Automated Generation Algorithm for Synthetic Medium Voltage Radial Distribution Systems

Eran Schweitzer, *Student Member, IEEE*, Anna Scaglione, *Fellow, IEEE*, Antonello Monti, *Senior Member, IEEE*, and Giuliano Andrea Pagani, *Member, IEEE*

***Abstract*— To introduce more automation in the distribution level of the power system, increasingly more data are needed to serve as input to control algorithms. To examine the complex interaction between the various layers of the system and verify the effectiveness of automation, difficult to obtain, realistic test systems are necessary. An algorithm for automatically synthesizing realistic medium voltage distribution feeders, and thus circumvent the data access problem, is presented. The algorithm treats distribution system feeders as graphs with nodes and edges, each with various properties, and leverages this structure to search for emerging statistical patterns. Using a large data set from a DSO in The Netherlands, clear statistical distributions are identified linking properties, such as load, node degree, or cable length, to the feeder structure. Specifically, many properties are linked to a node's or edge's distance, in hops, from the primary substation. With consideration for standard engineering practices, the statistical trends are exploited in the synthesis process, to generate feeders, which display similar characteristics to the real samples. The KL-divergence is used in the evaluation of analysis and synthesis results. Beyond solving the data access problem, the use of automatically generated, synthetic, distribution systems will enable testing and validation techniques, such as Monte Carlo simulations, which are currently not possible in this field, where single test cases are the norm.**

***Index Terms*— Distribution systems, statistical distributions, synthetic test cases, KL-divergence.**

I. Introduction

COMPLEXITY in distribution systems is increasing, driven by increasing distributed generation penetration. In light of this trend, automations and controls from the transmission system are migrating to what was previously considered a passive system [1]. To meet the goals of future distribution systems such as [2]: 1) self-healing from disturbances, 2) enabling prosumers, 3) or accommodating various generation and storage options, algorithms for state estimation [3], fault location [4], [5], and voltage control [6] are actively being adapted to distribution feeders.

Development and testing of all the algorithms require test systems to verify behavior. Unfortunately, such data is often difficult to obtain due to security or proprietary concerns on the side of the utility. There is, therefore, a real need for synthetic systems [7], [8] including the recent Grid Data project from ARPA-E.[1] While much effort has been placed on transmission systems, several distribution test cases are available, [9]–[11]. Additionally, PNNL published a report in 2008 with some prototypical feeders [12].

Our approach here is decidedly different from the synthetic test cases mentioned. We view the distribution system as a graph whose nodes and edges can be imbued with various properties. In the spirit of Complex Network Science (CNS), we search for statistical patterns that emerge in this graph and its properties, and use these to synthesize similar systems. In doing so, the amount of data needed to generate a test case is reduced to the relatively few parameters of several distributions. Additionally, the process can be trivially automated, enabling the creation of many samples. As a result, previously impossible testing and validation regimes like Monte Carlo simulations, to observe in what percentage of cases an application functions as expected, become realizable. This paper focuses on radial feeders, although Section VIII signals how we plan to move beyond this simplification. A large majority of distribution feeders are radial, or at least operated radially [13], justifying the initial focus on this topology.

A. Related Work

CNS has been used extensively in transmission network analysis for over a decade [7], [14]–[18]. Its use on the distribution system, however, has been rather limited, with [19] as the main exemplar. We expand upon the analysis in [19] using a more complete dataset than was available at that time. Our analysis targets more of the electrical properties of the system and exploits the simplified topology of a radial feeder to merge these with classic CNS measures like the degree distribution. This hybrid approach seeks to bridge the gap between purely topology-oriented works and power engineering methods, which has been suggested as a beneficial approach for future study in [20] and [21].

Beyond a few metrics, focus on statistically emergent properties links our work and traditional CNS studies. Our basic premise is that clear, statistical patterns emerge in large complex systems. We show that certain properties, like branch voltage drops or power flows, which can be seen as edge weights, emerge naturally from the synthesis without explicit inclusion in the algorithm.

Fully, or largely, automated generation of synthetic power systems is not entirely new, with several exam-

Manuscript received September 29, 2016; revised January 1, 2017; accepted February 17, 2017. Date of publication April 12, 2017; date of current version June 10, 2017. This work was supported in part by the Advanced Research Projects Agency-Energy through the U.S. Department of Energy under Award DE-AR0000714 and in part by Flexible Electrische Netze GmbH through Forschungscampus Elektrische Netze der Zukunft. This paper was recommended by Guest Editor H. H.-C. Iu.

E. Schweitzer and A. Scaglione are with Arizona State University, Tempe, AZ 85287-5706 USA (e-mail: eranschweitzer@gmail.com).

A. Monti is with the Rheinisch-Westfalische Technische Hochschule Aachen, 52072 Aachen, Germany.

G. A. Pagani is with the Rijksuniversiteit Groningen, 9747 AG Groningen, the Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JETCAS.2017.2682934

[1] http://arpa-e.energy.gov/?q=arpa-e-programs/grid-data

ples in [7], [22], and [23]. Recently, [24] described a method for generating stochastic feeder data in PNNL's GridLAB-D[2] environment. However, this approach is mainly designed to enhance the load model as seen from the transmission system, and does not, therefore, go into much detail on how the distribution system is constructed. Beyond this, and our previous work, which focused on topology [25], we are unaware of other automated methods for generating distribution systems.

In [25], our focus centered on matching topological features. This was achieved by embedding the graph in a two dimensional plane as in [26] and [27]. However, we found that the topology oriented geometric embedding applied very restrictive constraints to matching reasonable electrical parameters. This motivated the approach of linking topology and electric parameters from the beginning in the modeling.

The novelty in this paper is to propose a methodology that is able to easily generate varied sets of test systems with minimal input. Several works in the field of planning [28], [29], also automate the process of power system creation. However, their framework and objectives are quite different. First, we are not trying to optimize. Our goal is not to find the *best* feeder, but rather a wide set of feeders one might encounter. Distribution feeders are the result of a complex set of decisions, constraints and optimizations, as illustrated by the planning literature. We claim these lead to the emergent behaviors we observe and model directly in our methodology. Additionally, the planning approach generally starts from fixed points in space, i.e., the load is taken to be known both spatially and in quantity for a given scenario. Our feeders are in principle agnostic to location, beyond the fact that certain regions will exhibit varying construction styles, reflected in modified distributions.

The remainder of the paper is structured as follows. Section II introduces a few notational conventions. Section III describes the data and how it is processed and analyzed. Section IV describes the findings of the analysis alongside the synthesis. Section V addresses some modifications during synthesis, and Section VI presents results from the generation algorithm. Section VII discusses the results, and the generality of the methodology presented. Finally, Section VIII concludes the paper, suggests possible applications for the feeders, and addresses future research directions.

II. Notation

The following notational conventions are adopted for clarity:

Variable	Explanation
N	Total number of nodes in a feeder
n	Single node object
M	Total number of branches in a feeder
m	Single branch object
$\mathbb{1}(\cdot)$	The indicator function, which evaluates to one if the argument is true and zero if it is false
$\mathcal{U}(0, 1)$	Uniform distribution on the unit interval

We use the dot notation to represent an object's properties. For example, $n.P$ is the the real load attached to node n, and $m.I_{\text{est}}$ is the estimated current in branch m. Finally, to refer to multiple individual nodes or branches, subindices are occasionally used for clarity, so that n_i and n_j are two distinct nodes on a feeder.

[2]http://www.gridlabd.org/

TABLE I
Data Component Overview

Buses	21 118
220 kV	6
110 kV	53
20 kV	708
10 kV	18 357
3 kV	1979
400 V	15
Branches	23 041
Underground Cables	21 274
Transformers	711
Link	996
Overhead Lines	7
Reactance Coils	53
Node Objects	
HV Grid Connection	64
Transformer Loads	17 548
Loads	1494
Generators	461

III. Analysis Overview

The dataset comprises the Medium Voltage (MV) system from one of the DSOs in the Netherlands, covering an area around 8200 square kilometers. Summary statistics are provided in Table I.

The data was provided in several `.vnf` files, the proprietary format of the Vision software from Phase2Phase.[3] From there, the data was exported to Excel and imported to a PostgreSQL[4] database for easier manipulation.

A. Feeder Identification

For the purposes of our analysis, we define a feeder as a section of the distribution system fed by a single primary substation MV bus, plus the High Voltage (HV) source bus on the other side of the distribution transformer. To identify the feeders, the complete system data was gathered into a large graph, $G(V, E)$,[5] with buses as the vertices, V, and all the branch elements as the edges, E. Importantly, only branches that are connected at both ends are used. There are 20 903 of these branches as opposed to the full count shown in Table I.

Beginning at each HV source, all its neighbors, η_i, in G are identified. Two nodes, v_i and v_j are neighbors, if there exists an edge $e = \{v_i, v_j\}$, with $e \in E$. Each η_i is used as the starting point of a Breadth First Search (BFS) [30] that excludes the HV source and its other neighbors. All of the nodes found in the BFS constitute the feeder. We refer to the High Voltage (HV) node as the *source*, and to η_i as the *root* of the feeder. Around 100 such feeders are identified in the data.

An additional set of feeders was generated by grouping nodes that are separated by very small impedances.[6] These

[3]http://www.phasetophase.nl/en_products/vision_network_analysis.html

[4]https://www.postgresql.org/

[5]We use V and E here to differentiate between the full distribution system graph and the individual feeders with N and M, which are subgraphs of G.

[6]Primarily the Links that have $R = X = 1\ \mu\Omega$.

"reduced" feeders are used for much of the analysis since the difference between a large busbar or two smaller busbars connected by negligible impedance is, for us, immaterial.

B. Feeder Analysis

Each feeder is analyzed individually for various node and edges properties. These include topological properties such as node degree and hop distance from the source, h, where hop distance is the number of edges along a path between two nodes. Note that, in contrast to meshed transmission networks [7], our construction of distribution circuits starts from trees, which are representative of feeders. Therefore, since at this stage open connections are ignored, there is no ambiguity about the class of graph we are dealing with.

For tree graphs representing distribution feeders, h can be thought of as similar in spirit to the betweenness centrality, while the distribution of h gives a sense for the average path length (both metrics are common in much CNS literature). In fact, in the context of a radial feeder, the one path of interest for each node is the one between it and the substation. Other common topological features such as clustering coefficients do not make sense for this analysis, since clustering on a tree is, by definition, zero.

Additionally, electrical properties like load at nodes, as well as actual and nominal branch currents are collected. From the graph perspective, these are different weights. Analysis of specific properties can then be conducted over all the nodes or edges in all of the feeders, granting access to a larger sample pool and therefore, more reliable statistics.

The main objective of the analysis is to identify clear distributions in the data that can be exploited in a synthesis process. Distributions that are a good fit to the cumulative data are compared to each individual feeder to determine the range of deviation at the feeder level from the cumulative trend. Note that although the dataset comprises one distribution system, it comprises about a hundred independent feeders, which are the meaningful cases under analysis.

C. Verification Methodology

Our tool of choice for testing how well the synthetic feeders match the real data is the KL-Divergence [31],

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

which is often used to characterize the distance between two distributions.[7]

Meaningful ranges for the KL-Divergence are determined in the following way:

1) The functional law is determined by considering aggregate data from *all* the feeders, for higher statistical relevance. The distribution that exhibits the lowest D_{KL} with respect to the data is selected.
2) We consider the distribution of KL-Divergences between *each* individual feeder and the selected functional law. This provides a weighted range for D_{KL}, given the selected functional law.

[7]The operational meaning of KL distance is as follows: an observer trying to determine if data come from the distribution $p(x)$ rather than $q(x)$ will be wrong with a probability that decays exponentially in the number of independent observations, with a rate that is the KL distance. Therefore, a small KL distance means that a significant number of samples can be generated form distribution $p(x)$, that look indistinguishable from data generated from the statistic $q(x)$ [32].

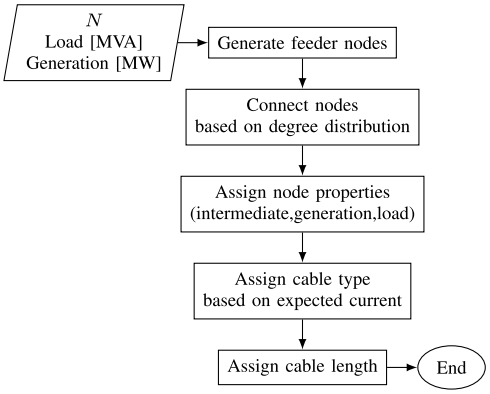


Fig. 1. Overview of feeder generation algorithm.

IV. Data Analysis and Synthesis Algorithm

In this section, we present the data analysis and synthesis alongside each other. The flowchart in Figure 1 outlines the main steps in the feeder generation algorithm. In each step, trends in the form of distributions are identified and then exploited for synthesis. Throughout the paper we provide some intuition as to why a particular distribution is a reasonable modeling choice for the data. This intuition is important for potential expansion and manipulation of the algorithm. By adjusting the parameters of the various distributions, the generation logic is preserved while more extreme or conservative results are achieved, which could be of interest.

A. Node Generation

1: **procedure** GENERATE NODES(Power Factor cdf, Negative Binomial distribution)
2: The first node is by design the source at $n.h = 0$
3: The second node is by design the only node at $n.h = 1$
4: **for** $n = 3, 4, \ldots, N$ **do**
5: n.power factor $\leftarrow$ power factor from input cdf
6: $n.h \leftarrow$ sample from the Negative Binomial distribution
7: Adjust hop distances so there are no gaps

The radial assumption lies at the foundation of the synthesis algorithm because is allows each node to be characterized in terms of distance in *hops* away from the HV source, which is by design the first node in the feeder. For example, the root as described in Section III-A is by definition one hop away from the source, which we denote as $n.h = 1$. Figure 2 shows the distribution of hop distances in the dataset as well as a fit line following the Negative Binomial distribution,

$$f(x; r, p) = \frac{\Gamma(r + x)}{x!\Gamma(r)} p^r (1 - p)^x, \tag{2}$$

TABLE II
KL-DIVERGENCES

Property	Distribution	Cumulative D_{KL}	Per Feeder D_{KL}† $< 90\%$	$< 95\%$	< 1	Synthetic Samples
Hop Distance	Negative Binomial	0.0173	0.3903	2.3022	92%	0.0101
No-Load						
Fraction	Beta	0.0014	—	—	—	0.0242
Hop Distance	Bimodal Poisson	0.0755	—	—	—	0.0233
Power Injection						
Fraction	Beta	0.0620	—	—	—	0.2968
Hop Distance	Bimodal Normal	0.1706	—	—	—	0.4240
Deviation From Uniform	Normal	0.0459	—	—	—	0.2031
Load Deviation From Uniform	tLocationScale	0.0008	3.4103	4.5785	83%	0.1329
Degree Distribution	Bimodal Gamma	0.0211	0.1457	0.2701	99%	0.0147
$I_{\text{est}}/I_{\text{nom}}$	Exponential	0.0098	0.2010	0.3795	98%	0.0102
Cable Length	Modified Cauchy	0.0247	0.6967	1.1387	95%	0.0108
Downstream Power	Generalized Pareto	0.0111	0.6691	1.0766	94%	0.0243
Voltage Drop	Generalize Pareto	0.0917	0.9961	1.5091	90%	0.0216

† The number in column $< 90\%$ says that 90% of the individual feeders have a KL-Divergence with the functional law below this number, similarly for column $< 95\%$. Column < 1 reports the percent of feeders whose KL distance to the functional law is less than 1.

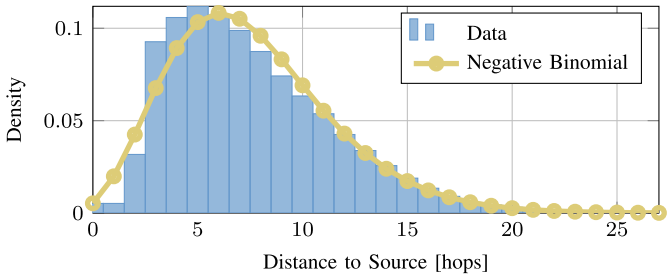


Fig. 2. Distribution of hop distances and the Negative Binomial fit.

where $r > 0$, $0 \leq p \leq 1$, $x = 0, 1, \ldots, \infty$, and $\Gamma(\cdot)$ is the Gamma function. The KL-Divergence for this fit, as well as the other distributions discussed in this paper, is given in Table II, and the values for the parameters of (2) are reported in Table III.

The intuition behind the Negative Binomial is its interpretation as an over-dispersed Poisson distribution. In other words, in the random process of deciding how far a node is from its source, the variance does not equal the mean, however, a mean and a variance are sufficient to describe the process.

The GENERATE NODES procedure uses samples from (2) to assign the hop distance property to each node. In addition to the hop distance, a power factor is assigned to each node, from an empirical cdf from the data, shown in Table IV. This greatly simplifies further manipulations, allowing to focus on real power.

B. Feeder Connection

Once the nodes have been created, we begin to connect them into a tree rooted at the root and by extension the source. By restricting the topology, the degree distribution actually reveals a fair amount about the feeder. The degree distribution, $P(k)$, describes the frequency of each degree in the graph, and is widely used in Complex Network Analysis [33],

$$P(k) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}(n.d = k), \tag{3}$$

procedure CONNECT NODES(Mixture Gamma, $g_{d_{\max}}(h)$)
  for all Nodes where $n.h = \max_n n.h$ or $n.h = 0$ **do**
    $n.d \leftarrow 1$
  $n.d \leftarrow 1 + \sum_n \mathbb{1}(n.h = 2)$ for the single node with $n.h = 1$
  for $n = 1, 2, \ldots, N$ **do**
    if No degree assigned **then**
      repeat
        $d_{\text{tmp}} \leftarrow$ sample from the mixture Gamma distribution
      until $d_{\text{tmp}} \leq \lceil g_{d_{\max}}(n.h) \rceil$
      $n.d^* \leftarrow d_{\text{tmp}}$
  Sort nodes into ascending order in h.
  for n=N,N-1,...,2 **do** ▷ moving from furthest nodes toward source
    Connect node n to a viable predecessor, p, ($p.h = n.h - 1$) for whom the difference between the current degree, $p.d$, and assigned degree, $p.d^*$ is most negative:

$$n.\text{predecessor} \leftarrow \min_p p.d - p.d^*$$

where $n.d$ is the degree of node n—the number of incident branches on node n.

The empirical degree distribution for all feeders is fit by a mixture of Gamma distribution,

$$f(x; p, a_1, b_1, a_2, b_2) = p \cdot g(x; a_1, b_1) + (1 - p) \cdot g(x; a_2, b_2) \tag{4}$$

with,

$$g(x; a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}, \tag{5}$$

where $a_{1,2}, b_{1,2} > 0$, $x > 0$, and $g(x; a, b)$ is the Gamma distribution pdf. The exponential degree distribution of transmission grids has been widely discussed in literature [14], [17],

TABLE III
FIT PARAMETERS

Property	Parameter Values
Hop Distance	$r = 7.46$, $p = 0.50$
No-Load	
Fraction	$\alpha = 3.03$, $\beta = 49.54$
Hop Distance	$p = 0.53$, $\mu_1 = 3.55$, $\mu_2 = 10.50$
Power Injection	
Fraction	$\alpha = 4.28$, $\beta = 246.19$
Hop Distance	$p = 0.92$, $\mu_1 = 0.12$, $\sigma_1 = 0.04$, $\mu_2 = 0.32$, $\sigma_2 = 0.32$
Deviation From Uniform	$\mu = 0$, $\sigma = 0.15$
Load Deviation From Uniform	$\mu = -0.001$, $\sigma = 0.002$, $\nu = 1.46$
Degree Distribution	$p = 0.03$, $a_1 = 5.30$, $b_1 = 1.24$, $a_2 = 9.00$, $b_2 = 0.21$
$I_{\text{est}}/I_{\text{nom}}$	$\mu = 0.17$
Cable Length	$x_0 = 0.4807$, $\gamma = 0.3595$
Downstream Power	$k = 0.27$, $\sigma = 0.015$, $\theta = 0$
Voltage Drop	$k = 0.67$, $\sigma = 4.12 \times 10^{-4}$, $\theta = 0$
Maximum Degree	$a = 23.47$, $b = -0.68$
Maximum Length	$a = 26.97$, $b = -0.13$

TABLE IV
POWER FACTOR cdf

Power Factor	cdf(Power Factor)
0.85	0.1649
0.90	0.2700
0.95	1

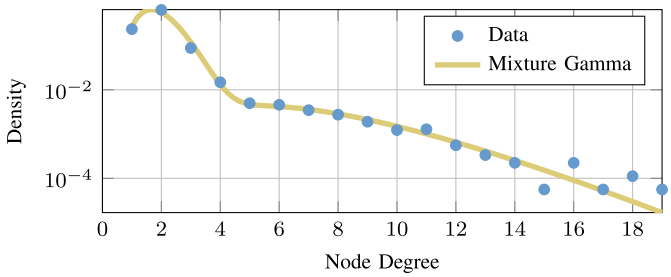


Fig. 3. Degree distribution with a mixture of Gamma distributions fit line.

[34], [35], while in [19] a more split view is given on the appropriateness of an exponential decay versus a power law for distribution systems.

The data displays a clear bimodal behavior as seen in Figure 3, with two very evident rates of decay. As the conjugate prior of the Exponential distribution, a mixture of Gamma distributions is a natural choice for modeling the two rates. This also fits with the findings in [36] that a sum of Exponential distributions provided a good fit to the degree distribution.

Due to the radial structure, nodes with maximum distance from the source must be leaves and therefore have degree one. Since by design there is only one node with hop distance one, the root, the degree of the source at hop distance zero must also be one. Finally, all nodes with hop distance two and the source must connect to the root, so its degree is also deterministically known following the hop distance assignment. For the remaining nodes, a degree is assigned based on the bimodal Gamma. The distribution is clipped based on the hop distance of each node using function $g_{d_{\max}}(h)$, which is further discussed in Section V-A.

Once each node has an assigned degree, the algorithm starts at the furthest nodes and works its way up towards the root. A predecessor, p, is picked from the viable set for each node, by choosing the one with actual degree, $p.d$, furthest below its assigned degree, $p.d^*$: $\min_p p.d - p.d^*$.

C. Node Properties

At present, node properties are the powers associated with each node. These are obtained based on a real power assignment, described below, and the power factor assigned in Section IV-A. We identify three types of nodes: intermediate (no load), generation (negative load), consumption (positive load). Each is addressed by a separate procedure. Since the number of intermediate nodes and generation nodes is quite small, single feeder statistics are omitted in Table II.

1) Intermediate Nodes: Some nodes in the data have neither positive nor negative load. Such intermediate nodes are normally either junctions from which several sub-feeders spring, or nodes associated with normally open connections. The algorithm marks these points so that load will not be assigned to them by the later procedures discussed in Sections IV-C.2 and IV-C.3.

1: **procedure** INTERMEDIATE(Intermediate Beta Distribution, Mixture Poisson Distribution)
2: $N_{\text{intermediate}} \leftarrow \lfloor N \cdot \epsilon \rfloor$, where $\epsilon \sim$ Beta distribution
3: Mark source node ($n = 1$) as intermediate.
4: **for** $i = 1, 2, \ldots, N_{\text{intermediate}} - 1$ **do**
5: $\epsilon \leftarrow$ sample from mixture Poisson distribution
6: Mark a node with $n.h = \epsilon$ as intermediate.

The INTERMEDIATE procedure first sets the number of zero load nodes, $N_{\text{intermediate}}$, by sampling a Beta distribution for the fraction of intermediate nodes (cf. Figure 4a). The Beta distribution,

$$f(x; , \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{6}$$

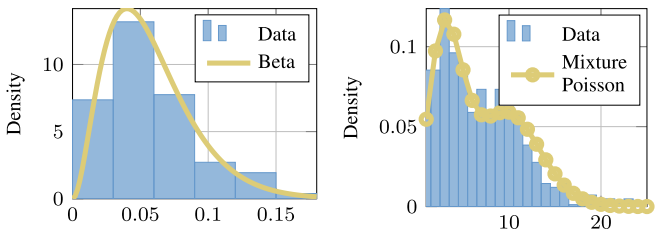
with $0 < x < 1$, and $\mathrm{B}(\cdot)$ the Beta function, is a common choice when modeling fractional quantities.

The source is designated to have zero load. For each of the remaining intermediate nodes, a sample is chosen from a mixture Poisson distribution,

$$f(x; p, \mu_1, \mu_2) = p\frac{\mu_1^x}{x!}e^{-\mu_1} + (1-p)\frac{\mu_2^x}{x!}e^{-\mu_2}, \tag{7}$$

where $x = 0, 1, \ldots, \infty$ and $\mu_{1,2} > 0$, to determine at what hop distance the node should be (cf. Figure 4b). Nodes serving as feeder junctions occur predominantly close to the primary substation, where the main sub-feeders separate from each other. Less frequently, junction points occur one third to halfway down the feeder, which may reflect further geographical splitting, or even a transition to another voltage level.[8] This physical interpretation helps justify the mixture

[8]Incorporating secondary voltage levels is currently left for future work.



(a) $\frac{1}{N}\sum_n \mathbb{1}(n.P = 0)$ (b) Distance to Source [hops]

Fig. 4. Distributions for intermediate node assignment. (a) Fraction of intermediate (zero load) nodes. (b) Hop distance of intermediate nodes.

model, and the Poisson distribution is a natural choice for a random process on the integers.

Assignment of node properties based on h exemplifies the CNS inspired nature of the procedure. The feeder's complexity lies in coupling between the structure and the node and edge properties. Conditioning on h addresses this coupling.

2) Negative Load Nodes: Negative load, or power injections, represent the "active" part of the feeder. In principle, the load at a given node is a combination of the power injected and consumed at that node. Presently, the algorithm only produces the sum total and as such, several nodes are picked based on observations from the data to have a net negative load.

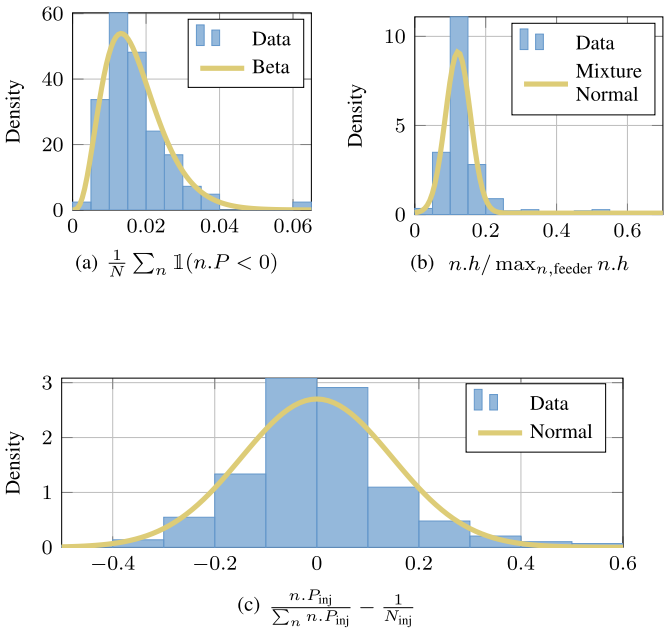
procedure POWER INJECTION(Injection Beta Distribution, Mixture Normal Distribution, Normal Distribution)
  $N_{\text{inj}} \leftarrow \text{round}(N \cdot \epsilon)$, where $\epsilon \sim$ Beta distribution
  for $i = 1, 2, \ldots, N_{\text{inj}}$ **do**
    $\epsilon \leftarrow$ sample from mixture Normal distribution
    Select one node, n, with $n.h = \lceil \epsilon \cdot \max_n n.h \rceil$
    repeat
      $\epsilon \leftarrow X \sim$ Normal
    until $1/N_{\text{inj}} + \epsilon > 0$
    $n.P \leftarrow -P_{\text{inj,total}}\left(1/N_{\text{inj}} + \epsilon\right)$
    $n.Q \leftarrow n.P \cdot \tan\left[\cos^{-1}(n.\text{power factor})\right]$

Again, the number of injection nodes, N_{inj}, is determined using a ratio sampled from a Beta distribution (cf. Figure 5a). The hop distance for each injection node is then selected by sampling a mixture Normal distribution,

$$f(x; p, \mu_1, \sigma_1, \mu_2, \sigma_2) = p \cdot g(x; \mu_1, \sigma_1) + (1 - p) \cdot g(x; \mu_2, \sigma_2), \quad (8)$$

where $0 \leq p \leq 1$, and $g(x; \mu, \sigma)$ is the normal distribution pdf with mean μ and standard deviation σ. Figure 5b shows the result of fitting (8) to the histogram of normalized hop distances, $h/h_{\max}$, where $h_{\max}$ is determined on a per feeder basis. The normalization was found to help rectify discrepancies between the longer and shorter feeders.

As would be expected, the main mode is close to the primary substation, since this is where small generators, larger PV installations, or even single wind turbines are likely to connect. The slight bump further down the feeder is most likely caused by LV feeders that are feeding back power due to the current loading scenario. While more rare, this does happen



(a) $\frac{1}{N}\sum_n \mathbb{1}(n.P < 0)$ (b) $n.h/\max_{n,\text{feeder}} n.h$

(c) $\frac{n.P_{\text{inj}}}{\sum_n n.P_{\text{inj}}} - \frac{1}{N_{\text{inj}}}$

Fig. 5. Distributions for power injection assignment. (a) Fraction of injection nodes. (b) Normalized hop distance of injection nodes. (c) Deviation in (9) of power injection from uniform distribution.

and it is expected to become more frequent as distributed generation penetration increases.

Finally, Figure 5c shows the distribution of deviation between each power injection, normalized so that all injections on a single feeder sum to one, and the uniform distribution $1/N_{\text{inj}}$. The error,

$$\epsilon = \frac{n.P_{\text{inj}}}{\sum_n n.P_{\text{inj}}} - \frac{1}{N_{\text{inj}}} \quad (9)$$

is found to be normally distributed. Real power injection is assigned by solving (9) for $n.P_{inj}$, where the while loop in the procedure is simply used to avoid sign reversals.

The POWER INJECTION module is an instance where the statistical distributions could potentially be modified to achieve progressively more "active" feeders. One simple way would be to vary the parameters of the Beta distribution, thus increasing the fraction of injection nodes.

3) Positive Load Nodes: We know from the design principles of distribution feeders that the utility attempts to distribute the load evenly across a feeder [37]. Therefore, the error, ϵ, between the actual power consumed and the uniform distribution is an interesting quantity to consider,

$$\epsilon = \frac{n.P}{\sum n.P} - \frac{1}{N}, \quad (10)$$

where each load has been normalized so that all loads on a single feeder sum to one.

Figure 6 shows the histogram generated by (10), which is indeed tightly centered around zero. The t-Location-Scale distribution,

$$f(x; \mu, \sigma, \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}, \quad (11)$$

where $\sigma, \nu > 0$, is used to fit the data, which reflects the fact that the load is symmetrically distributed around the Uniform

procedure POSITIVE LOAD(t-Location-Scale Distribution)
  for $n = 2, 3, \ldots, N$ **do**
    if Node is not an intermediate or an injection node **then**
      repeat
        $\epsilon \leftarrow X \sim$ t-Location-Scale
      until $1/N + \epsilon > 0$
      $n.P \leftarrow P_{\text{total}}\,(1/N + \epsilon)$
      $n.Q \leftarrow n.P \cdot \tan\left[\cos^{-1}(n.\text{power factor})\right]$

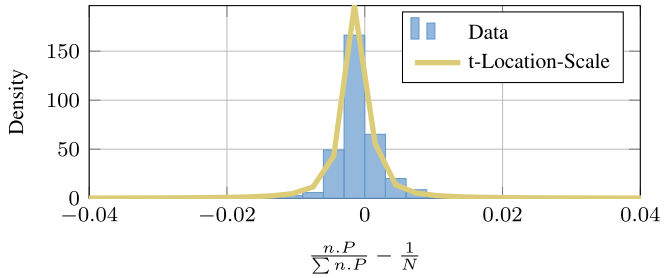


Fig. 6. Histogram of the deviation in (10) of the load from the uniform distribution.

distribution, but with heavier tails. In fact, as can be seen from the parameters in Table III, the distribution is close to being Cauchy, which is the case when $\nu = 1$. All the POSITIVE LOAD procedure does is solve (10) for $n.P$, generating ϵ by sampling the t-Location-Scale distribution. Once more, the while loop is used to avoid sign reversals.

D. Cable Type

procedure CABLE TYPE(Cable Library, Exponential Distribution)
  for m=M,M-1,…,1 **do** ▷ moves from the furthest branches towards the source.
    if $m.I_{\text{est}} \neq 0$ **then**
      $r \leftarrow \mathcal{U}(0, 1)$
      if $r < 2/3$ and some nominal current has been attached to the downstream node **then**
        $m.I_{\text{nom}} \leftarrow$ Maximum nominal current of downstream node
      else
        $I_{\text{nom,tmp}} \leftarrow m.I_{\text{est}}/\epsilon$, where $\epsilon \sim$ Exponential.
        Pick the cable from the library with closest I_{nom} taking parallel cable options into consideration as well as the expected frequencies of each cable in the feeder.
  for m=1,2,…,M **do**
    if $m.I_{\text{est}} = 0$ **then**
      $m.I_{\text{nom}} \leftarrow$ average of maximum and minimum I_{nom} attached to *upstream* node
      Pick cable from library with closest I_{nom}

Neglecting losses, the amount of power flowing in each branch of the feeder can be estimated by simply summing all downstream powers. A node n_i is *downstream* of node n_j if the path between n_i and the source passes through n_j. Similarly,

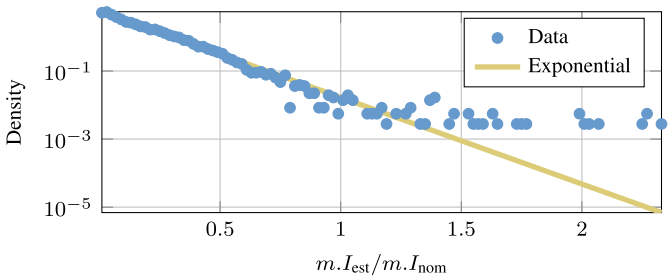


Fig. 7. Exponential fit the ratio of estimated current, I_{est}, and nominal cable current, I_{nom}.

node n_j is said to be *upstream* of node n_i. By assuming nominal voltage, the current magnitude can be calculated as:

$$\|m.I_{\text{est}}\| = \frac{\|m.S_{\text{downstream}}\|}{\sqrt{3}\|m.V_{\text{nom}}\|}. \tag{12}$$

For notational simplicity we drop the magnitude signs in the following. The Exponential distribution,

$$f(x;\mu) = \frac{1}{\mu}e^{-x/\mu}, \tag{13}$$

with $\mu > 0$, and $x \geq 0$, describes the ratio between estimated current and nominal cable current, $I_{\text{est}}/I_{\text{nom}}$, as shown in Figure 7. Since some of the feeders analyzed are not 100% radial, the calculation of $S_{\text{downstream}}$ is sometimes erroneous, leading to errors in I_{est}. These errors are largely responsible for the points that lie furthest from the fit line in Figure 7. The discrepancy is quite small, since its frequency is very small, and we find that there is no significant difference between using I_{est} as given in (12) or the currents calculated from the powerflow in Vision. Since the powerflow requires conductor parameters, which have not yet been assigned, using I_{est} offers a significant advantage.

This last point deserves reiteration. The most powerful result of the radial assumption is that we are able to calculate the powerflow *without* knowing line parameters. If the radial assumption is lifted, this is no longer valid. Since distribution systems are operated radially, we believe there is much utility even in radial models. In fact, most of the publicly available test systems, such as the IEEE8500 bus feeder [11] or all the feeders available from PNNLs project [12], are radial. Nonetheless, reconfiguration options are available and there are non radial distribution systems. We leave these for future work (see Section VIII).

In a separate analysis, all the nominal currents, I_{nom}, incident on a given node are considered. In roughly two-thirds of the cases, all are found to be the same. For implementation, a library of conductors is supplied, which was selected from the data via a k-means clustering algorithm based on the cables nominal current. The library contains all the cable data, as well as the frequency of occurrence for each cable type.

The key idea in the CABLE TYPE procedure, is that I_{nom} serves as a surrogate for the cable type. Once a desired I_{nom} is calculated, the cable which most closely matches out of the library is chosen.

The procedure performs three main functions. In two-thirds of the cases, a cable is assigned by picking the largest cable

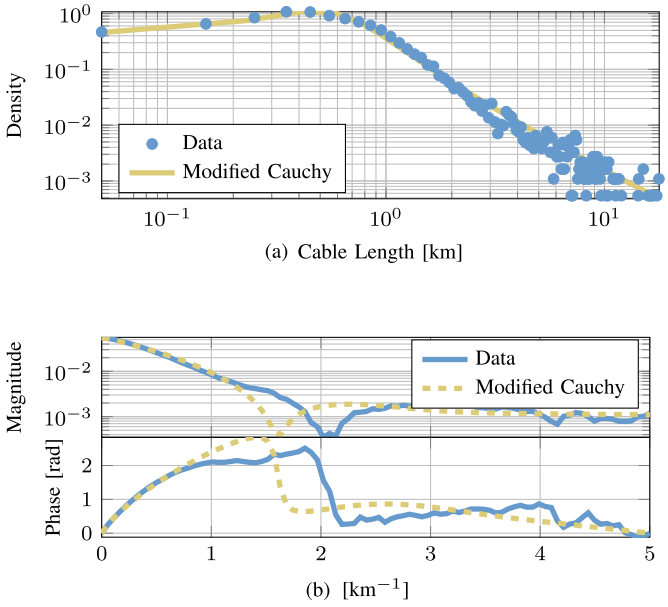


Fig. 8. Distribution of cable lengths and modified Cauchy fit. (a) Histogram. (b) FFT of histogram representing the characteristic function.

connected to the downstream node,[9] to match the above analysis. In the rest of the cases, the Exponential distribution is used to sample a ratio, $I_{\text{est}}/I_{\text{nom}}$, and then solve for I_{nom}. There are some implementation details regarding how parallel conductors are handled and how the cable type frequencies are used to weight the cable selection, but these are left out of the present discussion as they are strictly implementation issues. Finally, branches with no current are given an I_{nom} taken as the average over the incident branches on the upstream node, since the procedure using the ratio, $I_{\text{est}}/I_{\text{nom}}$, does not work in this case.

E. Conductor Length

procedure CABLE LENGTH(Modified Cauchy Distribution, $g_{\ell_{\max}}(h)$)
  for m=1,2,…,M **do**
    repeat
      $\ell_{\text{tmp}} \leftarrow$ Sample from Modified Cauchy Distribution.
    until $\ell_{\text{tmp}} \leq g_{\ell_{\max}}(m.h)$
    $m.\ell \leftarrow \ell_{\text{tmp}}$

Since cable types are assigned, and the cable library contains all the per distance parameters, all that remains is to assign length to each branch so that a total impedance could be calculated. During the investigation of the length distribution, we observe a clear exponential decay in the magnitude of the empirical characteristic function—the Fourier transform of the histogram—which can be seen in Figure 8b. Considering common characteristic functions, only the Cauchy distribution

[9]Assuming there *is* a downstream node, i.e., for non-leaf nodes.

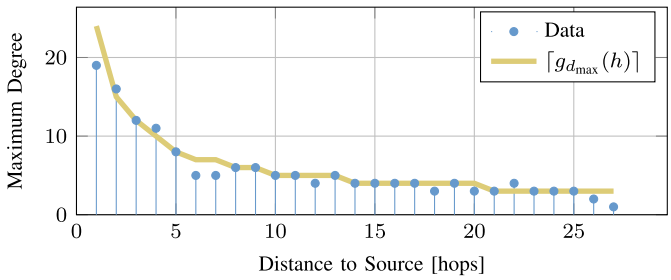


Fig. 9. Maximum degree at each hop distance along with a power law fit.

with characteristic function,

$$\phi_x(t; x_0, \gamma) = e^{jx_0 t - \gamma|t|}, \tag{14}$$

exhibits such a decay in magnitude. We therefore try to fit the data with a modified Cauchy distribution,

$$f(x; x_0, \gamma) = \left[\arctan\left(\frac{x_0}{\gamma}\right) + \frac{\pi}{2}\right]^{-1} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right] \tag{15}$$

where $x_0 \in \mathbb{R}$, $\gamma > 0$, and the modification cuts the support of the distribution from the real line to $x > 0$. Figure 8a shows very good fit to the data.

The CABLE LENGTH procedure, thus simply assigns length by sampling from (15). Since this is a heavy tailed distribution, extreme values will inevitably occur. However, there is a physical limit to how long a particular branch can be, which is addressed by function $g_{\ell_{\max}}(h)$. As already mentioned, a few of these clipping functions are necessary either due to physical, or common engineering practice constraints.

V. CLIPPING DISTRIBUTIONS

Most of the distributions in Section IV have either the whole real line or the positive real line as support. Since several of them are heavy tailed distributions, extreme values occur at non-negligible frequencies. However, from fundamental engineering principles certain situations do not make physical sense. For example, constraints on acceptable voltage drop limit the length of a distribution conductor, given the nominal voltage. Therefore, for several of the distributions, bounds are needed to restrict the range returned when sampling. All of these bounds are expressed in terms of the node's hop distance, $n.h$, from the source. In this way we again leverage the graph description of the feeder to identify trends in physical node and edge properties.

A. Maximum Degree

From the basic design principles of a distribution feeder, we expect branching occurrences to diminish as the distance from the primary substation increases [38]. The maximum degree for each hop level in the dataset, shown in Figure 9, exhibits this trend, which is fit by a power law function,

$$g_{d_{\max}}(h) = a \cdot h^b, \tag{16}$$

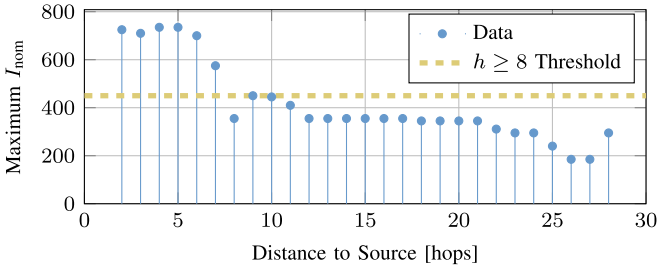


Fig. 10. Maximum I_{nom} for cables at each hop distance from the source. The dotted line is the threshold chosen for nodes at $h \geq 8$.

where h, is the hop distance, and a and b are fit to minimize squared error. The specific fit parameters can be found in Table III.

During degree assignment in the CONNECT NODES procedure, the bimodal Gamma is sampled until the the result is less than or equal to $\lceil g_{d_{\max}}(h) \rceil$. In this way excessive degrees further down the feeder are avoided.

B. Maximum Nominal Current

Since the Exponential distribution for $I_{\text{est}}/I_{\text{nom}}$ places a high weight on very low ratios, it is possible that very high I_{nom} will be assigned in the CABLE TYPE procedure. However, as Figure 10 shows, the largest cables are not used beyond several hops away from the source. This is, if nothing else, an economics issue, since larger capacity cables are much more expensive. Therefore, a threshold is picked that for hop distances, $h \geq 8$, the nominal current is $I_{\text{nom}} \leq 450$A.

Procedure CABLE TYPE does not explicitly show this threshold due to clarity considerations. In implementation, however, if the threshold is exceeded, a new sample is drawn from the Exponential distribution. While the threshold is currently a scalar, it could be expanded to a step function if finer control is desired.

C. Length Maximum

Given the heavy tails of the modified Cauchy distribution, physically unrealizable lengths are occasionally drawn in procedure CABLE LENGTH. Figure 11a shows the maximum length at each hop distance, h, as well as an exponential fit,

$$g_{\ell_{\max}}(h) = a \cdot e^{b \cdot h}. \tag{17}$$

Since the data falls on both sides of $g_{\ell_{\max}}(h)$, we further consider what errors are made by using the function instead of the empirical data. Our goal is to not overly constrain the algorithm. That is, we do not want to force a cable to be much shorter than it could be. Figure 11b plots two different error functions. The first shows the percentage of cables that are longer than the value returned by (17) at each hop distance,

$$g_1(h) = \frac{\sum_{m=1}^{M} \mathbb{1}\left(m.\ell > g_{\ell_{\max}}(m.h) \cap m.h = h\right)}{\sum_{m=1}^{M} \mathbb{1}(m.h = h)}. \tag{18}$$

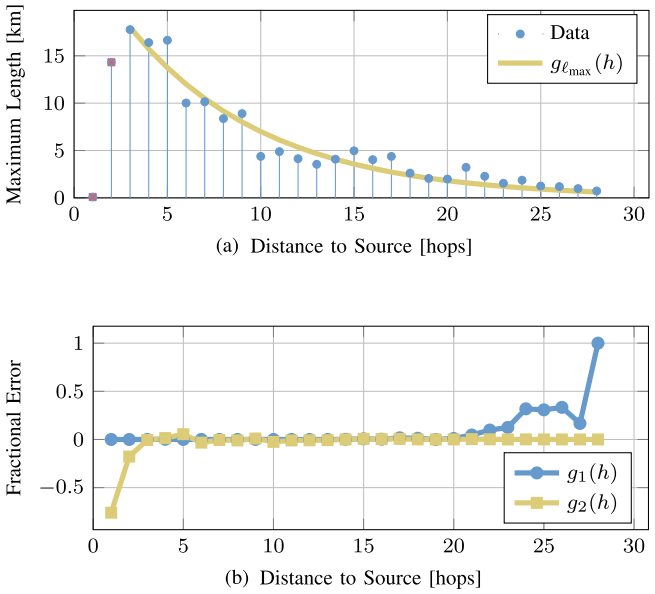


Fig. 11. Clipping function for length assignment. (a) Maximum cable length at each hop distance an Exponential fit to the data. (b) Analysis of the Exponential fit to maximum length.

The second shows the maximum percent error in length with respect to (17),

$$g_2(h) = \frac{\max_m \mathbb{1}(m.h = h) m.\ell - g_{\ell_{\max}}(h)}{g_{\ell_{\max}}(h)}. \tag{19}$$

These two tests reveal that when the percent error is large, Equation (19), the percent of cables that are *longer* than the maximum $g_{\ell_{\max}}(h)$, is negligible, $g_1(h) \approx 0$. Alternatively, as the value of Equation (18) increases, meaning there are more cables that are longer than the maximum returned by $g_{\ell_{\max}}(h)$, the percent error in length is negligible, $g_2(h) \approx 0$. Therefore, we conclude that (17) is a good bounding function for the length assignment. In the CABLE LENGTH procedure, the modified Cauchy distribution is sampled for each cable m, until the result falls below $g_{\ell_{\max}}(m.h)$.

D. The Effect of Clipping

From the modeling perspective, applying these bounds is akin to applying a condition to the distributions, from $f(x)$ to $f(x|x < x_{max}(h))$. The effect of such conditioning is to redistribute the weight from outside the constrained domain, to the domain, depending on parameter h. Another way of saying this is that there is a relationship between the support of the distribution and h. In the case of the degree distribution, the influence is fairly minimal, since so much is already dictated by the radial assumption. For example, the average degree is fixed to $2 - 2/N$.

Consider the weighted adjacency matrix A, where the nodes have been sorted based on $n.h$. The effect of clipping the degree based on h is to shift more of the non-zero entries of A to the upper rows. For edge properties, such as length, the clipping function is somewhat like a diagonal matrix with decreasing values that multiplies A. Note that clipping effects were trends observed in the real data.

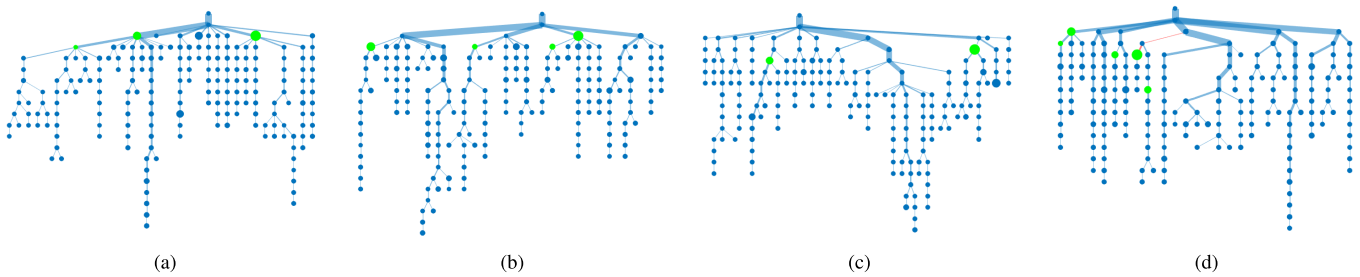
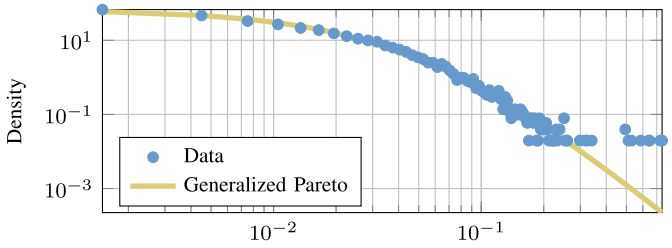
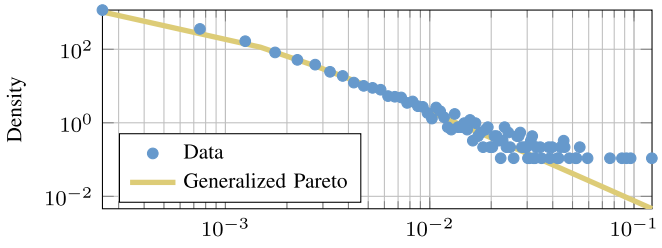


Fig. 12. Three samples generated with the following inputs: N: 195, Load: 23MVA, and Generation: 3MVA. The width of each line represents the relative real power flow magnitude. Edges with reverse flow are marked in red. The size of each node represents the relative magnitude of real load/injection. Injection nodes are identified with green. The fourth feeder is a real feeder from the data set with the same N, Load and Generation. The real feeder is identified at https://sine.fulton.asu.edu/~eran/RealAndSynthetic.pdf.



(a) Normalized Downstream Power



(b) Voltage Drop [p.u.]

Fig. 13. Two additional distributions, not explicitly used in synthesis, that are used to validate the effectiveness of the generation algorithm. (a) Downstream power distribution with generalized Pareto fit line. (b) Per unit voltage drop distribution with generalized Pareto fit line.

VI. Results

To test the algorithm, data from one of the real feeders is used to generate some samples. Figure 12 shows three generated samples as well as the real feeder from the dataset. As a fun exercise, we encourage the reader to try and pick out the real feeder before inspecting the solution provided online at: https://sine.fulton.asu.edu/~eran/RealAndSynthetic.pdf.

In addition to visual comparison of individual samples, 427 synthetic samples are generated to observe the cumulative statistics. Input parameters to the algorithm are drawn from a three dimensional Kernel Density Estimate (KDE) for the data vector (N, Load, Generation). The input variables should therefore, be similar to the dataset. Slices from the KDE for fixed generation are shown in Figure 15a.

Using the cumulative dataset, the distributions identified in Section IV can be evaluated. Because our intent is to create synthetic data that reproduce the real behavior of distribution feeders, we are interested in going beyond the statistics that were directly fit for synthesis. Therefore, in addition to the distributions introduced in Section IV, which are used in synthesis, two additional distributions are considered. As we show next, these distributions naturally emerge with the same trends observed in the data, and further validate the algorithm's ability to synthesize realistic distribution system feeders. The emergence of statistical behavior for edge and node properties is the main validation of our work.

The first new trend is the downstream power distribution (cf. Figure 13a). We define downstream power of given node, $n_i.P_{\text{downstream}}$, as the sum of all real power that must flow past this node to reach its destination. The same concept can be applied to reactive and apparent power, as well as to branches. Since the feeder is radial, downstream power is simply,

$$n_i.P_{\text{downstream}} = \sum_{n_j \leftarrow n_i} n_j.P, \tag{20}$$

where $n_j \leftarrow n_i$ is used to denote a node n_j that is a downstream node of node n_i. For example, the HV source has downstream power equal to the sum of all loads minus generation in the feeder. The quantity, whose histogram is plotted in Figure 13a, is a normalization of downstream power by the total load in the feeder. Each node in this distribution is highly dependent on the others, which is the main reason why this distribution is not used directly in the algorithm.

The second emergent distribution considered is the estimated voltage drop magnitude over a cable, expressed as a fraction of the nominal voltage. This can be calculated using the estimated current and impedance of branch m:

$$m.\Delta V = \frac{\|m.I_{\text{est}}\| \cdot \|m.Z\|}{m.V_{\text{nom}}}, \tag{21}$$

where $m.I_{\text{est}}$ is as in (12) and $m.Z$ is calculated using the per distance cable data and length, $m.\ell$.

Both the downstream power, Figure 13a, and the voltage drop, Figure 13b, distributions are fit by a Generalized Pareto distribution:

$$f(x;k,\sigma,\theta) = \frac{1}{\sigma}\left(1 + k \cdot \frac{x-\theta}{\sigma}\right)^{-1-\frac{1}{k}}, \tag{22}$$

where, $x > \theta$, and $k > 0$. The KL-Divergence of both fits is reported in Table II.

Figure 14 plots the various distributions from the synthetic data along with the original functions fit to the real data.

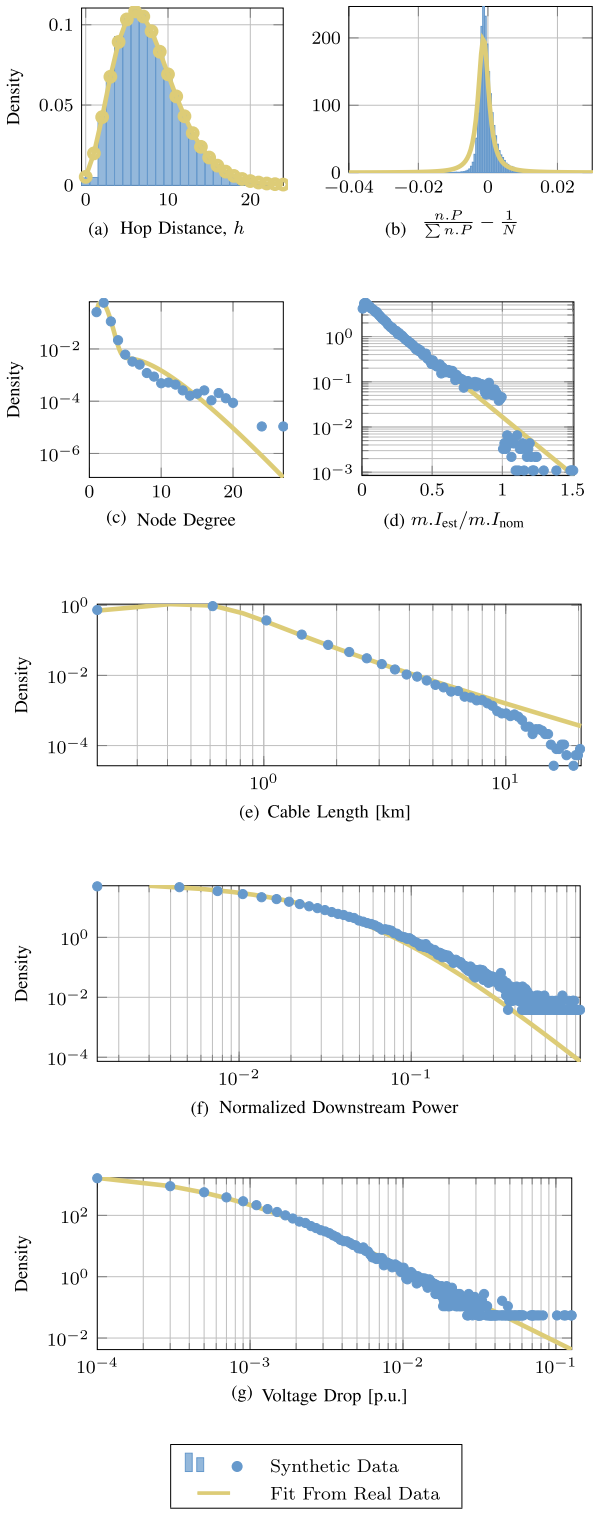


Fig. 14. Results from the generated synthetic samples using the inputs in Figure 15b. The quantities in (a)–(g) are identified on their x-axes.

Visual inspection suggests fairly good matches, including for the emergent downstream power, Figure 14f, and the voltage drop, Figure 14g, distributions. The KL-Divergence for each sample is reported in Table II and the relatively low values further help to indicate a good match.

A second set of feeders is created with the load for each input in Figure 15b doubled. Two illuminating results are

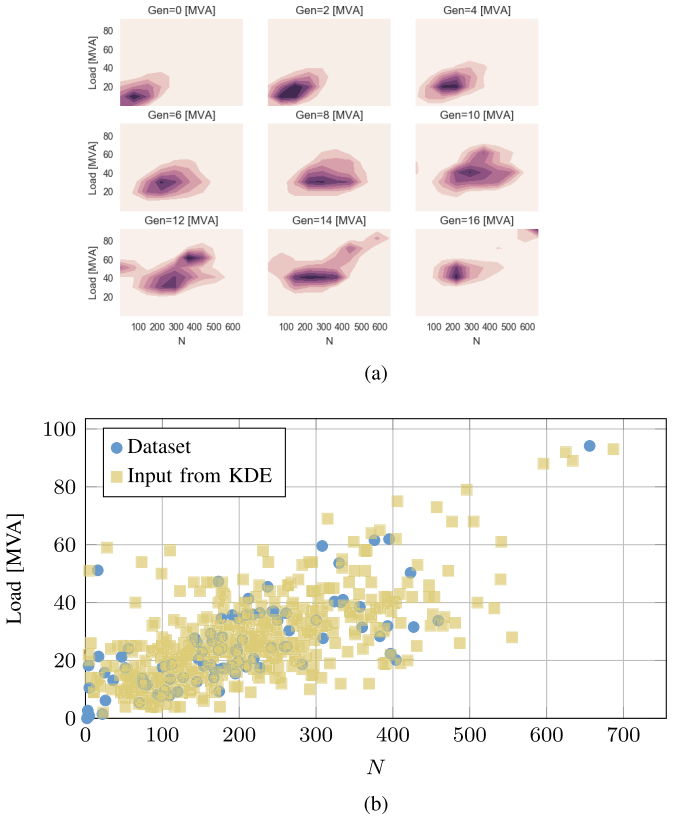


Fig. 15. A Kernel Density Estimate (KDE) is created for input vector (N, load, generation) and then sampled. Slices of the KDE at fixed generation levels are shown in (a). The 427 samples from this KDE used as inputs to generate the synthetic dataset are shown along with the original data points in (b). (a) Kernel Density Estimate (KDE). (b) Inputs to algorithm sampled from (a), as well as original data points.

shown in Figure 16. Because the input vectors are now further separated from the actual data, the ensemble contains a larger concentration of extreme cases. As a result, some emergent distributions diverge more strongly from their expected trend. If the load on a given feeder were to double we would expect more heavily loaded conductors and larger voltage drops, exactly as seen in Figure 16, where the data lies further above the expected trend line than in Figure 14. Correspondingly, the KL divergence between the empirical distribution and expected trends has increased by an order of magnitude.

VII. Discussion

The results in Section VI show that the algorithm is capable of reproducing feeders that exhibit the distributions across various properties seen in the data. The i_{est}/i_{nom}, downstream power, and voltage drop distributions are closely linked to the results of a power flow calculation. Therefore, by meeting these distributions, we argue the feeders behave realistically, at least from the steady-state power flow perspective. Of course there are many other measures of realistic behavior that we did not test, that will be the subject of future investigations.

Some deviation is to be expected, especially since, as can be seen in Figure 15b the inputs span a somewhat larger space than the real data. However, the fact that a distribution like the branch voltage drop emerges naturally suggest that the complex interactions are in fact correctly calibrated.

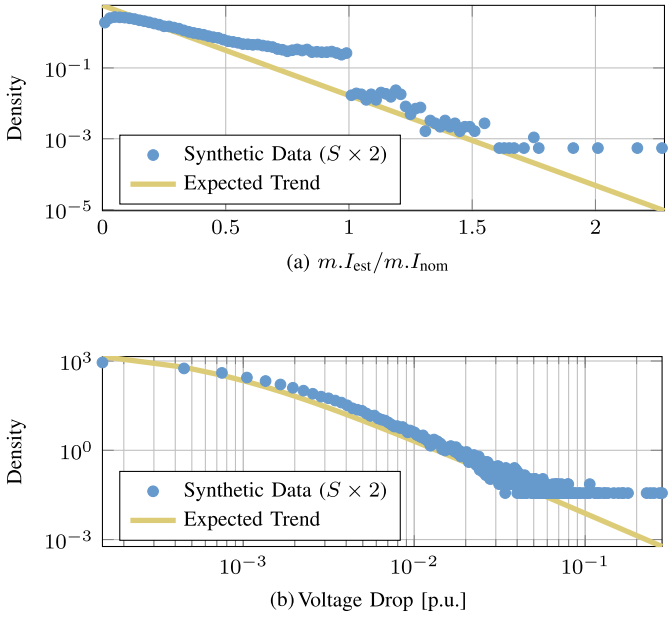


Fig. 16. As the input vectors to the algorithm are more distant from those seen in the dataset, certain properties begin to diverge from the expected trend. This can be observed numerically in an increasing D_{KL}: 0.21 for $m.I_{\text{est}}/m.I_{\text{nom}}$ instead of 0.01 obtained with the original inputs, and 0.13 instead of 0.02 for voltage drop. (a) Distribution of $m.I_{\text{est}}/m.I_{\text{nom}}$ when Load input is doubled. (b) Distribution of branch voltage drops when Load input is doubled.

The effects of extreme cases is further investigated by doubling the input load. We notice that the KL-Divergence for some of the distributions increases by an order of magnitude, suggesting it can be used to measure how extreme a case is. It is perhaps not correct to say that the cases resulting in Figure 16 are "wrong". Instead, they are more extreme than normal. Therefore, the model both generates the feeders and provides a measure of how "normal" they are via the KL distances.

There are several notable omissions in the model that require clarification. Transformers are currently omitted since each feeder has only one transformer between the source and the root and thus plays a fairly minimal role. For powerflow considerations, the HV node can be omitted and the MV bus set as the swing bus. In future work, when multiple voltage levels are added, and multiple feeders are connected, transformers will be included through a library approach similar to cables.

Capacitors are also omitted, primarily because none are in the dataset. The utility explained that due to the dominance of underground cables and their natural capacitance, additional devices are not needed. As we expand analysis to feeders from different locations, where capacitors are heavily used, these devices will clearly have to be incorporated.

A. Generality of the Model

While the previous section shows that the algorithm returns ensembles of feeders that are similar to those in the dataset, one might question the generality of the approach to other datasets. First of all, we note again that the data covers over 8000km^2, so it is reasonable to assume that this utility is not insular and the results immediately translate to other parts of at least northern Europe.

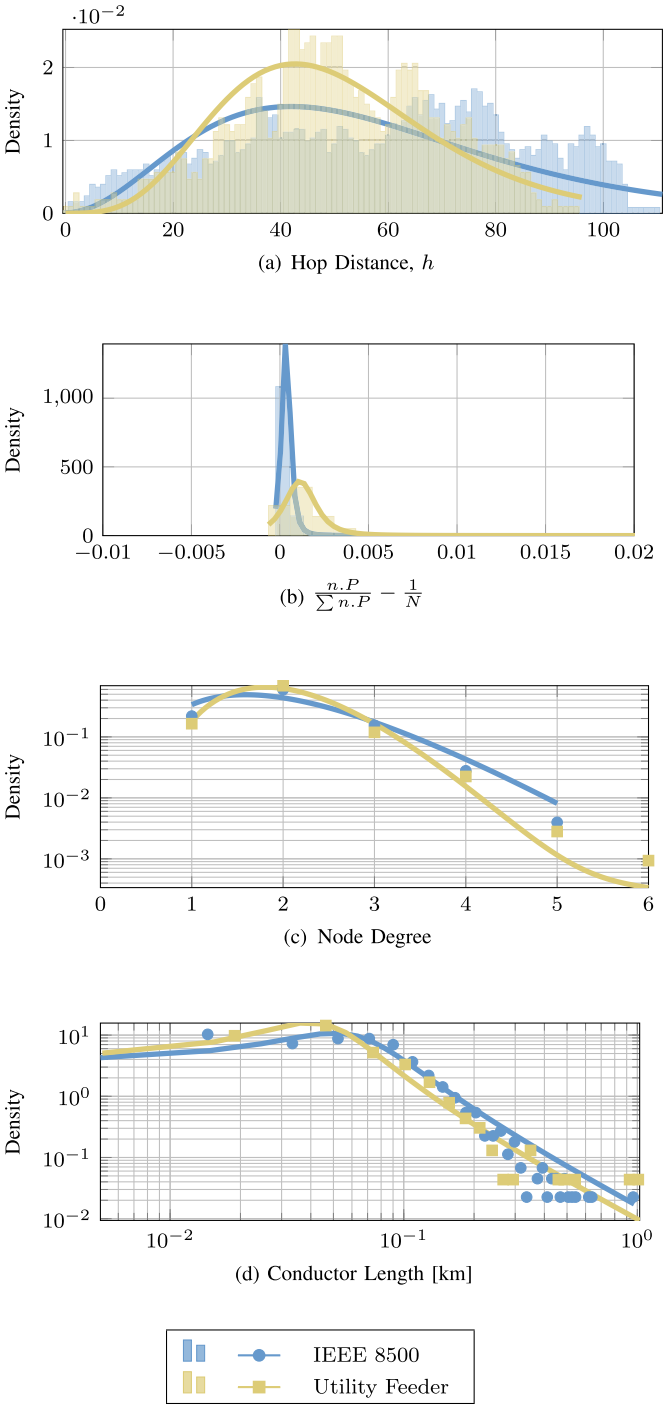


Fig. 17. Preliminary analysis of two distribution feeders from the U.S. numerical values are given in Table V.

For a more detailed comparison, other high quality datasets are needed. We include in Figure 17 preliminary analysis of two feeders from the U.S., the IEEE 8500 bus test feeder [11], and one feeder we were able to obtain from a utility in California. With the exception of h, at first blush the other trends seem to fit, albeit with different parameters.

The maximum h is significantly greater than in Figure 2, which raises a question. Combined with the shorter conductor lengths, we suspect there are modeling differences at play. Additional evidence is the fraction of intermediate nodes, which is over 50% in comparison to less than 10% for most

TABLE V
RESULTS OF PRELIMINARY U.S. FEEDER ANALYSIS

Property	Parameter Values	D_{KL}
Hop Distance		
IEEE 8500	$r = 3.75$, $p = 0.06$	0.1667
Utility Feeder	$r = 6.73$, $p = 0.12$	0.1158
Load Deviation From Uniform		
IEEE 8500	$\mu = 3.5 \times 10^{-4}$, $\sigma = 2.5 \times 10^{-4}$, $\nu = 2.5$	0.0988
Utility Feeder	$\mu = 1.1 \times 10^{-3}$, $\sigma = 9.2 \times 10^{-4}$, $\nu = 2.7$	0.1846
Degree Distribution		
IEEE 8500	$p = 0.99$, $a_1 = 5.09$, $b_1 = 0.39$, $a_2 = 1.6 \times 10^{-9}$, $b_2 = 5.2 \times 10^{-3}$	0.0558
Utility Feeder	$p = 0.996$, $a_1 = 9.95$, $b_1 = 0.20$, $a_2 = 1.62$, $b_2 = 3.1$	0.0304
Conductor Length		
IEEE 8500	$x_0 = 0.0505$, $\gamma = 0.0369$	0.1063
Utility Feeder	$x_0 = 0.0402$, $\gamma = 0.0236$	0.0630

of the Netherland feeders. Further investigation is needed, however, we argue that the appearance of similar statistical laws supports our claim that the methodology can be extended more generally. In fact, the variation in parameters could even help in classifying systems. For example, one could note the difference in conductor lengths between the European and American samples by noting that γ is an order of magnitude smaller.

VIII. CONCLUSION & FUTURE WORK

An automated algorithm for generating synthetic distribution feeders is presented. The synthetic dataset shows good statistical compliance with trends identified in a real distribution system. Furthermore, increasing KL-Divergence signals input deviation form "normal."

Synthetic feeders could find applications in areas such as:

- More nuanced algorithm (eg. state estimation) evaluation via Monte Carlo testing.
- Providing a system behind the common point of coupling for hardware in the loop simulations. This enhances testbeds and introduces uncertainty in test conditions, thus discouraging "for the test" designs.
- Provide fast access to new cases when none are available. In this context a utility could apply our approach and provide the distributions to generate synthetic samples.

In all applications, pairing the synthetic generation algorithm with specific platforms and simulation software used for analysis will be critical. We therefore believe the implementation intricacies of specific applications merit their own treatment and we will address them in future work.

While the majority of distribution systems are radially operated [13], the system as a whole is not, due to normally open branches that enable reconfiguration. This is evident by the discrepancy between total branches in Table I and the total number of branches in the final set of feeders, 20 903. Our future work will study where open branches are found by considering topology, such as the cycle distribution, as well as typical power engineering practices. The analysis will enable connection of several feeders into a full distribution system.

Differences between the North American and European systems are often mentioned in passing [13], [37]. The distributions in Section IV reflect the Netherlands distribution system structure, but the algorithm's modular approach enables altering these while maintaining the construction logic. Our future efforts will extend the North American results.

REFERENCES

[1] J. Northcote-Green and R. G. Wilson, *Control and Automation of Electrical Power Distribution Systems*, vol. 28. Boca Raton, FL, USA: CRC Press, 2006.

[2] G. T. Heydt, "The next generation of power distribution systems," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 225–235, Dec. 2010.

[3] M. Pau, P. A. Pegoraro, and S. Sulis, "Efficient branch-current-based distribution system state estimation including synchronized measurements," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 9, pp. 2419–2429, Sep. 2013.

[4] S. Hossain, H. Zhu, and T. Overbye, "Distribution fault location using wide-area voltage magnitude measurements," in *Proc. North Amer. Power Symp. (NAPS)*, Sep. 2013, pp. 1–5.

[5] P. Jamborsalamati, A. Sadu, F. Ponci, and A. Monti, "Implementation of an agent based distributed FLISR algorithm using IEC 61850 in active distribution grids," in *Proc. Int. Conf. Renew. Energy Res. Appl. (ICRERA)*, Nov. 2015, pp. 606–611.

[6] H. Zhu and H. J. Liu, "Fast local voltage control under limited reactive power: Optimality and stability analysis," *IEEE Trans. Power Syst.*, vol. 31, no. 5, pp. 3794–3803, Sep. 2016.

[7] 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, Jun. 2010.

[8] 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. Power Energy Conf. Illinois (PECI)*, Urbana, IL, USA, Feb. 2016, pp. 1–6.

[9] *IEEE PES Distribution Test Feeders*, accessed on Sep. 29, 2016. [Online]. Available: http://ewh.ieee.org/soc/pes/dsacom/testfeeders/index.html

[10] W. H. Kersting, "Radial distribution test feeders," in *Proc. IEEE Power Eng. Soc. Winter Meeting*, vol. 2. Jan./Feb. 2001, pp. 908–912.

[11] R. F. Arritt and R. C. Dugan, "The IEEE 8500-node test feeder," in *Proc. IEEE PES Transmiss. Distrib. Conf. Expo.*, Apr. 2010, pp. 1–6.

[12] K. P. Schneider, Y. Chen, D. P. Chassin, R. Pratt, D. Engel, and S. Thompson, "Modern grid initiative distribution taxonomy final report," Pacific Northwest Nat. Lab., Richland, WA, USA, Tech. Rep. PNNL-18035, 2008.

[13] T. A. Short, *Electric Power Distribution Handbook*. Boca Raton, FL, USA: CRC Press, 2003.

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

[15] P. Hines, S. Blumsack, E. C. Sanchez, and C. Barrows, "The topological and electrical structure of power grids," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2010, pp. 1–10.

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

[17] R. Albert, I. Albert, and G. L. Nakarado, "Structural vulnerability of the North American power grid," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, p. 025103, Feb. 2004.

[18] G. A. Pagani and M. Aiello, "The power grid as a complex network: A survey," *Phys. A, Statist. Mech. Appl.*, vol. 392, no. 11, pp. 2688–2700, 2013.

[19] G. A. Pagani and M. Aiello, "Towards decentralization: A topological investigation of the medium and low voltage grids," *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 538–547, Sep. 2011.
[20] C. D. Brummitt, P. D. H. Hines, I. Dobson, C. Moore, and R. M. D'Souza, "Transdisciplinary electric power grid science," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 30, p. 12159, 2013.
[21] M. Rosas-Casals *et al.*, "Knowing power grids and understanding complexity science," *Int. J. Critical Infrastruct.*, vol. 11, no. 1, pp. 4–14, 2015.
[22] J. Hu, L. Sankar, and D. J. Mir, "Cluster-and-connect: An algorithmic approach to generating synthetic electric power network graphs," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep./Oct. 2015, pp. 223–230.
[23] 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," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1502–1510, Mar. 2016.
[24] R. Kadavil, T. M. Hansen, and S. Suryanarayanan, "An algorithmic approach for creating diverse stochastic feeder datasets for power systems co-simulations," in *Proc. IEEE Power Energy Soc. General Meeting*, Jul. 2016, pp. 1–5.
[25] E. Schweitzer, K. Togawa, T. Schloesser, and A. Monti, "A MATLAB GUI for the generation of distribution grid models," in *Proc. ETG-Fachbericht-Int. ETG Congr.*, 2015, pp. 1–6.
[26] B. Cloteaux, "Limits in modeling power grid topology," in *Proc. IEEE 2nd Netw. Sci. Workshop (NSW)*, Apr. 2013, pp. 16–22.
[27] D. Deka, S. Vishwanath, and R. Baldick, "Analytical models for power networks: The case of the Western US and ERCOT grids," *IEEE Trans. Smart Grid*, to be published.
[28] N. Rotering, C. Schröders, J. Kellermann, and A. Moser, "Medium-voltage network planning with optimized power factor control of distributed generators," in *Proc. IEEE Power Energy Soc. General Meeting*, Jul. 2011, pp. 1–8.
[29] 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 Trans. Power Del.*, vol. 21, no. 2, pp. 995–1005, Apr. 2006.
[30] J. Kepner and J. Gilbert, *Graph Algorithms in the Language of Linear Algebra* (Software, Environments, and Tools). Philadelphia, PA, USA: SIAM, 2011.
[31] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.
[32] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Stochastic Modelling and Applied Probability). Berlin, Germany: Springer, 2009.
[33] T. G. Lewis, *Network Science: Theory and Applications*. Hoboken, NJ, USA: Wiley, 2009.
[34] Z. Wang, A. Scaglione, and R. J. Thomas, "The node degree distribution in power grid and its topology robustness under random and selective node removals," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, May 2010, pp. 1–5.
[35] 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 Syst. J.*, vol. 6, no. 4, pp. 616–626, Dec. 2012.
[36] D. Deka and S. Vishwanath, "Generative growth model for power grids," in *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Dec. 2013, pp. 591–598.
[37] W. H. Kersting, *Distribution System Modeling and Analysis*. Boca Raton, FL, USA: CRC Press, 2012.
[38] J. Dickert, M. Domagk, and P. Schegner, "Benchmark low voltage distribution networks based on cluster analysis of actual grid properties," in *Proc. IEEE Grenoble PowerTech (POWERTECH)*, Jun. 2013, pp. 1–6.

Eran Schweitzer received the M.Sc. degree in electrical power engineering from RWTH Aachen University in 2015. He is currently pursuing the Ph.D. degree in electrical engineering with the Prof. Scaglione's SINE Laboratory, Arizona State University.

His current research is focused on how statistical methods can be used to enhance and improve power systems research and analysis.

Anna Scaglione (F'11) was with the University of California at Davis, Cornell University, and the University of New Mexico. She is currently a Professor in electrical and computer engineering with Arizona State University. Her research focuses on various applications of signal processing in network and data science that include intelligent infrastructure for energy delivery and information systems.

Antonello Monti received his M.Sc degree (*summa cum laude)* and his PhD in Electrical Engineering from Politecnico di Milano, Italy in 1989 and 1994 respectively. He started his career in Ansaldo Industria and then moved in 1995 to Politecnico di Milano as Assistant Professor. Between 2000 and 2008 he was an Associate and then Full professor at the University of South Carolina. Since 2008 he is the director of the Institute for Automation of Complex Power System within the E.ON Energy Research Center at RWTH Aachen University.

Giuliano Andrea Pagani received the Ph.D. degree in computer science from the University of Groningen, Groningen, The Netherlands, in 2014. He has been a Postdoctoral Researcher at the University of Groningen and IBM Research T.J. Watson in 2014-2015 and worked for the Dutch DSO Alliander N.V. in 2015–2016.

He is currently a Data Scientist at the Royal Netherlands Meteorological Institute and is an external Postdoctoral Researcher with the University of Groningen.