the demands and supplies, we used the cities' populations and the power plants' capacities, as well as their locations. The locations and populations of the cities in the U.S. and Canada are publicly available. We obtained the information about the locations of the power plants and their capacities from U.S. Energy Information Administration (EIA) website [49]. We observed that the average residential power usage (obtained from EIA website [49]) is directly related to the population (see Fig. 2). By modeling this relationship using linear regression, we estimated the power demand at each city based on its population. Once we computed the total demand, we assigned the power generation level to each power plant according to its capacity to supply the demand.

In the WI, we assigned each city (along with its demand) or power plant (along with its supply) to its closest node.

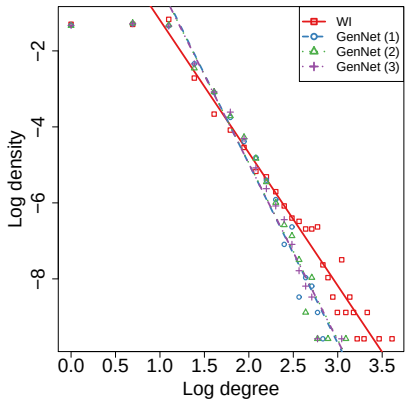
B. Degree and line length distributions comparison

To compare the degree distributions of the nodes in the WI and the generated networks, we use the Kolmogrov-Smirnov (D_{KS}) statistic [50]. If $P(x)$ and $Q(x)$ are two Cumulative Distribution Functions (CDFs), the KS statistic between these two is $D_{KS} = \max_x |P(x) - Q(x)|$.

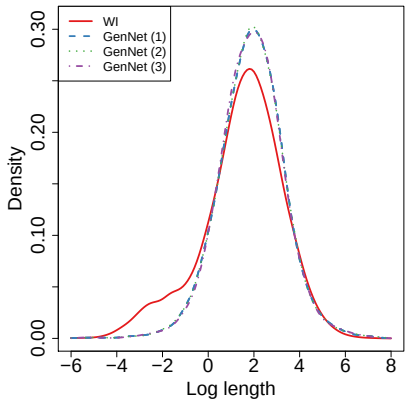
To measure the similarity between the length distributions of the lines in a given network and a generated network, we use the Kullback-Leibler (D_{KL}) divergence. Specifically, the KL-divergence of distribution q from p, denoted by $D_{KL}(p\|q)$, is a measure of the information lost when q is used to approximate p: $D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} dx$.

C. Cascading failures model

We adopt the DC power flow model, which is widely used as an approximation for the more accurate non-linear AC power flow model [51]. We represent the power grid by an *undirected graph* $G = (V, E)$ where V and E are the set of nodes and edges corresponding to the buses and transmission lines, respectively. p_v is the active power *supply* ($p_v > 0$) or *demand* ($p_v < 0$) at node $v \in V$ (for a *neutral node* $p_v = 0$). We assume *pure reactive* lines, implying that each edge $\{u, v\} \in E$ is characterized by its *reactance* $x_{uv} = x_{vu} > 0$.



(a) Degree distributions



(b) Length distributions

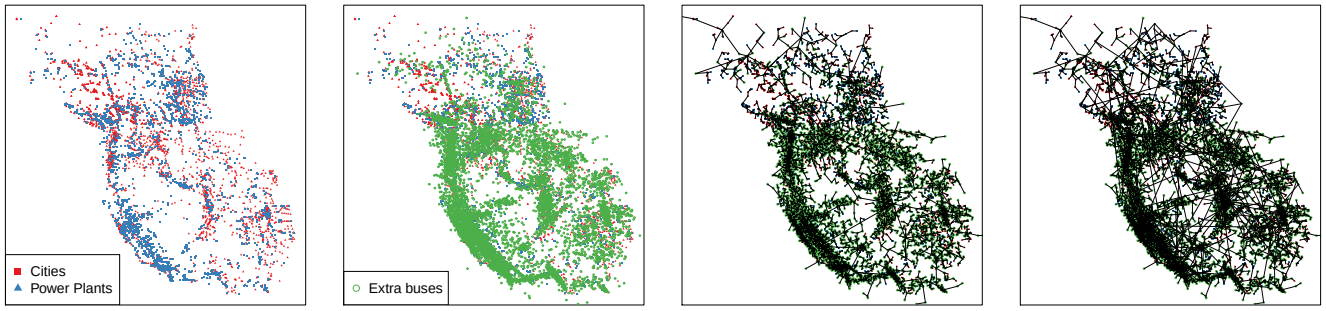
Fig. 3: The degree distributions of the nodes and length distributions of the lines in the WI and three generated networks. **(a)** The degree distributions of the nodes in log-log scale. In this figure, the slope of the fitted linear regression line to the distributions' tail is $\zeta = -3.49$ for the WI and $\zeta \approx -4.7$ for all the generated networks. The Kolmogrov-Smirnov statistic between the degree distributions in the WI and the generated networks is $D_{KS} \approx 0.05$. **(b)** The log length distributions of the lines (in *km*). In this figure, $D_{KL} \approx 0.1$ between the length distributions of the lines in the WI and the generated networks.

Given the power supply/demand vector $P \in \mathbb{R}^{|V|\times 1}$ and the reactance values, a *power flow* is a solution (f, θ) of:

$$\sum_{v \in N(u)} f_{uv} = p_u, \ \forall \ u \in V \tag{1}$$

$$f_{uv} = (\theta_u - \theta_v)/x_{uv}, \ \forall \ \{u, v\} \in E \tag{2}$$

where $N(u)$ is the set of neighbors of node u, f_{uv} is the power flow from node u to node v, and θ_u is the phase angle of node u. Eq. (1) guarantees (classical) flow conservation and (2) captures the dependency of the flow on the reactance values and phase angles. Additionally, (2) implies that $f_{uv} = -f_{vu}$. *Note that the edge capacities are not taken into account in determining the flows.* When the total supply equals the



(a) (b) (c) (d)

Fig. 4: NIMBLE's steps for generating a synthetic grid similar to the WI. **(a)** In the first step, it picks the locations of the cities and power plants as a subset of the nodes (buses) in the generated network. **(b)** In the second step, by estimating the spatial density of the nodes in the WI using GMM, it adds more nodes to the generated network to make the total number of nodes equal to the one in the WI. **(c)** In the third step, it finds an spanning tree of the nodes by connecting each node to its closest node with a higher index to ensure the connectivity of the network. **(d)** In the last step, it adds more lines to the network to increase the robustness of the generated network and adjust its properties (e.g., total number of lines and degree distribution) to resemble those of the WI. It does it by repeatedly selecting a low degree node in a dense area and connecting it to a high degree node which is also nearby.

total demand in each connected component of G, (1)-(2) has a unique solution [52, lemma 1.1]. The uniqueness is in the values of f_{uv}'s rather than θ_u's (shifting all θ_u's by equal amounts does not violate (2)).

We follow one of the previously suggested cascade models due to line overloads in power grids with a *deterministic outage rule* [7], [8], [31]: namely, a line $\{u, v\}$ fails when the magnitude $|f_{uv}|$ of the flow on that line exceeds its capacity c_{uv}. The line flow capacities are estimated as $c_{uv} = (1+\alpha)\max\{|f_{uv}|, \overline{f}\}$, where $\overline{f}$ is the median of the initial magnitude of line flows and α is the lines' factor of safety. For comparison purposes, we select the median of the initial magnitude of line flows in the WI as the minimum capacity for the lines in the generated networks as well.

When a line fails, it is removed from the network. As a result of this removal, the network topology is changed, and the network can be divided into one or more connected components. We assume that each connected component can operate autonomously. Therefore, within each connected component with non-zero supply and demand, the amounts of the supply and demand are balanced by either scaling down all the supply values (if supply is greater than demand in the connected component) or scaling down all the demand values (if demand is greater than supply in the connected component). If there is either no supply or no demand node within a connected component, all demands or supplies become zero.

After supply and demand balancing in each component, the power flow equations are solved to compute new flows on the lines. Using the deterministic outage rule, the new set of line failures are then found in all the components, and the cascade continues with the removal of those lines. If there are no overloaded lines in any of the components, the cascade terminates.

We use *yield (the ratio between the demand supplied at the end of a cascade and the original demand), number of failed lines, and number of connected components* at the end of the cascade to evaluate its severity.

TABLE I: Comparison between the structural properties of the WI and the generated networks. Three instances of GenNet are shown to illustrate that the metric values are similar in various generated networks. All networks have 14,430 nodes and 18,884 lines.

Networks	L	C	ζ	D_{KS}	$\mathcal{X}$	D_{KL}
WI	17.44	0.048	-3.49	0	7,358	0
GenNet (1)	16.28	0.048	-4.72	0.051	12,108	0.10
GenNet (2)	16.02	0.045	-4.65	0.043	12,132	0.10
GenNet (3)	16.28	0.050	-4.66	0.052	11,145	0.10

D. Computation

For analyzing the power grid topological properties, we used the `igraph` library in `R` [53]. This library provides a collection of network analysis tools.

To estimate the KL-divergence between distributions, we used the `FNN` library in `R` which utilizes the method introduced by Boltz, et al. [54] for estimating the KL-divergence between two distributions using their samples.

For fitting a Gaussian Mixture Model (GMM) (see Section IV), we used the `mclust` library in `R` [55]. 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.

IV. NIMBLE

The structural properties of the WI are shown in Fig. 3 and Table I. The main observations are: (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(a)), (ii) it is inefficient and unsafe for the power grids to include very long lines (e.g., Fig. 3(b)), and (iii) the average path length and the clustering coefficient demonstrate the *small-world* property of power grids [21].

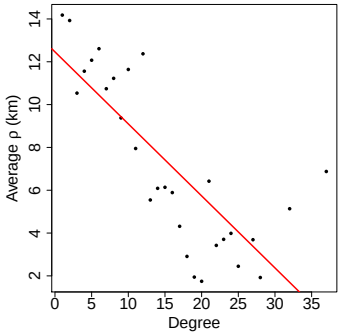


Fig. 5: The relationship between the degree of a node and its average ρ with $k = 10$, for the nodes in the WI (the red line is the linear regression fit to the data points).

Based on these characteristics, we introduce the Network Imitating Method Based on LEarning (NIMBLE) for generating synthetic networks similar to the real power grids. The NIMBLE steps are summarized in Fig. 4.

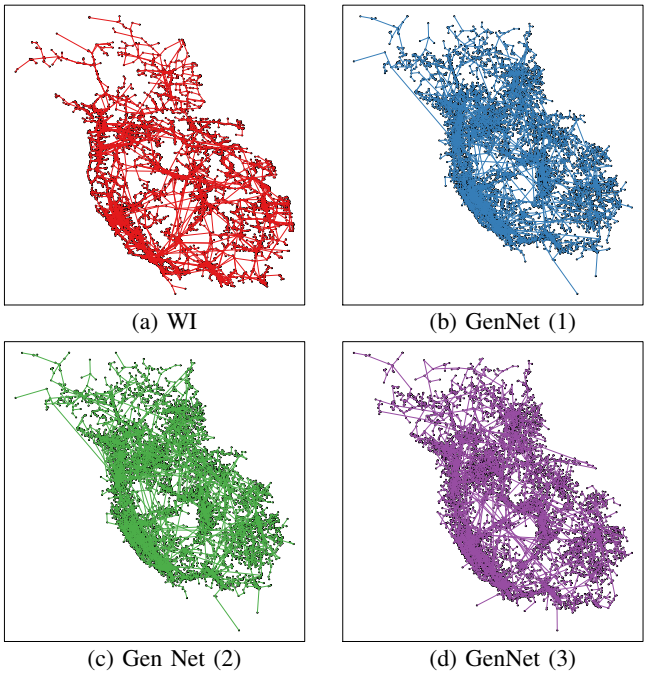
In the first step, the NIMBLE picks the locations of the cities and power plants as a subset of the nodes (buses) in the generated network. In the general case, if the locations of the loads are available for the real grid, those locations can be used instead of the cities. This step facilitates the mapping process of the supply and the demand values to the generated network, since the real supply and demand values (if available) can directly be used from the real network in the generated network at the same locations as the real grid.

Since the cities and the power plants are usually the end-points of the network, more nodes need to be added to the generated network to make the total number of nodes equal to the number of nodes in a given grid (here the WI). Notice that if the locations of the cities and power plants are not available and only the topological properties of the given network need to be imitated, step 1 of the method can be skipped (see the arXiv report [56], for a more general algorithm for this case).

The node (buses/substations) positions are correlated with the population and geographical properties. Thus, the nodes can be clustered into groups based on their geographical proximity using mixture models and in particular the Gaussian Mixture Models (GMM). Hence, in the second step, NIMBLE uses a 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 more nodes with similar spatial distribution as the nodes in a given network to make the total number of nodes equal in the given and generated networks. For the WI, we select $c = 55$ based on the BIC.

To connect the nodes, the NIMBLE takes two steps inspired by the historical evolution of power grids. The two main design consideration of the grid are (i) connectivity and (ii) robustness.

In order for the power grid to operate, the substations (nodes) should be connected. To satisfy the connectivity of the generated network, in the third step, the NIMBLE connects each node i to its closest (Euclidean distance) node j such that $i < j$ in order to form a spanning tree of the nodes. This is



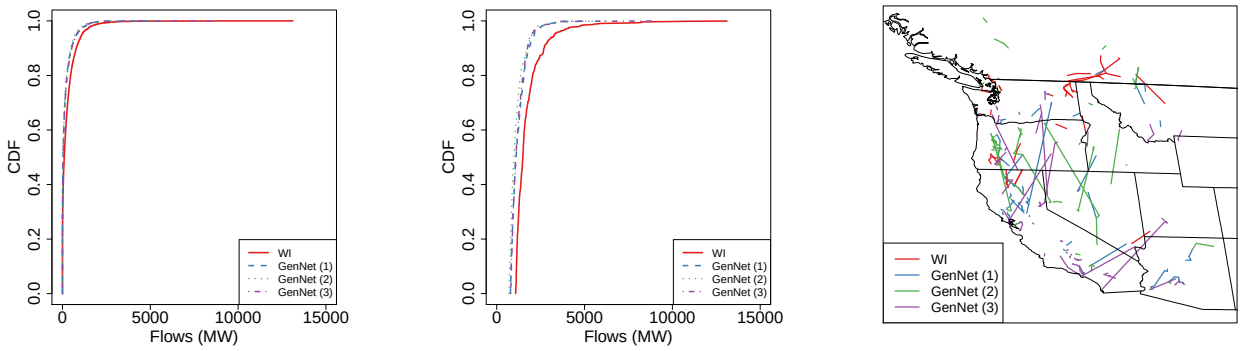
(a) WI (b) GenNet (1) (c) Gen Net (2) (d) GenNet (3)

Fig. 6: The Western Interconnection (WI) power grid and the three generated networks with 14,430 buses (nodes) and 18,884 lines.

inspired by the way the power grids are evolved. Each new bus or substation is connected to the closest bus and substation that already exist in the grid. Since the cities and the power plants are usually the end points in the power grid, in the NIMBLE, these nodes have the lowest index. At the end of this step, the network gets connected. Notice that the resulting tree differs from the minimum spanning of the nodes since the obtained tree depends on the order of the nodes (for more details see the arXiv report [56]).

In the last step, the NIMBLE repeatedly adds lines to the generated network to increase its robustness and adjust its properties (e.g., total number of lines, L, and C) to resemble those of a given network. This step 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(a)), (ii) it is inefficient and unsafe for the power grids to include very long lines (e.g., Fig. 3(b)), and (iii) nodes in denser areas are more likely to have higher degrees (see Fig. 5). To compute the density around node i, we define ρ_i as the average Euclidean distance of node i from its k nearest nodes. We select $k = 10$ in this work.

Hence, the NIMBLE repeats the following steps until the number of lines is equal to the number of lines in the given network: (1) select a low degree node in a dense area (observations (i) and (iii)), and (2) connect it to a high degree node which is also nearby (observations (i) and (ii)). *To select a low degree node in a dense area*, the NIMBLE samples a node i from all the degree 1 and 2 nodes with probability $\propto \rho_i^{-\eta}$, where η is a tunable parameter. *To connect the sampled node to a high degree but nearby node*, the NIMBLE connects node i to node j sampled from all other nodes with probability



(a) (b) (c)

Fig. 7: Power flows in the WI and the generated networks. **(a)** The CDF of the flows on the lines. **(b)** The CDF of the top 1,000 flows on the lines. **(c)** The geographical locations of the lines carrying the top 100 flows.

$\propto \|\mathbf{p}_i - \mathbf{p}_j\|^{-\beta} d_j^{\gamma}$, where $\mathbf{p}_i$ is the (x, y) coordinate of the node i, and β, γ are tunable parameters. 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 (as in the preferential attachment model [22], however distance was not considered there). For the generated networks in this paper, we empirically select $\eta = 0.5$, $\beta = -2.5$, and $\gamma = 1.5$.

A. Running Time

In this subsection, we briefly compute the running time of the NIMBLE. The first two steps of the NIMBLE take $O(|V|)$ time in total. The third step requires $O(|V|^2)$ time since for each node, it needs to find its closest node with a higher index which requires $0 + 1 + \cdots + |V| - 1 = O(|V|^2)$ time. The final step requires $O(|E| - |V| + 1)$ time to add $|E| - |V| + 1$ random additional lines to the grid. So the total running time of the method is $O(|V|^2 + |E|)$. Notice that the running time is without considering the GMM fitting step, since this step can be considered as a preprocessing step and only needs to be calculated once. Generating a synthetic grid based on the WI in our system with Intel Core i7-2600 @3.40GHz CPU and 16GB RAM, takes less than a minute.

V. Topological Evaluation of the Generated Networks

We demonstrate the performance of the NIMBLE in generating synthetic grids similar to the WI power grid. Since our method is probabilistic, we generate three networks for evaluation purposes to show the consistency of the generated networks' properties. We refer to the generated networks as: *GenNet (1), GenNet (2),* and *GenNet (3)*. The generated networks visually resemble the WI very well (see Fig. 6).[1]

Fig. 3 compares the degree distributions of the nodes and length distributions of the lines in the WI and the generated networks. As can be seen in Fig. 3(a), the tail of the degree distribution in the WI follows a power-law distribution. However, following the work by Clauset, et al. [57] and since these networks are finite, there is not 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. Table I provides a summary of the topological properties of the WI and the generated networks. The provided results demonstrate that the generated networks resemble the topological properties of the WI very well. However, the number of line intersections ($\mathcal{X}$) is about 50% more in the generated networks.

[1]A sample generated network is also publicly available at [38] and [39].

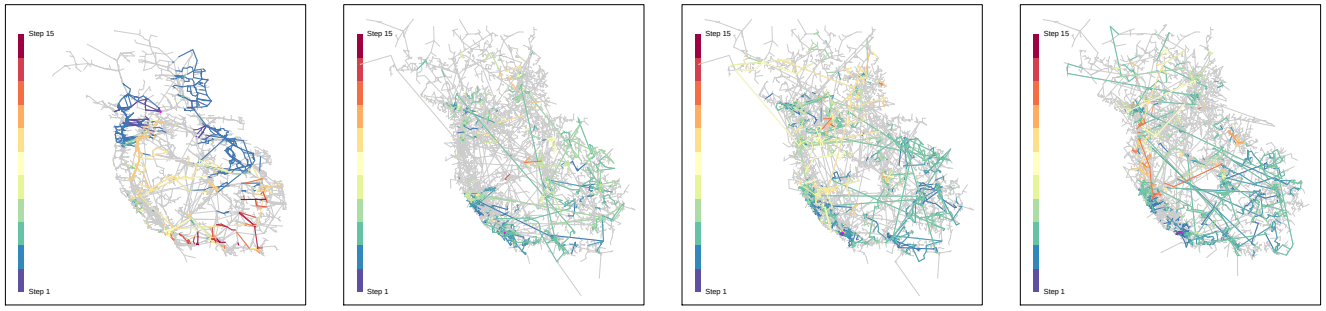
TABLE II: Statistics of the flows (MW). The *backup lines* are the lines that do not initially carry any significant flow.

Networks	Mean	Median	SD	Max	Backup lines
WI	282.37	98.54	488.79	13,111.68	3,558
GenNet (1)	168.93	33.32	320.93	4,777.07	3,419
GenNet (2)	153.78	29.53	295.83	6,171.67	3,464
GenNet (3)	172.27	34.71	321.12	8,867.31	3,289

VI. Robustness Evaluation of the Generated Networks

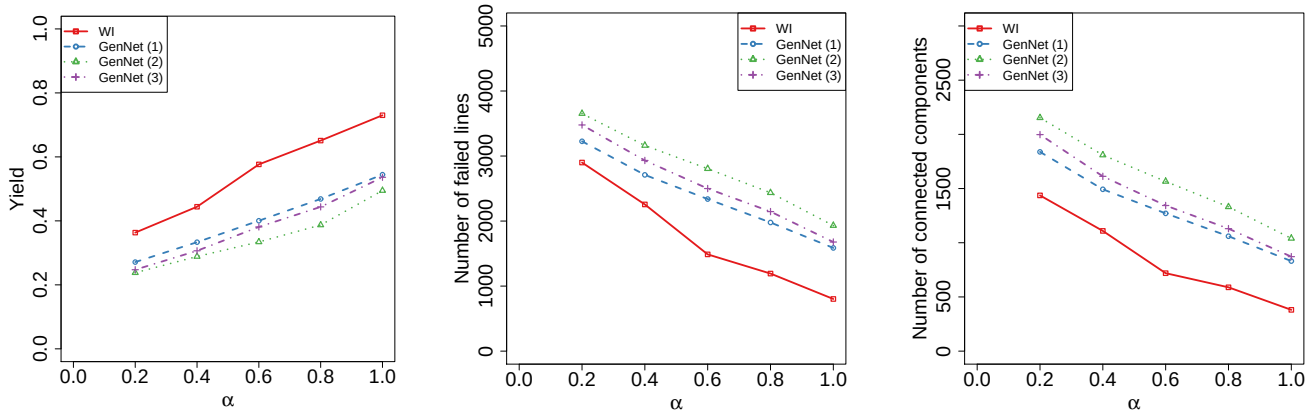
We use the DC power flow equations and compare the flow distribution as well as the robustness against cascading failures in the WI and the generated networks. The power demands and supplies are estimated based on the population of the cities and power plants' capacities as described in Subsection III-A. The flow statistics are very similar in the generated networks and the WI (see Table II). Despite the close similarity in the flow CDFs between the WI and the generated networks, the locations of the lines that carry the top 100 flows are uncorrelated in all the four networks (see Fig. 7). This demonstrates that NIMBLE generates networks that not only have very similar flow characteristics to the real network, but also do not reveal the locations of the potentially vulnerable lines in the real network.

To evaluate the robustness of the generated networks compared to the WI, we study cascading failures in these networks. As an example, the evolution of the cascade in the WI and the generated networks is depicted in Fig. 8. We consider cascades initiated by double line failures (see Fig. 9) and removal of all the nodes and lines in a circular area (see Figs. 10 and 11 and also Figs. 12 and 13).



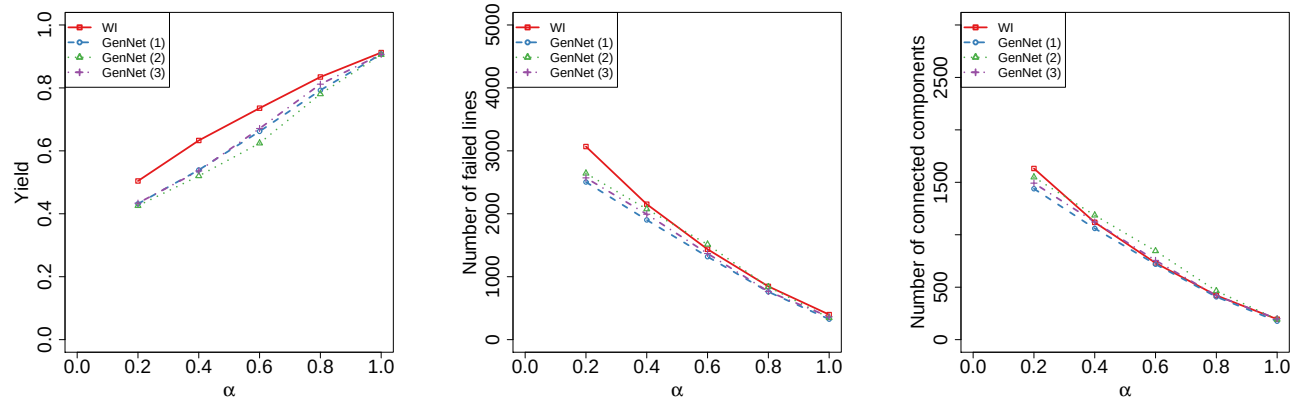
(a) WI (b) GenNet (1) (c) GenNet (2) (d) GenNet (3)

Fig. 8: The first 15 steps of the cascades in the WI and the three generated networks. In all the networks $\alpha = 0.2$ and the cascade is initiated by a failure in the line with the highest flow, indicated with a magenta colored point. **(a)** Evolution of the cascade in the WI, which continues over 25 steps, ending with the yield = .338, number of failed lines = 3,041, and number of connected components = 1,558. **(b)** Evolution of the cascade in the GenNet (1), which continues over 15 steps, ending with the yield = .357, number of failed lines = 2,621, and number of connected components = 1,366. **(c)** Evolution of the cascade in the GenNet (2), which continues over 15 steps, ending with the yield = .257, number of failed lines = 3,573, and number of connected components = 1,946. **(d)** Evolution of the cascade in the GenNet (3), which continues over 15 steps, ending with the yield = .277, number of failed lines = 3,434, and number of connected components = 1,889.



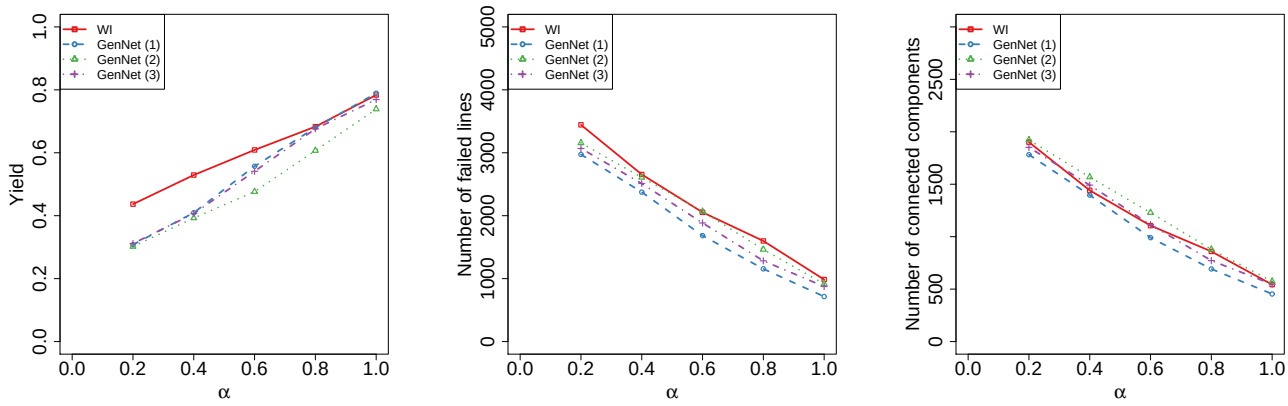
(a) (b) (c)

Fig. 9: The severity of the cascades initiated by all possible double line failures selected from the lines that carry the top 25 largest flows in the networks as a function of the lines' factor of safety (α). **(b)** The total number of failed lines at the end of the cascade. **(c)** Number of connected components at the end of the cascade. As the cascade proceeds, the networks may get partitioned into several parts.



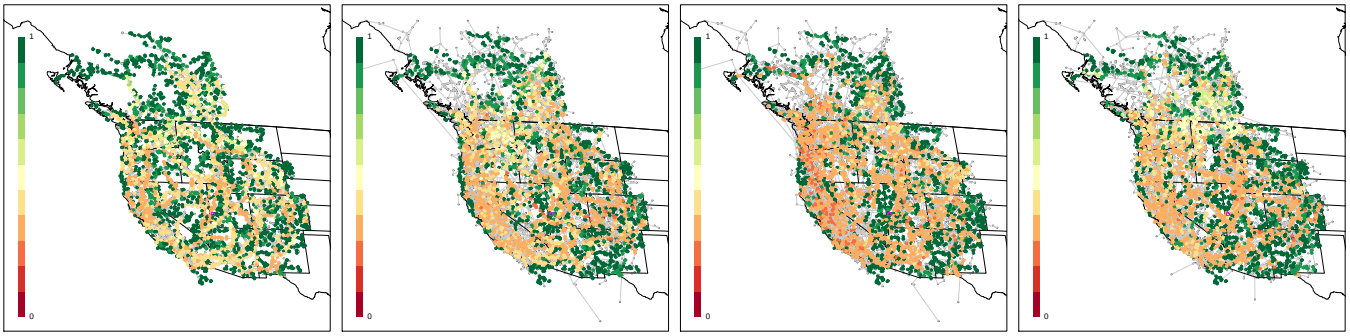
(a) (b) (c)

Fig. 10: The average severity of the cascades initiated by failures in 10,000 uniformly distributed regions of radius 20 *km* as a function of lines' factor of safety (α). All the nodes and lines are removed from the initial failed area. **(a)** Yield. **(b)** The total number of failed lines at the end of the cascade. **(c)** Number of connected components at the end of the cascade.



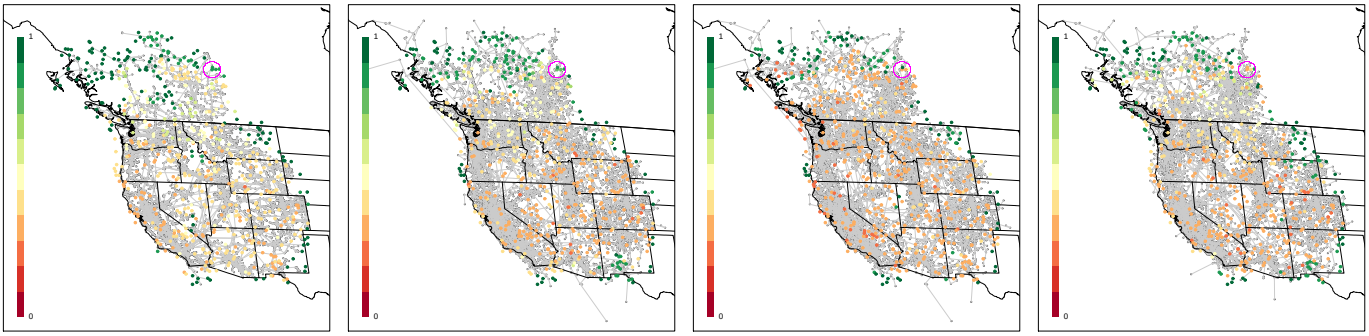
(a) (b) (c)

Fig. 11: The average severity of the cascades initiated by failures in 1,000 uniformly distributed regions of radius 100 *km* as a function of lines' factor of safety (α). All the nodes and lines are removed from the initial failed area. The locations of failed areas are exactly the same in all the four networks. **(a)** Yield. **(b)** The total number of failed lines at the end of the cascade. **(c)** Number of connected components at the end of the cascade.



(a) WI (b) GenNet (1) (c) GenNet (2) (d) GenNet (3)

Fig. 12: Yield of the cascades initiated by failures in 10,000 regions of radius 20 *km* uniformly distributed when $\alpha = 0.6$. The color of each point represents the yield of the cascade initiated by the failure in the area centered at that point. The size of the failed area in the map scale is shown by a magenta circle.



(a) WI (b) GenNet (1) (c) GenNet (2) (d) GenNet (3)

Fig. 13: Yield of the cascades initiated by failures in 1,000 regions of radius 100 *km* uniformly distributed when $\alpha = 0.6$. The color of each point represents the yield of the cascade initiated by the failure in the area centered at that point. The size of the failed area in the map scale is shown by a magenta circle.

In Fig. 9, we observe that the generated networks and the WI have similar level of robustness against failures in the lines that carry the highest flows (notice that as we mentioned in Fig. 7 these lines are in very different locations in the four networks). Figs. 10 and 11 also suggest that if a big area in the networks fails, the WI and the generated networks on *average* behave very similarly. Figs. 12 and 13 show the heatmaps of the yield for cascades initiated by area failures in the WI and the generated networks. As can be seen, despite having a very similar level of robustness, the generated networks do not reveal the vulnerabilities of the WI since the vulnerable areas in the generated networks and the WI are not in one to one correspondence.

VII. Conclusion

In this paper, we studied the structural and spatial properties of the WI and developed the NIMBLE for generating synthetic power grid networks. We showed that the generated networks have similar structural properties as well as the same level of robustness against failures to the WI. Since the generated grids are embedded on approximately the same coordinates as the real grid, they can provide realistic test networks to the researchers.

The NIMBLE is designed mostly to replicate the *average* structural properties of the real grid. Hence, the generated networks may not reflect the severity of a single event in the actual grid and should be used for inferring *average robustness* of the actual grid instead.

There are several engineering aspects of power grid networks such as line nominal voltages, reactive power demands, voltage and frequency stabilities, and transformer characteristics that were not considered in this work. In a recent paper [18], statistical properties of some of the related engineering parameters were studied in details. We plan to incorporate these findings in our future work to improve the quality of the generated networks to the level of industrial grade data sets. Moreover, we believe that if industrial grade data sets become available to the broad research community, the methods developed in this paper can be used in order to develop synthetic grids based on them.

We also believe that our approach can be extended for generating various other types of spatially distributed networks (e.g., water and gas pipe networks). It is part of our future work to explore these directions.

Acknowledgements

This work was supported in part by the U.S. DOE OE as part of the DOE Grid Modernization Initiative, DARPA RADICS under contract #FA-8750-16-C-0054, and DTRA grant HDTRA1-13-1-0021.

References

[1] S. Soltan and G. Zussman, "Generation of synthetic spatially embedded power grid networks," in *Proc. IEEE PES'16*, July 2016.

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

[3] 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.

[4] 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.

[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-GM'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, "Analysis of failures in power grids," *to appear in IEEE Trans. Control Netw. Syst.*, 2017.

[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/ (Date of access: 03/07/2017).

[12] "National Grid UK," available at http://www2.nationalgrid.com/uk/services/land-and-development/planning-authority/shape-files/ (Date of access: 03/07/2017).

[13] "Polish grid," available at http://www.pserc.cornell.edu/matpower/ (Date of access: 03/07/2017).

[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] K. M. Gegner, A. B. Birchfield, T. Xu, K. S. Shetye, and T. J. Overbye, "A methodology for the creation of geographically realistic synthetic power flow models," in *Proc. PECI'16*, Feb. 2016.

[17] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *to appear in IEEE Trans. Power Syst.*, 2017.

[18] A. B. Birchfield, K. M. Gegner, T. Xu, K. S. Shetye, and T. J. Overbye, "Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies," *to appear in IEEE Trans. Power Syst.*, 2017.

[19] 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.energy.gov/sites/default/files/documents/files/GRIDDATA_ProgramOverview.pdf (Date of access: 03/07/2017).

[20] Platts, "GIS Data," http://www.platts.com/Products/gisdata (Date of access: 03/07/2017).

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

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] I. Dobson, *Encyclopedia of Systems and Control*, ser. Cascading network failure in power grid blackouts. Springer, 2015.

[29] P. Hines, K. Balasubramaniam, and E. C. Sanchez, "Cascading failures in power grids," *IEEE Potentials*, vol. 28, no. 5, pp. 24–30, 2009.

[30] R. Baldick, B. Chowdhury, I. Dobson, Z. Dong, B. Gou, D. Hawkins, H. Huang, M. Joung, D. Kirschen, F. Li *et al.*, "Initial review of methods for cascading failure analysis in electric power transmission systems IEEE PES CAMS task force on understanding, prediction, mitigation

and restoration of cascading failures," in *Proc. IEEE PES-GM'08*, July 2008.
[31] D. Bienstock, *Electrical Transmission System Cascades and Vulnerability: An Operations Research Viewpoint*. SIAM, 2015.
[32] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.
[33] H. Xiao and E. M. Yeh, "Cascading link failure in the power grid: A percolation-based analysis," in *Proc. IEEE Int. Work. on Smart Grid Commun.*, June 2011.
[34] J. Zhao, D. Li, H. Sanhedrai, R. Cohen, and S. Havlin, "Spatio-temporal propagation of cascading overload failures in spatially embedded networks," *Nat. Commun.*, vol. 7, 2016.
[35] I. Dobson, "Electricity grid: When the lights go out," *Nature Energy*, vol. 1, p. 16059, 2016.
[36] L. Daqing, J. Yinan, K. Rui, and S. Havlin, "Spatial correlation analysis of cascading failures: congestions and blackouts," *Scientific reports*, vol. 4, 2014.
[37] A. E. Motter and Y.-C. Lai, "Cascade-based attacks on complex networks," *Phys. Rev. E*, vol. 66, no. 6, p. 065102, 2002.
[38] "Columbia synthetic power grid data set," avilable at: http://wimnet.ee.columbia.edu/portfolio/synthetic-power-grids-data-sets/.
[39] "Columbia University synthetic power grid with geographical coordinates," avilable at: https://egriddata.org/dataset/columbia-university-synthetic-power-grid-geographical-coordinates.
[40] 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.
[41] 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.
[42] 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.
[43] M. M. Danziger, L. M. Shekhtman, Y. Berezin, and S. Havlin, "Two distinct transitions in spatially embedded multiplex networks," *arXiv:1505.01688*, 2015.
[44] 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.
[45] 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.
[46] M. Barthélemy, "Spatial networks," *Physics Reports*, vol. 499, no. 1, pp. 1–101, 2011.
[47] S. S. Manna and P. Sen, "Modulated scale-free network in Euclidean space," *Phys. Rev. E*, vol. 66, no. 6, p. 066114, 2002.
[48] R. Xulvi-Brunet and I. M. Sokolov, "Evolving networks with disadvantaged long-range connections," *Phys. Rev. E*, vol. 66, no. 2, p. 026118, 2002.
[49] "U.S. Energy Information Administration," available at http://www.eia.gov/ (Date of access: 03/07/2017).
[50] W. H. Press, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.
[51] A. R. Bergen and V. Vittal, *Power Systems Analysis*. Prentice-Hall, 1999.
[52] D. Bienstock and A. Verma, "The $N - k$ problem in power grids: New models, formulations, and numerical experiments," *SIAM J. Optimiz.*, vol. 20, no. 5, pp. 2352–2380, 2010.
[53] R Core Team, *R: A Language and Environment for Statistical Computing*, 2014, available at https://www.R-project.org (Date of access: 03/07/2017).
[54] 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.
[55] 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.
[56] S. Soltan and G. Zussman, "Generation of synthetic spatially embedded power grid networks," *arXiv:1508.04447*, Aug. 2015.
[57] 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.

Saleh Soltan is a postdoctoral research associate in the department of Electrical Engineering at Princeton University. In 2017, he obtained the Ph.D. degree in Electrical Engineering from Columbia University. He received B.S. degrees in Electrical Engineering and Mathematics (double major) from Sharif University of Technology, Iran in 2011 and the M.S. degree in Electrical Engineering from Columbia University in 2012. He is the Gold Medalist of the 23rd National Mathematics Olympiad in Iran in 2005 and the recipient of Columbia University Electrical Engineering Armstrong Memorial Award in 2012.

Alexander Loh is currently a PhD student in the department of Electrical Engineering at Columbia University. He received a B.S. in Electrical Engineering from Rutgers University in 2015, and an M.S. in Electrical Engineering from Columbia University in 2016. His research interests include Algorithms, Stochastic Modeling, Network Science, Computer Networks, and Power Systems Analysis. He is the recipient of the Columbia University Presidential Fellowship (2015), NSF IGERT Fellowship (2015), and the NSF GRFP Honorable Mention (2017).

Gil Zussman is an Associate Professor of Electrical Engineering at Columbia University. He received the Ph.D. degree in electrical engineering from the Technion in 2004 and was a postdoctoral associate at MIT in 2004-2007. He is a co-recipient of 7 paper awards including the ACM SIGMETRICS06 Best Paper Award, the 2011 IEEE Communications Society Award for Advances in Communication, and the ACM CoNEXT'16 Best Paper Award. He received the Fulbright Fellowship, the DTRA Young Investigator Award, and the NSF CAREER Award, and was a member of a team that won first place in the 2009 Vodafone Foundation Wireless Innovation Project competition.