Quantification of Area Sparsity in Large-Scale Electric Grids

Sanjana Kunkolienkar, Farnaz Safdarian, Jonathan Snodgrass, Thomas Overbye
Department of Electrical and Computer Engineering Texas A&M University
College Station, TX
{sanjanakunkolienkar, fsafdarian, snodgrass, overbye}@tamu.edu

***Abstract*—This paper quantifies the likelihood of two substations being connected based on the topology to improve and build more realistic synthetic grids with evaluating the idea that in North American power grids, two substations are more likely to be connected if they belong to the same area than to different areas. Statistical methods are used to identify and visualize the topological differences between real and synthetic grids and how areas to which a substation belongs influence how likely the substations are to be connected. This paper defines a new term, 'Area Sparsity,' to quantify the relationship between substation connectedness and highlights the need to explicitly incorporate the power grid areas into creating more realistic synthetic grids. The results show that the actual grids are more connected in the same area; however, this is not the case for the existing large-scale synthetic grids.**

I. Introduction

Historically, electrification was highly regional, with independent power companies owning separate local grids. As the benefits of linking these mostly isolated local networks became evident, the stakeholders worked towards making a larger interconnected power grid a reality [1]. Today, even though the North American electric grid is highly interconnected, it is not owned by a single entity. Hundreds of entities have a stake in it and are responsible for its operation and maintenance, such as different independent system operators (ISOs), regional transmission organizations (RTOs), electric cooperatives (co-ops), and utilities.

While interconnected grids are a single electric circuit, they are often divided into operating areas or zones that traditionally correspond to a particular electric utility. The size and boundaries of these areas are not uniform because of geopolitical reasons, management issues, and technical limitations.

Each operational area is a more densely connected local sub-network that is also interconnected to other areas by transmission lines to form a more extensive grid network. Examining the grid, one could ask how likely are two given substations to be connected by a transmission line. Does this likelihood change when the substations belong to different areas, and if so, how to quantify this? This paper quantifies the relationship between the likelihood of two substations being connected, the areas to which they belong, and the distance between them. In doing so, this paper defines a new term, 'Area Sparsity,' which measures the connectedness of two substations based on their area affiliation.

Understanding the topological properties of a power grid is pertinent to creating robust power networks [2]. Interesting findings from the application of graph theory to the power grid are shown in [3], [4]. Important network characteristics are highlighted by treating the power grid as a complex network [5], [6]. While these works extract topological properties of the North American grid using statistical and graph measures like node degree distribution, characteristic path length, node clustering coefficient, and betweenness centrality, the impact of areas on actual grid topology is not discussed and ultimately not considered in the formation of synthetic grids. However, it is worth noting that [7] recognizes that the average node degree does not scale with an increase in network size and is area dependent.

Since the North American grid data in the United States is considered critical energy/electric infrastructure information (CEII), [8]–[11] introduced the creation of a fictitious but realistic power system model based on census data and U.S. Energy Information Association (EIA) generation data [12], which span on the actual geographic footprints of the United States. These large-scale grids are used to perform analysis and develop algorithms to be implemented in the planning or operation of actual power grids. Additionally, [13]–[15] validated synthetic grids using metrics extracted from actual North American electric grid models to ensure the realism and usefulness of the synthetic grids for studies and research to develop new algorithms and applications. These metrics follow the structure, proportions, and parameters of key power system elements, which can be used in assessing and validating the quality of synthetic power grids. However, these metrics do not evaluate the nature of North American power grid areas and the likelihood of a transmission line connecting two substations; hence, their impact is not considered when creating synthetic grids.

This paper tests the hypothesis that in North American power grids two substations are more likely to be connected if they belong to the same area than to different areas. Once this is demonstrated, the paper then provides a way to quantify this relationship so that it can be used to generate more realistic synthetic grids. The paper is divided as follows: Section II discusses the synthetic and real-world grids considered for the analysis. Two indices are defined in Section III. Section

IV analyses the results using box plots and quantifies the relationship between the likelihood of two substations being connected and whether they belong to the same area. Finally, Section V discusses potential future work.

II. Studied Synthetic Grids and North American Grids

Since the paper focuses on understanding how areas in the North American grid dictate the likelihood of two substations being connected, extracting information from these North American grids is essential. This real-world power grid is divided into three interconnections – the Western Electricity Coordinating Council (WECC), Eastern Interconnect (EI), and the Electric Reliability Council of Texas (ERCOT). The recent WECC power flow models have about 25,000 buses divided into 25 areas, ranging in size from less than 100 buses to more than 4000. The EI model has more than 90,000 buses divided into 135 areas, with sizes ranging from 1 bus up to 5000. Some of these areas represent an aggregation of multiple utilities. These grids are studied, and the WECC and the EI system are investigated more closely.

Overall, the paper aims to quantify the likelihood of two substations being connected based on the topology to improve and build more realistic synthetic grids. Hence, it is required to show that the current large-scale synthetic grids do not incorporate this relationship. Three synthetic test grids are used. These are the 2000-bus (2K) case covering Texas, the 10,000-bus (10K) case covering the WECC footprint, and the 70,000-bus (70K) grid over the EI footprint in the US. The creation and validation of these test grids are discussed in [8]–[10]. These grids are selected since they are large-scale electric grids with topological and electrical characteristics of North American power grids.

III. Methodology

It is necessary to establish two indices to quantify the relationship in question. In this paper and future works, a term is needed to define how likely two substations are to be connected if they are *x*-miles apart. A new term, based on an existing term, is made up called "Area Sparsity" and is calculated as shown in Eq. (2). In graph theory, network density is defined as the ratio between existing edges to the maximum possible number of edges [16]. This term is modified into Area Sparsity to make the definition relevant to a power grid in Eq. (1), where edges are equivalent to transmission lines between two substations while substations act as nodes.

$$Area\ Sparsity = \frac{N_{connected}}{N_{total}} \tag{1}$$

where $N_{connected}$ = Number of substations *connected* by a transmission line and *x*-miles apart. N_{total} = Total number of substations that are x-miles apart. N_{total} represents the total number of pairs of substations that are *x*-miles apart, regardless of whether they are connected or not connected.

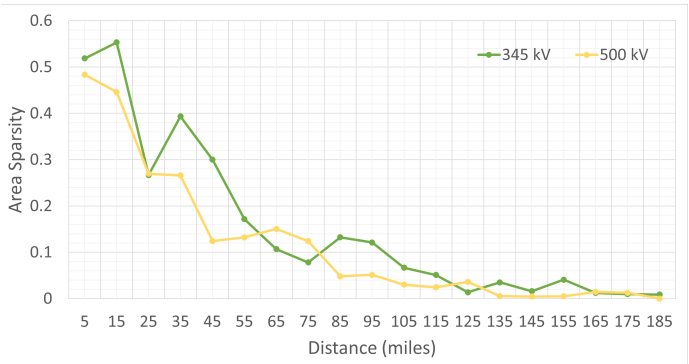

Fig. 1. Area Sparsity for WECC for the 345kV and 500kV network.

For example, for the 345kV transmission network of the WECC case, first, $N_{connected}$ is calculated as the number of substation pairs that are 10 miles apart and have a transmission line connecting them to each other. Then N_{total} is calculated as the total possible number of substation pairs that are 10 miles apart. N_{total}, contains all substation pairs satisfying the distance category regardless of whether they have a transmission line between them. Thus, for the 345kv network in WECC, for a 10-mile distance, $N_{connected}$ = 14 and N_{total} = 27. Thus Area Sparsity for the 10-mile bin is 0.51. This can be seen in Fig. 1.

In this paper, *x* is chosen as 10 miles. A lower bin size may introduce lower values of Area Sparsity, skewing the results to show lower connectedness. Whereas, in reality, there are not enough substations that are less than 5 miles apart and connected. The idea here is not to see if Area Sparsity or connectedness reduces with distance but to observe the impact of the area on the likelihood of connectedness. On the other side, a higher bin size may cause a loss of points, also leading to a misrepresentation of the trend.

As expected, the likelihood of two substations being connected decreases as the distance between them increases, as shown in Fig. 1. It is important to note that this plot is not a probability density function. At first glance, it can be identified that the likelihood of two substations being connected reduces as the distance between them increases. This idea has been identified by Watts and Strogatz [17], where power grids are categorized as small-world networks. However, networks built by assuming power grids as small-world networks are not realistic, as explained in [4].

Hence, we define Area Sparsity by modifying Eq. (1) to include whether the substations are located in the same or different areas. Here, the area number or name to which a substation belongs is unimportant as long as it is investigated whether the two substations are in the same area or not. Thus, depending on whether the substations belong to the same area or a different area, the equation for Area Sparsity is modified as follows:

$$Intra - Area\ Sparsity = \frac{N_{connected\ and\ in\ same\ area}}{N_{total\ in\ the\ same\ area}} \tag{2}$$

$$Inter - Area\ Sparsity = \frac{N_{connected\ and\ in\ diff.\ area}}{N_{total\ in\ diff.\ area}} \quad (3)$$

The procedure for plotting Intra-Area and Inter-Area Sparsity for a system is to take all the substations in one voltage network, with their latitude and longitude information, and calculate a geographical distance matrix between them. Then, for a total distance of x-miles, calculate the number of substations in each category: $N_{connected}$; N_{total}; $N_{connected\ and\ in\ same\ area}$; $N_{connected\ and\ in\ different\ area}$; $N_{total\ in\ the\ same\ area}$ and $N_{total\ in\ different\ area}$. The distance is increased by x-miles, and the calculation is repeated.

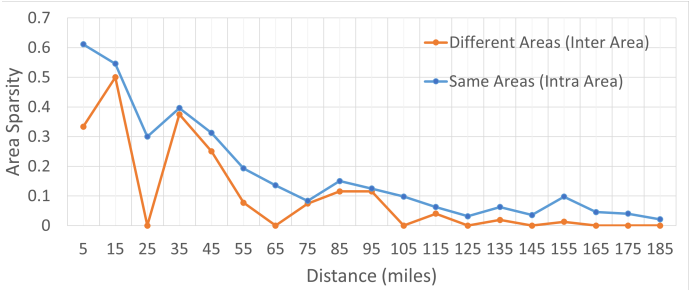

Fig. 2. Area Sparsity for the 345kV network of the WECC Grid

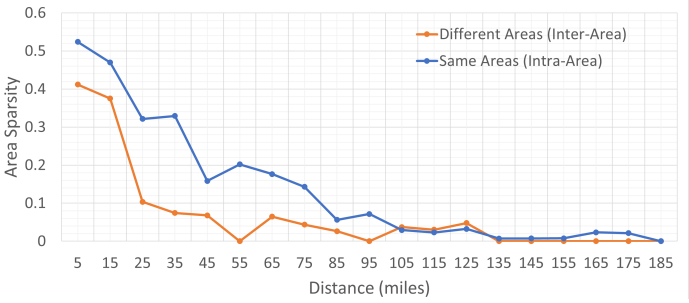

Fig. 3. Area Sparsity for the 500kV network of the WECC Grid

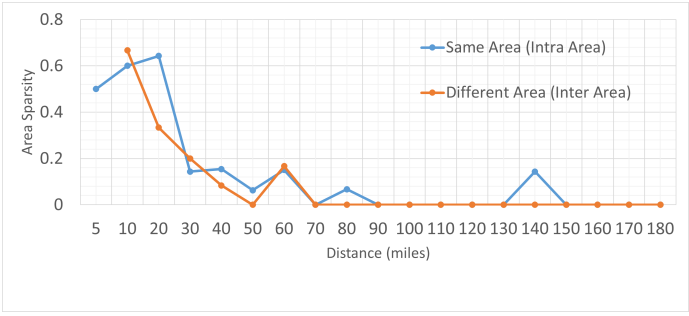

Fig. 4. Area Sparsity for the 230kV network of the EI grid

Using Eq. (2), and (3), the indices are calculated and plotted for each transmission level voltage network. The WECC cases have multiple transmission level voltage networks, out of which the plot for 345kV and 500kV are shown in Fig. 2 and in Fig. 3. Fig. 4 is the line plot for Inter and Intra Area Sparsity for the 230kV network of the EI case.

The same technique and equations are used to calculate the Area Sparsity for the synthetic grids. In contrast to Fig. 2, it can be seen that in Fig. 5, the plot looks a lot like a power law. It seems like the Area Sparsity, i.e., the likelihood of connectedness, depends mainly on the distance between the two substations: As the distance between the two substations increases, the Area Sparsity decreases. The 230kV network plot in Fig. 6 is peculiar, given that the Area Sparsity is higher for substations that belong to two different areas than if they were in the same area. This means for the 230kV transmission network of the 24k synthetic grid, if two substations belong to two different areas and are 10 miles apart, then they are more likely to have a transmission line connecting them than if they belonged to the same area.

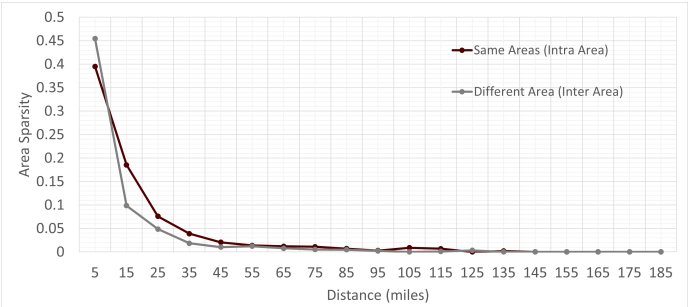

Fig. 5. Area Sparsity for the 345kV network of the synthetic 10k Grid

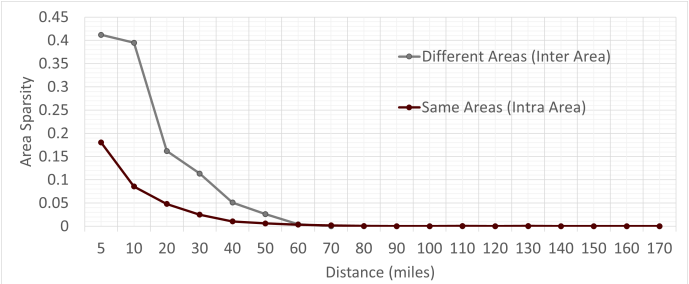

Fig. 6. Area Sparsity for the 230kV network of the synthetic 24k Grid

IV. Results and Analysis

Area Sparsity is calculated for the synthetic and North American WECC grids using Eq. 2 and 3 defined in section III. For the WECC and EI networks, from Fig. 2 and 3, it is seen that generally, the Area Sparsity for substations belonging to the same area (in blue) is higher than when they belong to different areas (in orange). That is, Intra-Area Sparsity is higher than Inter-Area sparsity. These differences are illustrated using box plots.

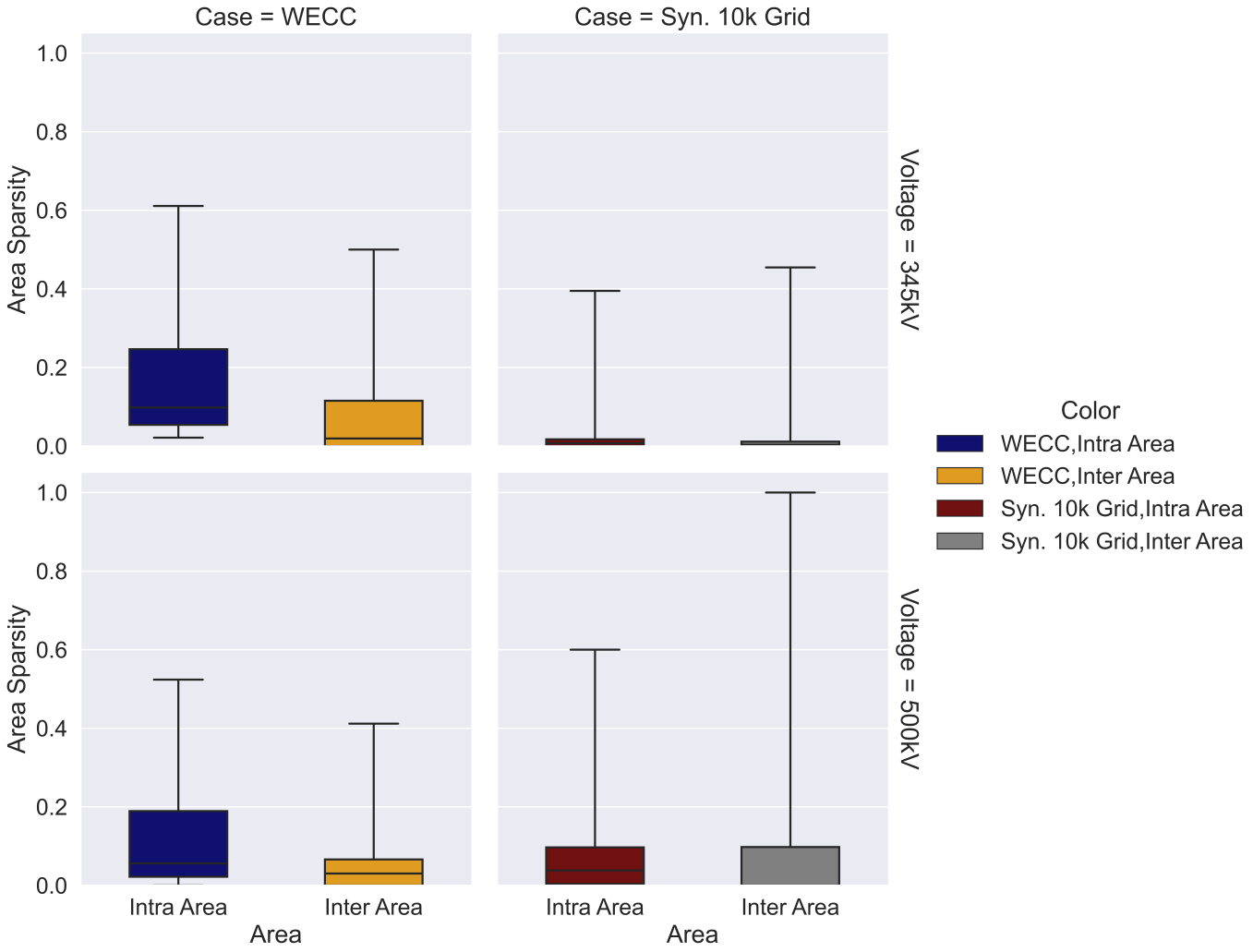

Fig. 7. Intra and Inter-Area Sparsity for the actual WECC grid and the synthetic 10k grid.

Using the Area Sparsity equations (2 and 3), line plots for all voltage networks can be computed. However, only a visual representation of the Area Sparsity of these networks is not enough. A tangible plot is required to identify notable differences between the Area Sparsity of actual North American grids and the large-scale synthetic grids. Thus, in this paper, box plots are utilized to visualize these differences. Box plots are a tool to display a dataset's statistical properties, mainly the dataset's dispersion [18]. The lower and upper ends protruding from the box are called whiskers and show the maximum and the minimum, respectively. Fig. 7 shows the corresponding box plots of Area Sparsity for the 345kV and 500kV voltage network of WECC. Similarly, Fig. 8 shows the corresponding box plots of Area Sparsity for the 230kV and 345kV voltage network of EI case. The x-axis indicates whether the substations belong to the same or different areas.

From preliminary observation of Fig. 7, it can be seen that in the actual WECC case, the Intra Area Sparsity is higher than Inter Area Sparsity. The result is consistent for the 500 kV network. On the other hand, for the 10k synthetic grid, the substations that belong to different areas are more likely to be connected than those in the same area. This is because the synthetic grid creation algorithms emphasize the geographic proximity of buses deciding whether two substations should be connected. Note the difference in the size of the box plots across different voltage networks. A reason for this difference is that the number of transmission lines and substations differs across the voltage networks. As the box plots are negatively skewed, overall, there is lower Inter-Area Sparsity for synthetic grids than the real-world grids. These plots aim to show the difference in Area Sparsity for substations in the same versus different areas.

A glaring point of note from both Fig. 7 and Fig. 8 is that Inter and Intra Area Sparsity is inconsistent for synthetic grids. It essentially implies that area has not been an explicit factor in determining the addition of transmission lines when creating synthetic grids. Thus, to make synthetic grids have realistic properties of the actual power grids, there is a need to consider

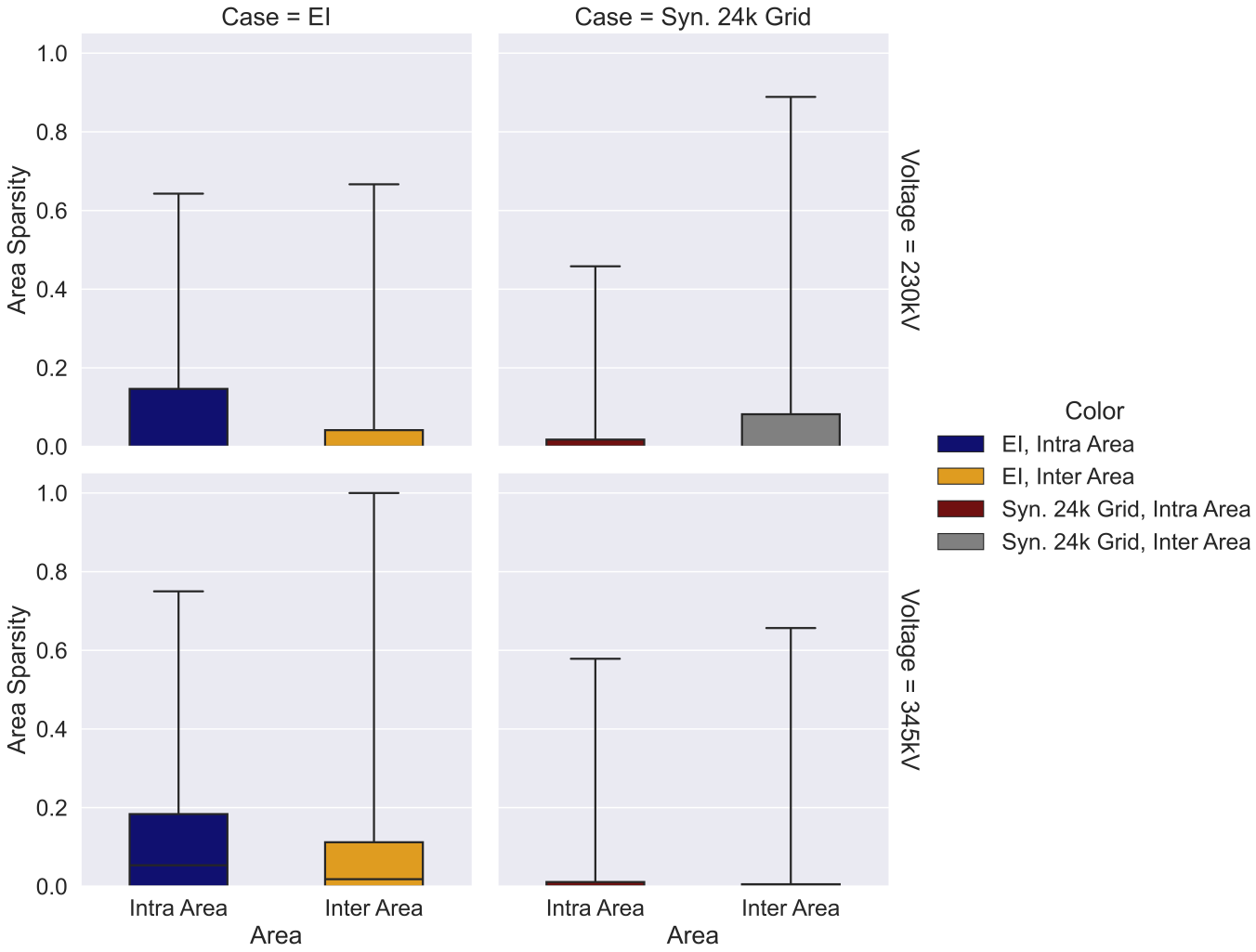

Fig. 8. Intra and Inter-Area Sparsity for the actual EI grid and the synthetic 24k grid.

areas when deciding which substations should be connected by a transmission line.

Now that it has been established that the location of substations with respect to areas is essential to determine whether two substations are connected, the next step is to quantify this relationship. Thus to determine the correlation, multivariate linear regression is performed on the data points plotted by Area Sparsity and Distance as shown in Fig. 9.

TABLE I
COEFFICIENTS OF MULTIVARIATE LINEAR REGRESSION

	WECC
Distance	-0.0023
Area	0.075

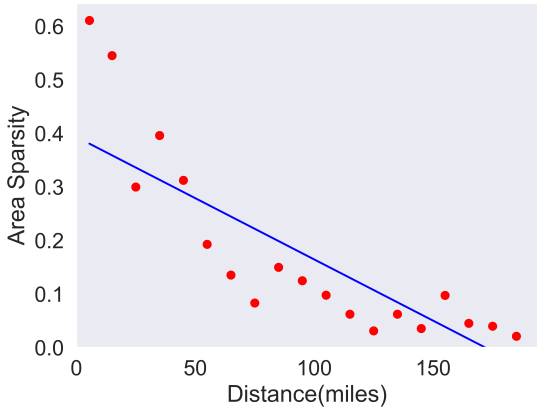

Fig. 9. Multivariate Linear Regression fit for the 345 kV Network(WECC)

There are two independent variables when setting up the multivariate linear regression problem. First is the Distance variable, which is a numeric variable. Second, is the Area variable, which is a Boolean value where 1 represents that the substations are intra-area and 0 represents that the substations are inter-area. After linear regression, the coefficients of both independent variables are shown in Tables I. This clearly shows that the Area variable has a higher weight (coefficient)

in determining whether there will be a connection between two substations than the Distance variable.

As observed from Tables II and III, the results of multivariate linear regression for the 345kV and 230kV transmission networks show that Area Sparsity carries a higher weight than the distance in deciding whether two substations are connected. Similar results are obtained for the 500kV transmission network of the WECC and 10k synthetic grid case.

As expected, the coefficient of the Distance variable is consistent across the actual power grid cases and their synthetic counterparts. This coefficient is negative, correctly signifying the negative correlation between the likelihood of two substations being connected as the distance between them increases. For the synthetic cases, there are multiple inaccuracies for the Area variable. In the 345 kV networks, the Area variables' negative signs agree with the real grids but are too close in magnitude to the Distance variable. For the 230kV network, the relative magnitudes of the Distance and Area variables agree for the synthetic grids, but the signs are reversed compared to the real networks.

TABLE II
COEFFICIENTS FOR THE 345KV NETWORK

	WECC	**EI**	**Syn. 10k grid**	**Syn. 24k grid**
Distance	-0.0023	-0.0020	-0.0013	-0.0014
Area	0.075	0.317	0.007	0.0009

TABLE III
COEFFICIENTS FOR THE 230KV NETWORK

	WECC	**EI**	**Syn. 10k grid**	**Syn. 24k grid**
Distance	-0.00048	-0.0022	-0.002	-0.0019
Area	0.0162	0.0532	-0.0519	-0.0648

V. SUMMARY AND FUTURE WORK

This paper provides evidence to evaluate the idea that substations that belong to the same area exhibit a higher probability of being connected to each other than the substations that belong to two different areas. Its key contributions are as follows. First, a new term called 'Area Sparsity' is defined to quantify the likelihood of connectedness of two substations x-miles apart. This term is modified to include whether substations are Inter Area or Intra Area. Then, the topology of the actual North American grids and the synthetic grids are analyzed. Box plots are used as a statistical measure to visualize and identify the differences. Finally, multivariate linear regression is used to quantify these differences into usable weights.

The coefficient of the Distance variable follows the same pattern for both the actual and the synthetic grids. This implies that the current large-scale electric grids are accurate in this regard. The results also prove the hypothesis that the actual grids are more connected within the same area than in different areas. However, this is not the case for the existing large-scale synthetic grids.

In future work, these weights will be used in synthetic grid creation algorithms to create topologically enhanced synthetic grids. The authors believe improved studies can be performed on the new synthetic grids created using the additional area weights and provide valuable insights into the topology and operation of power grids.

VI. ACKNOWLEDGEMENT

This work is partially supported through funding provided by the U.S. National Science Foundation (NSF) in Award 1916142, the U.S. Department of Energy (DOE) under award DE-OE0000895, the US ARPA-E Grant No. DE-AR0001366, and the Power Systems Engineering Research Center (PSERC)

REFERENCES

[1] J. Cohn, "When the grid was the grid: the history of North America's brief coast-to-coast interconnected machine [scanning our past]," *Proceedings of the IEEE*, vol. 107, no. 1, p. 232–243, 2019.

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

[3] G. A. Pagani and M. Aiello, "The Power Grid as a complex network: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688–2700, 2013. [Online]. Available: https://dx.doi.org/10.1016/j.physa.2013.01.023

[4] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, and S. Blumsack, "Comparing the Topological and Electrical Structure of the North American Electric Power Infrastructure," *IEEE Systems Journal*, vol. 6, no. 4, p. 616–626, 2012.

[5] A. B. Birchfield and T. J. Overbye, "Graph crossings in electric transmission grids," in *2021 North American Power Symposium (NAPS)*, Texas A&M University, College Station, TX, 2021, pp. 1–6.

[6] M. H. Mohammadi and K. Saleh, "Synthetic Benchmarks for Power Systems," *IEEE Access*, vol. 9, pp. 162 706–162 730, 2021. [Online]. Available: https://dx.doi.org/10.1109/access.2021.3124477

[7] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating Statistically Correct Random Topologies for Testing Smart Grid Communication and Control Networks," *IEEE Transactions on Smart Grid*, vol. 1, no. 1, pp. 28–39, 2010.

[8] T. Xu, A. B. Birchfield, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Application of large-scale synthetic power system models for energy economic studies," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[9] 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 *2016 IEEE Power and Energy Conference at Illinois (PECI)*. IEEE, 2016, pp. 1–6.

[10] A. B. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid structural characteristics as validation criteria for synthetic networks," *IEEE Transactions on power systems*, vol. 32, no. 4, pp. 3258–3265, 2016.

[11] J. M. Snodgrass, "Tractable Algorithms for Constructing Electric Power Network Models," Ph.D. dissertation, 2021.

[12] (2021) U.S. Energy Information Administration (EIA). [Online]. Available: https://www.eia.gov/electricity/data/eia860/

[13] A. B. Birchfield, E. Schweitzer, M. H. Athari, T. Xu, T. J. Overbye, A. Scaglione, and Z. Wang, "A Metric-based Validation Process to Assess the Realism of Synthetic Power Grids," *Energies*, vol. 10, no. 8, p. 1233, 2017.

[14] A. B. Birchfield, "The Creation, Validation, and Application of Synthetic Power Grids," Ph.D. dissertation, 2018.

[15] V. Krishnan, B. Bugbee, T. Elgindy, C. Mateo, P. Duenas, F. Postigo, J.-S. Lacroix, T. G. San Roman, and B. Palmintier, "Validation of synthetic US electric power distribution system data sets," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4477–4489, 2020.

[16] S. Wasserman, K. Faust, and K. Faust, *Social Network Analysis: Structural Analysis in the Social Sciences*. Cambridge UK: Cambridge University Press, 1994.

[17] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. [Online]. Available: https://dx.doi.org/10.1038/30918

[18] J. W. Tukey *et al.*, *Exploratory Data Analysis*. Addison-Wesley Publishing, 1977, vol. 2.