The Validation of Synthetic Power System Cases

Eran Schweitzer
Arizona State University

Anna Scaglione
Arizona State University

Robert J. Thomas
Cornell University

***Abstract*—Recently, there has been a surge in interest in generating synthetic power systems samples of large size to provide power systems engineering researchers with realistic methods to evaluate their new algorithms and discoveries. Many authors have shown that real network samples show marked trends that are easily discernible. The synthesis of new samples poses important questions: can the trends observed be reproduced? what is the right metric to ascertain the reproduction quality? This motivate this work, whose focus is the validation of synthetic power systems samples.**

***Index Terms*—Synthetic Test Cases, Validation, KL-Divergence**

I. Introduction

As the power system evolves and changes, there is a need for modern test cases that match the complexity, size, and peculiarities of power systems. Synthetic cases such as [1] or [2] offer the benefit of openly sharing model data, which is currently often lacking in power systems research.

There are two main approaches to generating large scale synthetic models. The first, with [1] as an example, applies the wealth of engineering knowledge to the synthesis, to form something akin to a large optimization problem, constructing the samples imposing reasonable constraints and selecting the remaining features optimally based on cost metrics. The second, with [2] as an example, views the system agnostically, and identifies structural, statistical trends, which are then replicated through a randomized generation technique. Both approaches have their shortcomings and strengths. Engineering knowledge can lead to over-optimization, since the real world often forces suboptimal compromises that are very hard to capture. The random approach, on the other hand, struggles to capture rigid structures enforced by engineering practice that end up creating dependencies, which a purely random approach has difficulty capturing in multivariate distributions. Therefore, it runs the risk of creating nonsensical results. It should be further noted that the envisioned uses of these two approaches tend to differ. The engineering-knowledge-optimization approach tends to produce single cases, while the random approach strives to create ensembles that enable statistical test, e.g. Monte Carlo, to be performed.

We view the role of validation in this sense as twofold. First and foremost, it should offer some assurances and confidence to users that the synthetic systems are in fact reasonable. This is done by demonstrating matching trends, which are generally reported as distributions. A secondary purpose though, is to discover signature differences between synthetic and real systems and suggest *corrections* that randomize or constrain and correlate some features to better reflect reality. In other words, validation asks the question where and how are signatures of synthetic girds manifesting themselves, and suggest recourse measures. We delve further into this iterative nature of validation in the second half of the paper.

A. Related Work

There has been quite a bit of literature in the last fifteen years examining the structural nature of the power system as a graph using statistical methods for the analysis and synthesis of grid samples. References such as [2]–[6] all examine the topological characteristics of the power system focusing on metrics like the degree distribution, average path length, clustering coefficients and so forth. Delving further into the power system perspective, [7]–[10] emphasize in various ways that more detailed attention must be paid to the electrical characteristics that underpin the topological ones, essentially arguing that engineering practice must play a greater role in the analysis.

There are several works from the planning literature that examine how synthetic networks can be created [11]–[13]. A recent project focused on dataset synthesis is [1], that in general optimizes various costs, such as minimizing conductor length, while ensuring reliability, over a given geographic footprint.

As will be shown, the synthetic systems start to show certain signatures when the correlation between operation and structure are examined. In [14] and [15] the correlation between the system topology and bus type (load, generation, or connection) is addressed via an entropy type measure. In [16] we made somewhat similar observations about the coupling of a power systems operating point and its topology, by examining properties of the Z_{bus} with respect to the generation vector. These observation are thematically linked to other works in literature that deal with the so-called electrical distance described in [17]—which is essentially the pseudo inverse of the DC powerflow's B matrix. In [18] transmission expansion is explored based on considering the singular value decomposition of the pseudo inverse of the B matrix. Similarly, [7], [19], [20] all describe the electrical distance as a preferable description for power system graphs. In [21], based on simulations of operating conditions, another learned graph is utilized to identify risks for cascading failures.

Before providing some examples of tests, we describe the basic validation methodology in Section II and provide some further intuition on the divergence measures used in Section III. Section IV provides examples comparing real data and the ACTIVSg2k test case project[1], an earlier version of which was also used in [22]. It is worth noting that the ACTIVSg2k model is generated following the first approach we mentioned, of constructing a single case that is geographically embedded in a region and uses power systems engineering guidelines to guide the synthesis. Finally, section V concludes and points at future directions of investigation.

II. Methodology

Validation, is the process by which one confirms, authenticates, or verifies whether something is true. Sometimes it involves a single metric, but more often it is an agreed upon process or set of procedures with recourse. At hand is the question of ensuring that synthetic systems, that claim to be representative of real power systems, are indeed valid representations. That is, if one is presented with a set of data that the owner claims is a valid power system the validation process should be able to say yes or no. We present here a methodology for a systematic evaluation of data sets, as a first step towards the long term validation process objective.

A. KL-Divergence and Hellinger Distance

When analyzing the features of large power-grid samples some clear statistical trends emerge. For a single feature, there is a wide variety of possible distribution to capture the empirical statistics with a law that depends on few tunable parameters. Whenever possible, we wish to fit a statistical law to the particular quantity under study because that naturally allows to *denoise* the empirical statistics. If a good fit is found, also the comparison between cases becomes more rigorous and can be handled by the Kullback-Leibler Divergence [23]:

$$D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \,\mathrm{d}x. \tag{1}$$

The KL-divergence has the following operational meaning. Suppose an observer seeing independent samples x_i wants to decide if they are from distribution $q(x)$ or $p(x)$. If they truly are from $p(x)$ the probability that the observer will reject the hypothesis that they come from $q(x)$ decreases exponentially with the number of samples x_i at a rate equal to $D_{KL}(p\|q)$. That means that if the KL-divergence of the two distributions is small, for many samples x_i from $q(x)$ will appear to be compatible with $p(x)$.

An often cited shortcoming of the KL-divergence is that it is not symmetric an does not obey the triangle inequality, meaning that it is not strictly speaking a "distance." This is not too problematic in the current application since we are always comparing an empirical distribution to a functional fit, i.e., the direction is clear. Nonetheless, the Hellinger [24] distance is used as another divergence measure. The Hellinger distance can be written as,

$$D_H(p||q) = \sqrt{1 - BC(p,q)}, \tag{2}$$

where,

$$BC(p,q) = \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\,\mathrm{d}x \tag{3}$$

is the Bhattacharyya coefficient. This is shown in [25] to obey the triangle inequality and is therefore a valid distance. Interestingly, [26] show that both the KL-divergence and the Hellinger divergence are special cases of the α-divergence family. While this is mainly a nice observation, it further helps support the fact that, as will be shown, the two behave similarly. That is, when the KL-divergence is small so is the Hellinger distance and vice versa.

An added benefit of using the Hellinger distance, is that it relates to the total variation distance, $\delta(p,q)$ as:

$$D_H^2(p,q) \le \delta(p,q) \le \sqrt{2} D_H(p,q). \tag{4}$$

Therefore, the Hellinger distance provides a bound on how far the two distributions could differ from each other at any given point. For all $D_H(p,q) < 1/\sqrt{2}$ this gives a non-trivial upper bound on the most extreme error between an estimate q and the empirical distribution p.

Once a function, $q(x;\theta^*)$, is identified with optimal values for parameters θ^*[2], the following steps are used for validation:

1) $q(x;\theta)$ is fit to the synthetic data
2) Parameter values are compared to ensure they are similar
3) Divergence values are examined to ensure they are sufficiently small.

What constitutes "similar" for parameters or "sufficiently small" for the divergences is somewhat subjective. Instead of focusing on the strict thresholds, the goal of this paper is to illustrate how the methodology is applied. Specific numerical values can always be decided on based on the specific application and may even be modified as more data becomes available and helps illuminate the extent of common ranges.

B. Quantile Function

For certain properties, no clear statistical law can be found, and empirical statistics are more reasonable to use. In these cases, a suitable methodology is to use the cumulative distribution function and its inverse, the quantile function, to establish validation ranges. The validation criteria are formulated as:

- quartile range $[x, y]$ maps to values in the range $[a, b]$,

where some tolerance may be added. For example, we might say that the 40–80% quartiles should map to values in the range of $[0.5, 1.1]$ for some attribute under investigation. This approach has the simultaneous benefit and drawback that the number of ranges identified can be increased or decreased for more or less precision. This adds flexibility, but also allows to blur the differences in trends, masking some differences.

[1]https://electricgrids.engr.tamu.edu/electric-grid-test-cases/activsg2k/

[2]Parameters are found by maximum likelihood or similar fitting approaches.

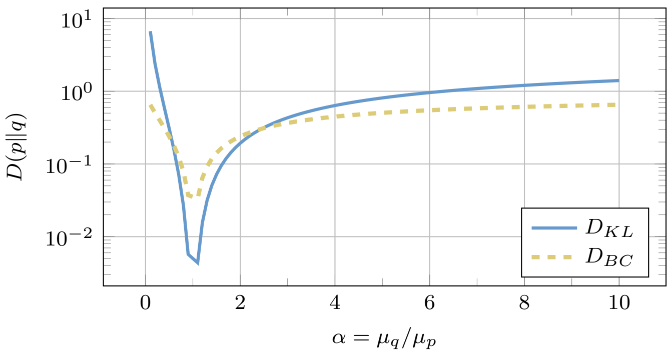


Fig. 1. Illustrative example for exponential distributions, of how the divergence measures change as the parameters of the underlying distributions diverge from each other.

III. Meaning of Divergence

This section is meant to provide some intuition for the divergence values. The values of the divergences, in general, are hard to interpret. Therefore, an example is useful to establish a rough baseline for the examples that come later.

Consider two Exponential distributions p and q, with parameters μ_p and μ_q respectively. In this case both D_{KL} and D_H can be expressed in closed form,

$$D_{KL}(p\|q) = \ln\left(\frac{\mu_q}{\mu_p}\right) + \frac{\mu_p + \mu_q}{\mu_q} \tag{5}$$

$$D_H(p\|q) = \sqrt{1 - \frac{2\mu_p\mu_q}{\sqrt{\mu_p\mu_q}(\mu_p + \mu_q)}}. \tag{6}$$

Letting $\alpha = \mu_q/\mu_p$ we can study how the divergences behave when two distributions belong to the *same* family, but have different parameters,

$$D_{KL}(p\|q) = \ln(\alpha) + \frac{1-\alpha}{\alpha} \tag{7}$$

$$D_H(p\|q) = \sqrt{1 - \frac{2\alpha}{\sqrt{\alpha}(1+\alpha)}}. \tag{8}$$

These relationships are plotted in Figure 1. Note that at $\alpha = 1$ both are zero, which cannot be shown on the log-scale plot. For example if μ_q is about about double μ_p both divergences evaluate to about 0.2.

IV. Examples

The following examples consider data from the three U.S. interconnects: Eastern Interconnect (EI), Western Electric Coordinating Council (WECC), and the Electric Reliability Council of Texas (ERCOT). Additionally, the synthetic ACTIVSg2k case created on the ERCOT footprint [1] is evaluated. Unless otherwise noted, the following scheme is used to identify each case:

Case Name	Empirical Data	Functional Fit
ACTIVSg2k	●	—
EI	■	- - -
WECC	▲	·····
ERCOT	♦	-·-·

A note on case size: The cases compared here are of different sizes and one could rightfully argue this may affect the statistical results. EI has around 60 000 buses, WECC around 20 000, ERCOT around 6000, and ACTIVSg2k has 2000 buses. Also, a basic feature of the ACTIVSg2k is that it is a *compressed* version of the ERCOT case in the sense that it supports a total load in the same ball-park as the ERCOT case and it has the same generation fleet but it has a third of its buses, meaning that some load buses aggregate load that in the ERCOT case is dispersed. Therefore, the goal of the analysis is to identify trends which fit all three interconnect cases. Since the three interconnects are an order of magnitude apart, if a trend fits all three, it is reasonable to assume that scaling does not play a major role.

A. Degree Distribution

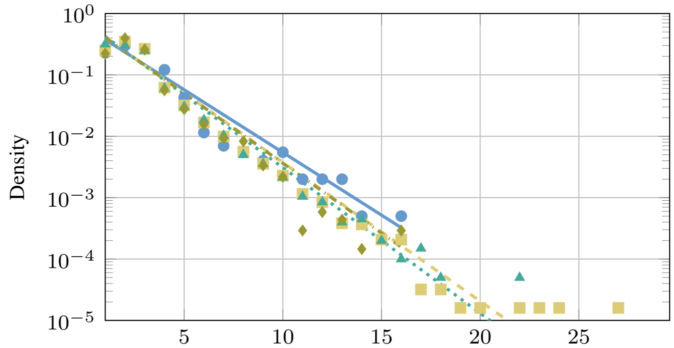
Due to its prevalence in literature, we begin by considering the degree distribution of the networks, which are fit with a geometric distribution,

$$\mathbb{P}(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots \tag{9}$$

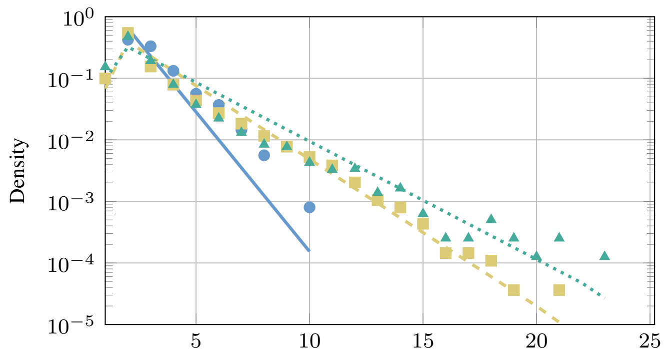
by letting p equal the $1/\bar{k}$, where $\bar{k}$ is the average node degree. Figure 2a shows the degree distributions for all four cases, which are quite similar. Table IIa reports the estimated p parameter, as well as the divergence measures from Section II. The parameters are closely clustered together, meaning that the average degrees of all four cases are similar ($\bar{k} \approx 2.5$), as are the divergence measures. Using the largest D_H from Table IIa, Equation (4) places a bound of 28% as the worst possible variation, although the actual worst variation error is about 10% lower.

Since the values for the synthetic ACTIVSg2k grid are similar to the real cases, this particular test suggests it is realistic. We note that the ACTIVSg2k grid did not consider the degree distribution in its generation process. Instead, it appears that this statistic emerges as a result of the geographic embedding and utilization of the Delaunay Triangulation as described in [1].

A recurring issue with the treatment of degree distributions in the power systems literature, is that the physical representation of a node is often not clearly defined. Although commonly buses are treated as nodes, due to the typical representation of cases in software, for topology analysis, we believe that considering substations as nodes is a more reasonable approach. From the power systems engineering perspective, the reasoning is that substations are much more frequently reconfigured for various reasons, from maintenance to transformer balancing. The topological question, however, is whether a particular region is connected to another and how, which is best answered at the substation level.



(a) Bus Degree Distribution



(b) Substation Degree Distribution

Fig. 2. Degree Distributions

TABLE I
DEGREE DISTRIBUTION PARAMETERS AND DIVERGENCES

Case	p	D_{KL}	D_{BC}
ACTIVSg2k	0.38	0.11	0.17
EI	0.41	0.11	0.17
WECC	0.42	0.07	0.13
ERCOT	0.41	0.15	0.2

(a) Buses as nodes: Geometric distribution

Case	p	Discrete Coeff.	D_{KL}	D_{BC}
ACTIVSg2k	0.65	[1, 0]	0.13	0.17
EI	0.42	[0.84, 0.16]	0.07	0.13
WECC	0.36	[0.73, 0.27]	0.11	0.17

(b) Substations as nodes: Mixture Geometric and Discrete

The degree distribution is therefore reconsidered, where buses inside a substation are grouped into "super-nodes." Unfortunately, substation data is not available for the ERCOT case, which is therefore left out of this analysis. However, since bus degree distribution in Figure 2a of ERCOT is similar to the other interconnects, it seems reasonable to assume that the substation picture should not alter dramatically. Results of the analysis are reported in Figure 2b and Table IIb.

In an attempt to achieve a better fit to the distributions, the method of combining a clipped geometric distribution with a finite discrete distribution from [2] is applied in a slightly modified manner. The coefficients of the discrete distribution/polynomial in Table IIb are reported in decreasing degree order so that the rightmost element corresponds to the constant term. For example, $[0.3, 0.7]$ represents $0.3z + 0.7$.

While the divergence measures are reasonably small for all cases, the parameters for the ACTIVSg2k case differ more starkly from the other two. There is, however, a design explanation for this discrepancy. In the ACTIVSg2k case no substations are allowed to have a degree of one, which is consistent with normal practice. The real cases do, however, exhibit such radial substations. The lack of radial substations creates an additional pole at 0 for the discrete probability generating function, as seen by the coefficients in Table IIb, which means a fundamentally different convolution in the formation of the final probability mass function.

B. Cycle Distribution

In addition to the degree distribution, we consider the minimum cycle distribution of the graphs. Cycles offer a way to describe the "meshness" of the system. They also directly relate to the classic circuit mesh analysis. In fact, Kirchhoff's work [27] used fundamental cycle bases for the application of his famous voltage law. To obtain a better level of consistency, we use minimum cycle bases in the following analysis, which are calculated using the algorithm in [28].

The cycles of both the buses and substation graphs are found to follow a Negative Binomial distribution, which is defined as,

$$f(k; p, r) = \frac{\Gamma(k+r)}{k!\Gamma r} p^k (1-p)^r \quad k = 0, 1, \ldots, \tag{10}$$

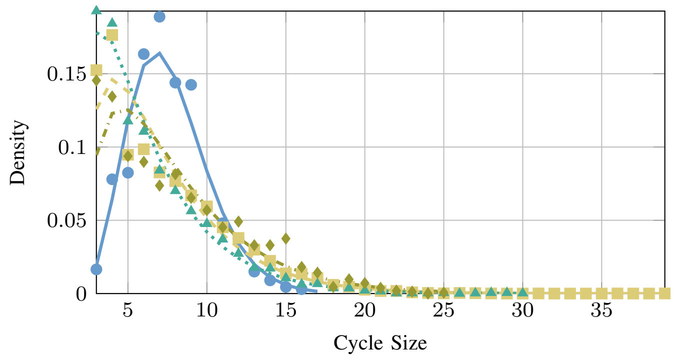
where $\Gamma(\cdot)$ is the gamma function, $p \in (0, 1)$ and $r > 0$. On a technical implementation note, for this particular application we map zero to 3, meaning that $k = 0$ implies that there are 3 edges in the cycle, since there are no cycles smaller than 3.

Once again the parameters are tightly clustered for the three interconnects according the values in Table II. Interestingly, the parameters do not vary too dramatically between the bus (Table IIIa) and substation (Table IIIb) graphs, which means that analysis can be focused on the bus level graph, where more cases are available. ACTIVSg2k has quite different parameters, which can also be observed visually in Figure 3, where its mode is clearly to the right of the interconnect cases.

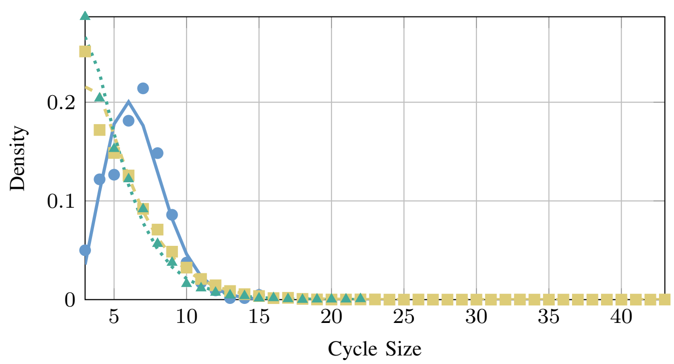
The operational implication of a higher concentration of larger cycles, as in the ACTIVSg2k case, in still an open question. Therefore, from the validation perspective, it merely warrants a caution. It is entirely possible that this degree of variation will not result in a large operational difference. Future work will investigate the effect of altering these metrics on the power system operations.

C. Surge Impedance Loading

At this point, attention is turned to a couple operational tests. While the topological tests can indicate potential structural issues with synthetic samples, the operational ones are of most interest for the researchers who will be using the cases. Line loading is considered as a fraction of its surge impedance loading (SIL). SIL represents an impedance matched termination

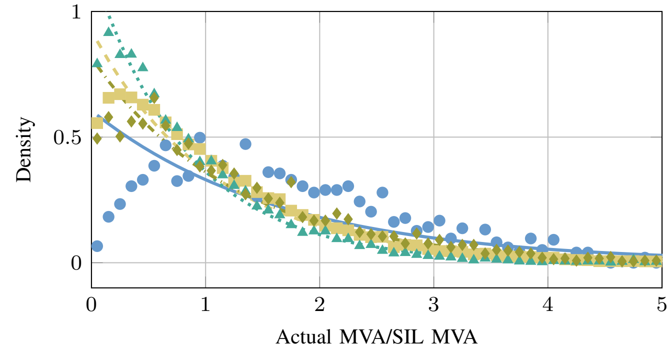


(a) Buses as nodes

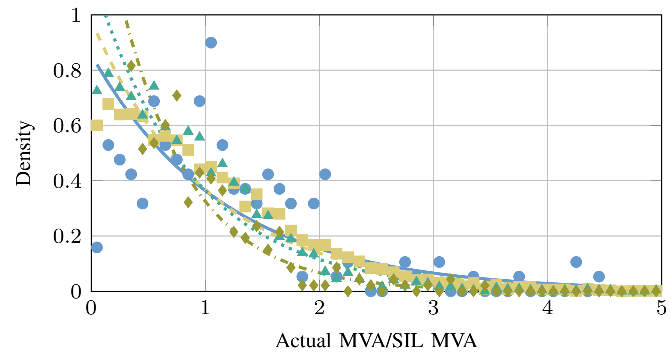


(b) Substations as nodes

Fig. 3. Minimum cycle distributions and corresponding Negative Binomial fit using both buses and substations as graph nodes.



(a) $80 < V_{\text{nom}} < 200\text{kV}$



(b) $200 < V_{\text{nom}} < 400\text{kV}$

Fig. 4. Fraction of SIL loading

TABLE II
NEGATIVE BINOMIAL FIT PARAMETERS AND DIVERGENCE MEASURES TO CYCLE DISTRIBUTION

	p	r	D_{KL}	D_H
ACTIVSg2k	0.27	12.66	0.03	0.08
EI	0.73	1.60	0.02	0.08
WECC	0.73	1.32	0.01	0.05
ERCOT	0.74	1.76	0.04	0.11

(a) Buses as nodes

	p	r	D_{KL}	D_H
ACTIVSg2k	0.14	21.86	0.03	0.08
EI	0.64	1.52	0.01	0.05
WECC	0.61	1.42	0.01	0.05

(b) Substations as nodes

for a lossless line, in the sense that the voltage profile is flat at the nominal voltage, when the line delivers its SIL [29]. The fraction of SIL loading, shown in Figure 4 for two different voltage ranges, roughly follows an exponential distribution,

$$f(x) = \frac{1}{\mu} e^{-x/\mu}, \quad x, \mu > 0. \tag{11}$$

Values for parameter μ and divergences are given in Table III. It should be noted that $\mu \approx 1$ is to be expected, since it suggests that lines are *on average* loaded at their SIL rating. Since at SIL the voltage profile is flat and power systems operation attempt to maintain voltages close to $1\,\text{pu}$, an average SIL around one is quite logical. With the exception of ERCOT for the higher voltage range, all the interconnects exhibit μ close to one.

The ACTIVSg2k case has both a poorer fit to the exponential trend as well as a significantly higher average loading than the other interconnects in both voltage ranges. This violation is a bit more concerning as it could have impacts on other studies. For example, the authors of [22] show that the LMPs on the synthetic network follow similar patters as on a real system, however, the figures appear to show different absolute values for the LMPs. Since LMPs are closely related to system congestion, the higher loading level is likely to impact these sort of studies. It should be emphasized however, that we do not claim to show this impact, as with the cycle distribution, causal implication of differing distributions are outside the scope of this paper, not currently known, and are left for future work.

Our suspition is that the discrepancy may be attributed to the optimization approach described in [1], as it tries to place lines to maximize the power they will carry. This analysis suggest this is not exactly what happens in the system, and in fact, many lines are highly underutilized.

D. Angle Difference

The next operational feature considered is the voltage angle difference, $\Delta\theta$, over an active line. Based on the DC-

TABLE III
EXPONENTIAL FIT TO SIL LOADING FRACTION

Case	$80 < V_{nom} < 200$kV			$200 < V_{nom} < 400$kV		
	μ	D_{KL}	D_H	μ	D_{KL}	D_H
EI	1.08	0.02	0.08	1.04	0.04	0.11
WECC	0.85	0.02	0.07	0.88	0.06	0.13
ERCOT	1.23	0.04	0.11	0.63	0.05	0.13
ACTIVSg2k	1.65	0.19	0.26	1.17	0.25	0.29

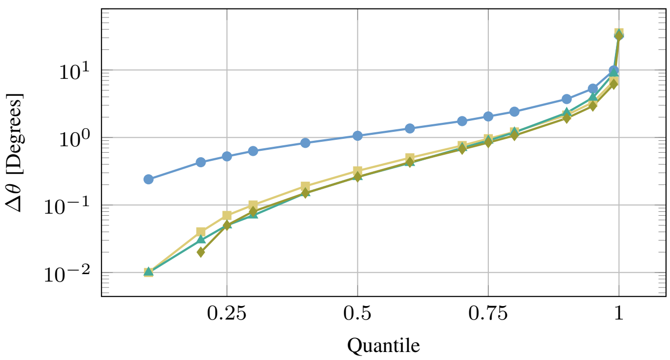


Fig. 5. Quantile function of branch $\Delta\theta$.

Powerflow assumptions [30], $\Delta\theta$ is directly proportional to the power transfer. It is also closely related to system stability [31].

A good functional fit was not found for the observed distributions of $\Delta\theta$ and therefore the quantile and numerical view introduced in Section II is considered. Figure 5 shows the quantile function or inverse cdf for the interconnect case and ACTIVSg2k.

Based on the interconnect cases in Figure 5 reasonable quantile mileposts are judged to be:

First quarter	$0° \leq \Delta\theta \leq 0.1°$
Second quarter	$0.1° \leq \Delta\theta \leq 0.5°$
Third quarter	$0.4° \leq \Delta\theta \leq 1.0°$
Fourth quarter	$1° \leq \Delta\theta \leq 35°$

The ACTIVSg2k lies outside of these ranges for all but the last quarter. This corresponds with the observation in the previous section that the ACTIVSg2k grid is on average more loaded (with respect to SIL) than the other cases. It is important to point out that the differences are not that large, the first three quarters have very small angles. Given the scale differences between the cases it is quite possible that at least part of the difference is the degree of modeling of very short, more "local" lines, which will typically have much smaller angle differences. Nonetheless, this observations merits further study to tease out why the differences may be arising, and perhaps more importantly, if the differences have any meaningful impact on other applications like OPF, SCOPF, and so on.

To provide a similar quantitative comparison similar to the divergence measure, we use the two-sample Kolmogorov-Smirnov test, shown in Table IV. It is important to stress that we *do not* suggest that the distributions are actually the same, therefore we *expect* the null hypothesis to be rejected (i.e. large test statistic and very low p-values). The test only provides a *relative* interpretation of how a pair of cases compare versus a different pair. This can be best seen in the p-value, which indicates the probability that two samples from the same distribution would produce the same result or worse as the two test samples. Essentially, the lower the p-value the less likely that the two samples were drawn from the same distribution. While all the p-values in Table IV are very small, meaning the samples are not likely to be drawn from the same distribution, the exponent between ACTIVSg2k and the other is an order of magnitude smaller. Since this is an empirical, non-parametric test, the size of the samples is likely to play a role. However, the ERCOT case is much closer in size to ACTIVSg2k and likewise a similar order smaller than the WECC and EI cases. Nonetheless, the ERCOT case compares noticeably more favorably with the two other interconnects than ACTIVSg2k.

V. CONCLUSION

The paper lays out a methodology for testing various metrics and demonstrates a few useful ones. Presented examples show that the engineering optimization approach to validation is quite powerful and can match, without direct targeting, structural properties like the degree distribution. However, the coupled nature of operation and structure that the power system exhibits is subtle and susceptible to over optimization. As a remedy, some randomness championed by statistical synthesis methods like [2] and [32] may need to be borrowed and inserted into the optimization to account for certain observable trends.

Realistic synthetic power system cases, in terms of complexity and scale, are needed by the research community to accurately support simulations. While the metrics discussed here pertain to the static view of the grid, dynamics can be expanded to in the future. The set of useful validation metrics is an ever growing and changing set. It is therefore imperative that a consistent and clear treatment is used for validation, which can be improved over time and help iterate ever better synthetic cases.

VI. ACKNOWLEDGEMENTS

This work was funded in part by the Advanced Research Projects Agency- Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000714. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

[1] A. Birchfield, T. Xu, K. M. Gegner, K. S. Shetye, and T. J. Overbye, "Grid Structural Characteristics as Validation Criteria for Syntheti Netowrks," *IEEE Transactions on Power Systems*, 2106.

[2] Z. Wang, A. Scaglione, and R. 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, June 2010.

TABLE IV
TWO-SAMPLE KS TEST FOR ANGLE DIFFERENCES. ENTRIES ARE [STATISTIC, p-VALUE].

	ACTIVSG2000	EI	WECC	ERCOT
ACTIVSg2k	—	$[0.38, 8.37 \times 10^{-229}]$	$[0.42, 3.05 \times 10^{-254}]$	$[0.41, 8.93 \times 10^{-218}]$
EI	$[0.38, 8.37 \times 10^{-229}]$	—	$[0.05, 5.37 \times 10^{-20}]$	$[0.05, 1.36 \times 10^{-14}]$
WECC	$[0.42, 3.05 \times 10^{-254}]$	$[0.05, 5.37 \times 10^{-20}]$	—	$[0.05, 4.03 \times 10^{-10}]$
ERCOT	$[0.41, 8.93 \times 10^{-218}]$	$[0.05, 1.36 \times 10^{-14}]$	$[0.05, 4.03 \times 10^{-10}]$	—

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

[4] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Phys. Rev. E*, vol. 69, p. 025103, Feb 2004.

[5] V. Rosato, S. Bologna, and F. Tiriticco, "Topological properties of high-voltage electrical transmission networks," *Electric Power Systems Research*, vol. 77, no. 2, pp. 99 – 105, 2007.

[6] G. A. Pagani and M. Aiello, "The Power Grid as a Complex Network: a Survey," *ArXiv e-prints*, May 2011.

[7] E. Cotilla-Sanchez, P. 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, pp. 616–626, Dec 2012.

[8] P. Hines, E. Cotilla-Sanchez, and S. Blumsack, "Do topological models provide good information about electricity infrastructure vulnerability?" *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, pp. –, 2010. [Online]. Available: http://scitation.aip.org/content/aip/journal/chaos/20/3/10.1063/1.3489887

[9] C. D. Brummitt, P. D. H. Hines, I. Dobson, C. Moore, and R. M. D'Souza, "Transdisciplinary electric power grid science," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, p. 12159, 2013.

[10] M. Rosas-Casals, S. Bologna, E. F. Bompard, G. D'Agostino, W. Ellens, G. A. Pagani, A. Scala, and T. Verma, "Knowing power grids and understanding complexity science," *International Journal of Critical Infrastructures*, vol. 11, no. 1, pp. 4–14, 2015.

[11] L. Verheggen, R. Ferdinand, and A. Moser, "Planning of low voltage networks considering distributed generation and geographical constraints," in *2016 IEEE International Energy Conference (ENERGYCON)*, April 2016, pp. 1–6.

[12] N. Rotering, C. Schroders, J. Kellermann, and A. Moser, "Medium-voltage network planning with optimized power factor control of distributed generators," in *Power and Energy Society General Meeting, 2011 IEEE*, July 2011, pp. 1–8.

[13] E. G. Carrano, L. A. E. Soares, R. H. C. Takahashi, R. R. Saldanha, and O. M. Neto, "Electric distribution network multiobjective design using a problem-specific genetic algorithm," *IEEE Transactions on Power Delivery*, vol. 21, no. 2, pp. 995–1005, April 2006.

[14] S. H. Elyas and Z. Wang, "Improved synthetic power grid modeling with correlated bus type assignments," *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[15] Z. Wang, S. Elyas, and R. Thomas, "A novel measure to characterize bus type assignments of realistic power grids," in *PowerTech, 2015 IEEE Eindhoven*, June 2015, pp. 1–6.

[16] E. Schweitzer, A. Scaglione, R. Thomas, and T. Overbye, "Analysis of the Coupling Between Power System Topology and Operating Condition for Synthetic Test Case Validation," in *2016 Grid of the Future Symposium*. CIGRE US National Committee, 2016.

[17] D. J. Klein and M. Randić, "Resistance distance," *Journal of Mathematical Chemistry*, vol. 12, no. 1, pp. 81–95, 1993.

[18] F. Thiam and C. DeMarco, "Transmission expansion via maximization of the volume of feasible bus injections," *Electric Power Systems Research*, vol. 116, pp. 36 – 44, 2014.

[19] E. Cotilla-Sanchez, P. D. H. Hines, C. Barrows, S. Blumsack, and M. Patel, "Multi-attribute partitioning of power networks based on electrical distance," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4979–4987, Nov 2013.

[20] P. Cuffe and A. Keane, "Visualizing the electrical structure of power systems," *IEEE Systems Journal*, vol. PP, no. 99, pp. 1–12, 2015.

[21] P. Hines, I. Dobson, and P. Rezaei, "Cascading power outages propagate locally in an influence graph that is not the actual grid topology," *IEEE Transactions on Power Systems*, vol. PP, no. 99, pp. 1–1, 2016.

[22] 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.

[23] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951.

[24] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen." *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[25] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 564–577, May 2003.

[26] A. O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval$1/4$ (revised)," Communications and Signal Processing Laboratory, University of Michigan Ann Arbor, Tech. Rep. CSPL-328, 2001.

[27] G. Kirchhoff, "Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird," *Annalen der Physik*, vol. 148, no. 12, pp. 497–508, 1847.

[28] F. Berger, P. Gritzmann, and S. de Vries, "Minimum cycle bases for network graphs," *Algorithmica*, vol. 40, no. 1, pp. 51–62, 2004.

[29] J. D. Glover, M. Sarma, and T. Overbye, *Power System Analysis & Design*, 5th, Ed. Cengage Learning, 2012.

[30] A. Wood, B. Wollenberg, and G. Sheblé, *Power Generation, Operation, and Control*, 3rd ed. Wiley, 2014.

[31] J. Machowski, J. Bialek, and J. Bumby, *Power system dynamics: stability and control*. John Wiley & Sons, 2011.

[32] J. Hu, L. Sankar, and D. J. Mir, "Cluster-and-connect: An algorithmic approach to generating synthetic electric power network graphs," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sept 2015, pp. 223–230.