Flexible
Elektrische
Netze

Final Project Report

Tool Development For Distribution Grid Topology Generation

Conducted by

Univ.-Prof. Antonello Monti
Institute for Automation of Complex Power Systems
at RWTH Aachen University

In collaboration with
Eran Schweitzer, M.Sc.

Mathieustr. 10
52074 Aachen

post_acs@eonerc.rwth-aachen.de

Tel.: +49 241 80 49700
Fax: +49 241 80 49709

Contents

Executive Summary

Whether introducing more automation, or investigating a switch to DC operation, the distribution system is currently the focus of very intensive investigation. All these investigations require increasingly more input data to build up the necessary test cases, which serve as the foundations of power systems simulations. Due to lack of publicly available data, as well as stemming from a desire to facilitate ensemble level testing, the goal of this project is to develop an algorithm that will automatically generate the distribution systems (medium voltage specifically) test cases. These test cases are expected to greatly aid future investigations by enabling far more robust testing and verification of new applications.

This report outlines the analysis that was performed on a large dataset from a DSO in the Netherlands to better understand the structure of distribution systems. Building on the analysis, an algorithmic approach is developed to meet the same characteristics. The latter part of the report presents results demonstrating that the algorithm performs well, as well as an extension of the algorithm from feeders to full distribution systems.

1 Background and Objectives of the Study

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]:

- self-healing from disturbances,
- enabling prosumers,
- 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. Our approach begins with radial feeders, and then connects them with normally open branches. A large majority of distribution feeders are radial, or at

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

least operated radially [13], justifying the initial focus on this topology.

1.1 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, 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, such as distribution feeders. We show that certain properties, such as branch voltage drops or power flows, which can be seen as various edge weights, emerge naturally from the synthesis process without explicit control from the algorithm.

Fully, or at least largely, automated generation of synthetic power systems is not entirely new, with several examples in [7, 22, 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 effort, and our previous work, which focused on topology [25], we are unaware of any 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, 27]. However, we found that the topology oriented geometric embedding applied very restrictive constraints to matching reasonable electrical parameters. This experience motivated the approach taken here, to link the topology and electric parameters in the modeling from the beginning.

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

The novelty in our methodology is that it enables 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.

2 Overview of Methodology

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 1.

The data was provided in several `.vnf` files, the proprietary format of the Vision software from Phase2Phase[1]. From there, the data was exported to Excel and imported to a PostgreSQL[2] database for easier manipulation.

Table 1: 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

[1] `http://www.phasetophase.nl/en_products/vision_network_analysis.html`
[2] https://www.postgresql.org/

2.1 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, $G(V,E)$[3], 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 1.

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[4]. These "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.

2.2 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.

[3] 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.

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

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.

2.3 Verification Methodology

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

$$D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) \, \mathrm{d}x, \tag{2.1}$$

which is often used to characterize the distance between two distributions[5].

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 feed-

[5] 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].

ers, 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.

3 Data Analysis

In this section, we present the data analysis performed that is used to inform the synthesis algorithm. In each step, trends in the form of distributions are identified, which are later exploited for synthesis. Throughout, we try to 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.

3.1 Node Generation

The radial assumption lies at the foundation of the synthesis algorithm because it 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 2.1 is by definition one hop away from the source, which we denote as $n.h = 1$. Figure 1 shows the distribution of hop distances in the dataset as well as a fit line following the Negative Binomial distribution,

$$f(x;r,p) = \frac{\Gamma(r+x)}{x!\Gamma(r)} p^r (1-p)^x, \tag{3.1}$$

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 the report, is given in Table 2, and the values for the parameters of (3.1) are reported in Table 3.

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.

Table 2: KL-Divergences

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

[†] 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.

Table 3: 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	$a = 26.97$, $b = -0.13$

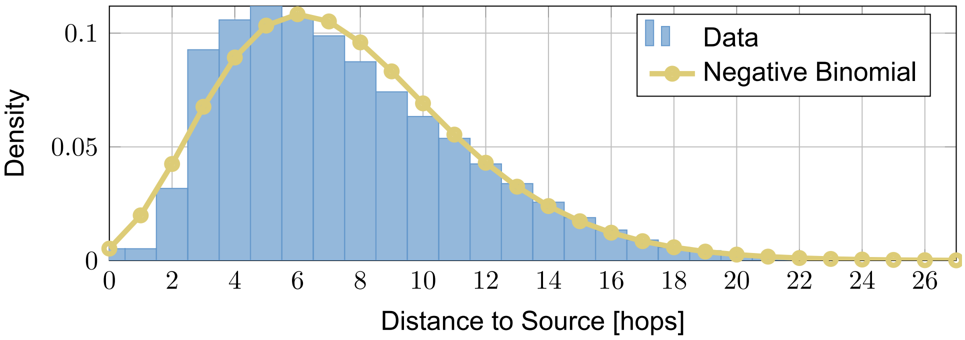


Figure 1: Distribution of hop distances and the Negative Binomial fit.

3.2 Feeder Connection

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],

$$P(k) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(n.d = k), \tag{3.2}$$

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,

$$f(x; p, a_1, b_1, a_2, b_2) = p \cdot g(x; a_1, b_1) + (1 - p) \cdot g(x; a_2, b_2) \tag{3.3}$$

with,

$$g(x; a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}, \tag{3.4}$$

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, 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 2, 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

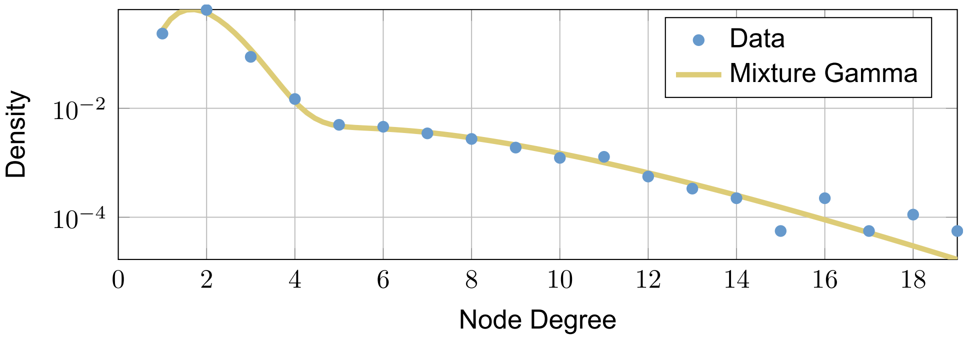


Figure 2: Degree distribution with a mixture of Gamma distributions fit line.

Table 4: Power Factor cdf

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

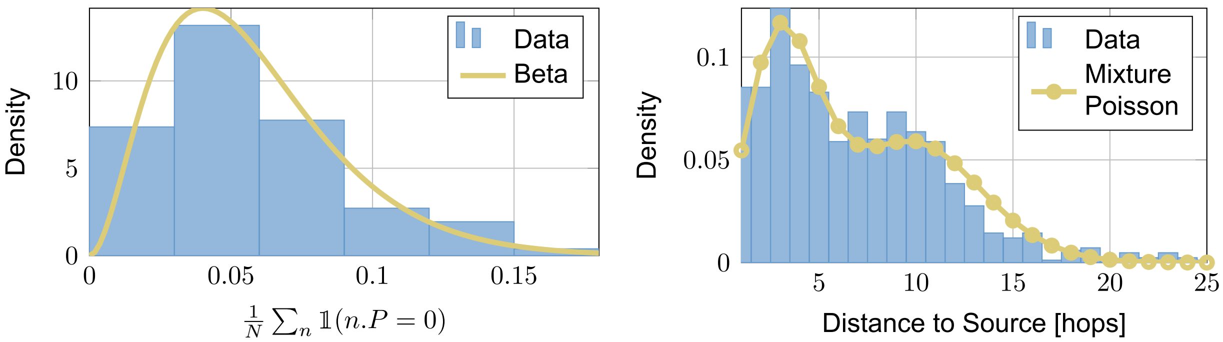
the findings in [36] that a sum of Exponential distributions provided a good fit to the degree distribution.

3.3 Node Properties

The node properties we consider are the powers associated with each node. Only the real power is considered, with the understanding that the reactive power is handled by a power factor distributed according to Table 4 . We identify three types of nodes: intermediate (no load), generation (negative load), consumption (positive load). Since the number of intermediate nodes and generation nodes is quite small, single feeder statistics are omitted in Table 2.

3.3.1 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. We consider the fraction of intermediate



(a) Fraction of intermediate (zero load) nodes

(b) Hop distance of intermediate nodes

Figure 3: Distributions for intermediate node assignment.

nodes per feeder and fit it with a Beta distribution as shown in Figure 3a. The Beta distribution,

$$f(x; , \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \tag{3.5}$$

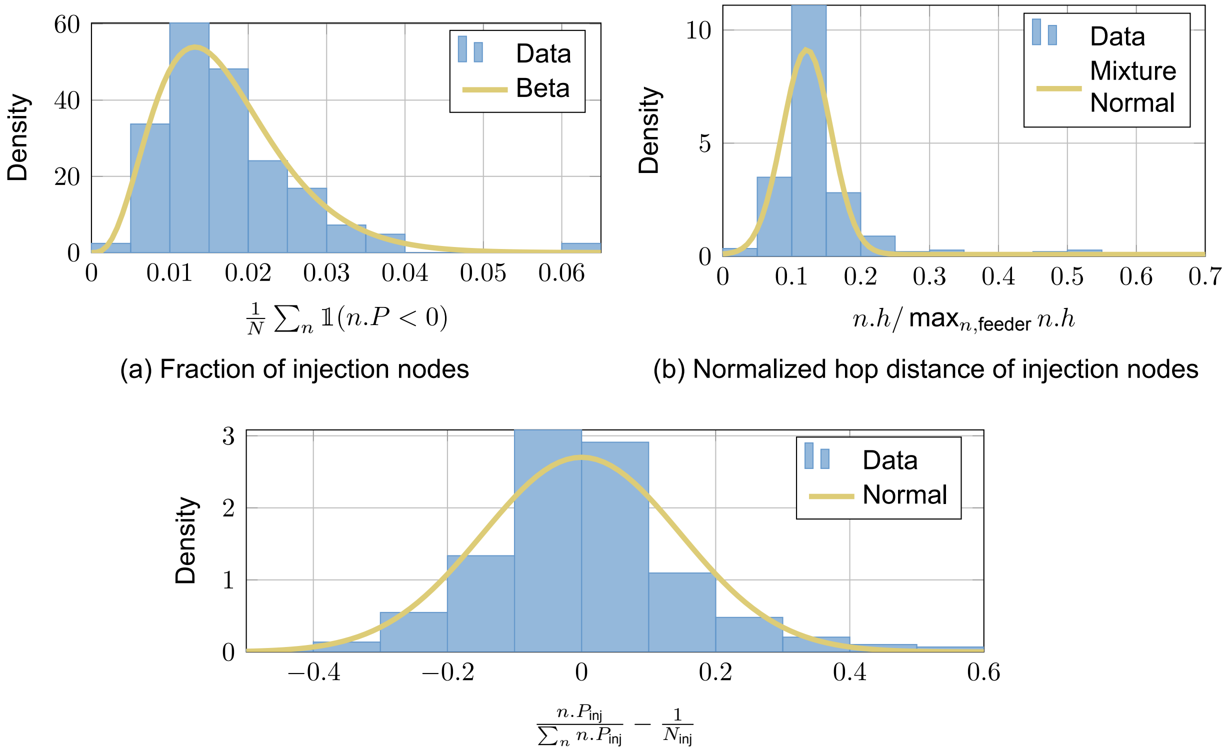
with $0 < x < 1$, and $\mathrm{B}(\cdot)$ the Beta function, is a common choice when modeling fractional quantities.

The source is designated to have zero load. For each of the remaining intermediate nodes, a sample is chosen from a mixture Poisson distribution, Next we consider how far the intermediate nodes are from the HV source in terms of hops. Figure 3b shows the histogram as well as a mixture Poisson distribution fit to the data. The mixture poisson is defined as,

$$f(x; p, \mu_1, \mu_2) = p\frac{\mu_1^x}{x!}e^{-\mu_1} + (1-p)\frac{\mu_2^x}{x!}e^{-\mu_2}, \tag{3.6}$$

where $x = 0, 1, \ldots, \infty$ and $\mu_{1,2} > 0$. 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[1]. This physical interpretation helps justify the mixture model, and the Poisson distribution is a natural choice for a random process on the integers.

[1] Secondary voltage levels are currently not implemented.



(a) Fraction of injection nodes

(b) Normalized hop distance of injection nodes

(c) Deviation in (3.8) of power injection from uniform distribution

Figure 4: Distributions for power injection assignment.

3.3.2 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, we consider the sum total and as such, are interested in nodes that have a net negative load.

Again, the fraction of injection nodes is analyzed and fit with a Beta distribution (cf. Figure 4a). Also similar to the intermediate nodes, we consider the hop distance for each injection node. However, however, here we normalize the hop distance by the maximum hop distance on the feeder. Figure 4b shows the histogram along with a mixture Normal distribution described as,

$$f(x; p, \mu_1, \sigma_1, \mu_2, \sigma_2) = p \cdot g(x; \mu_1, \sigma_1) + +(1 - p) \cdot g(x; \mu_2, \sigma_2), \tag{3.7}$$

where $0 \leq p \leq 1$, and $g(x; \mu, \sigma)$ is the normal distribution pdf with mean μ and standard deviation σ. The normalization was found to help to 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 and it is expected to become more frequent as distributed generation penetration increases.

Finally, Figure 4c 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",

$$\epsilon = \frac{n.P_{\text{inj}}}{\sum_n n.P_{\text{inj}}} - \frac{1}{N_{\text{inj}}} \tag{3.8}$$

is found to be normally distributed. This seems reasonable, suggesting that on a given feeder with power injections their magnitudes are somewhat consistent with one another.

3.3.3 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,

$$\epsilon = \frac{n.P}{\sum n.P} - \frac{1}{N}, \tag{3.9}$$

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

Figure 5 shows the histogram generated by (3.9), which is indeed tightly centered around zero. The t-Location-Scale distribution,

$$f(x;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}, \tag{3.10}$$

where $\sigma, \nu > 0$, is used to fit the data, which reflects the fact that the load is symmetrically distributed around the Uniform distribution, but with heavier tails. In fact, as can be seen from the parameters in Table 3, the distribution is close to being Cauchy, which

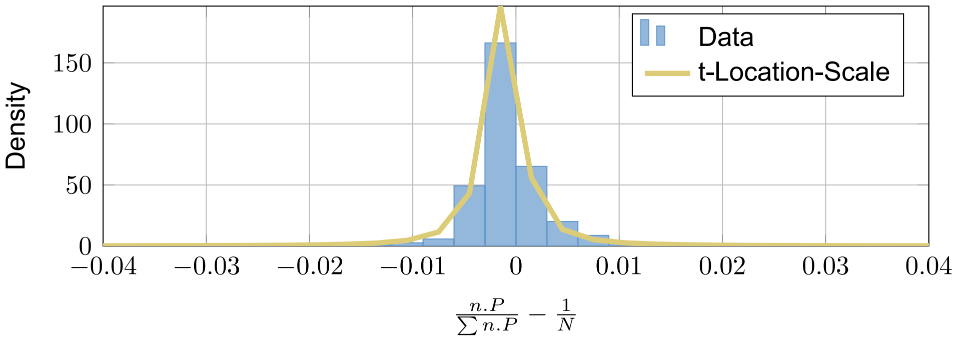


Figure 5: Histogram of the deviation in (3.9) of the load from the uniform distribution.

is the case when $\nu = 1$.

3.4 Cable Type

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

$$\|m.I_{\text{est}}\| = \frac{\|m.S_{\text{downstream}}\|}{\sqrt{3}\|m.V_{\text{nom}}\|}. \tag{3.11}$$

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

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu}, \tag{3.12}$$

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 6. 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 6. 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 (3.11) or the currents calculated from the powerflow in the Vision program. 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

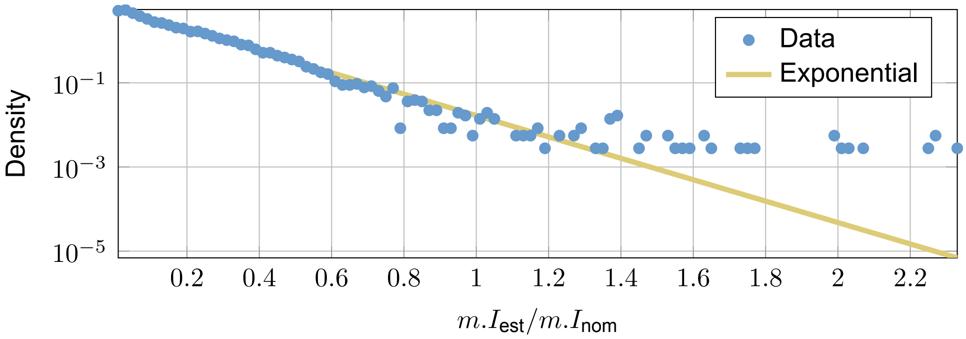


Figure 6: Exponential fit to the ratio of estimated current, I_{est}, and nominal cable current, I_{nom}.

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. This is addressed in Chapter 6.

3.5 Conductor Length

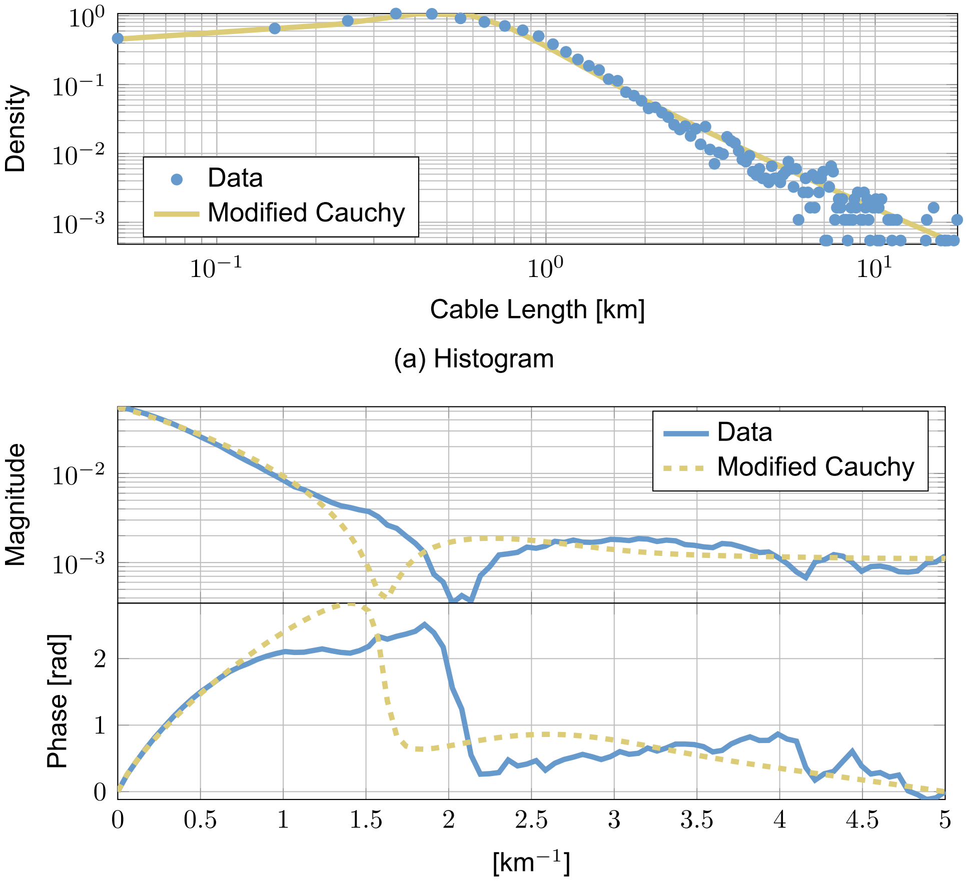
During the investigation of the length distribution, we observe a clear exponential decay in the magnitude of the empirical characteristic function—the Fourier transform of the histogram—which can be seen in Figure 7b. Considering common characteristic functions, only the Cauchy distribution with characteristic function,

$$\phi_x(t; x_0, \gamma) = e^{jx_0 t - \gamma |t|}, \tag{3.13}$$

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

$$f(x; x_0, \gamma) = \left[\text{arctan}\left(\frac{x_0}{\gamma}\right) + \frac{\pi}{2}\right]^{-1} \left[\frac{\gamma}{(x - x_0)^2 + \gamma^2}\right] \tag{3.14}$$

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 7a shows very good fit to the data.



(a) Histogram

(b) FFT of histogram representing the characteristic function

Figure 7: Distribution of cable lengths and modified Cauchy fit.

3.6 Clipping Distributions

Most of the distributions introduced thus far 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.

3.6.1 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 8, exhibits this trend, which is fit by a power law function,

$$g_{d_{\text{max}}}(h) = a \cdot h^b, \tag{3.15}$$

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 3. This function will be used in constraining the degree assigned to certain nodes.

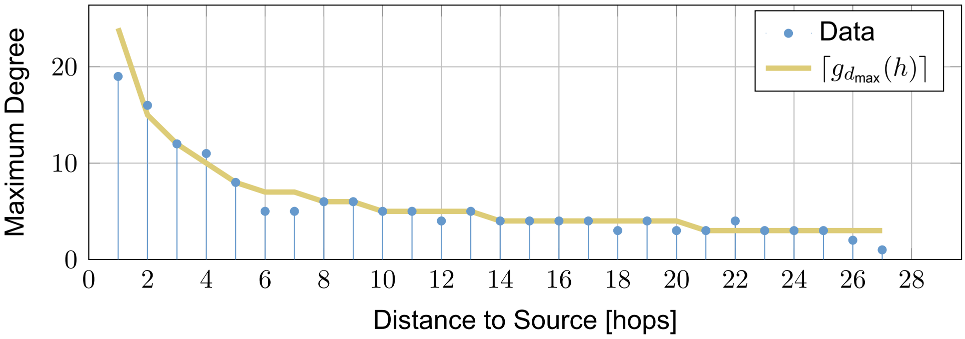


Figure 8: Maximum degree at each hop distance along with a power law fit

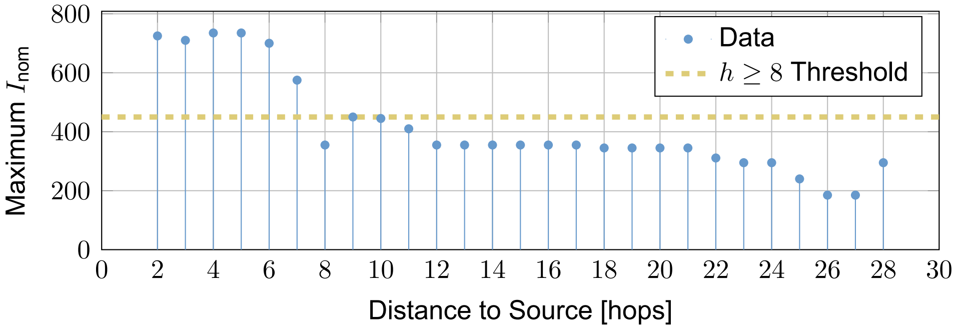


Figure 9: Maximum I_{nom} for cables at each hop distance from the source. The dotted line is the threshold chosen for nodes at $h \geq 8$.

3.6.2 Maximum Nominal Current

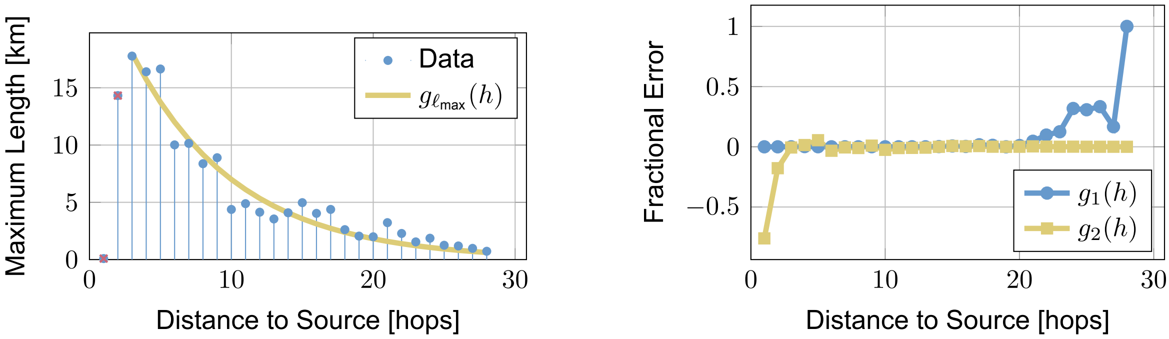
Since the Exponential distribution for $I_{\text{est}}/I_{\text{nom}}$ places a high weight on very low ratios, it is possible that sampling would result in very high I_{nom}. However, as Figure 9 shows, the largest cables are not used beyond several hops away from the source. This is, if nothing else, 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.

3.6.3 Length Maximum

Given the heavy tails of the modified Cauchy distribution, physically unrealizable lengths could be drawn. Figure 10a shows the maximum length at each hop distance, h, as well as an exponential fit,

$$g_{\ell_{\text{max}}}(h) = a \cdot e^{b \cdot h}. \tag{3.16}$$

Since the data falls on both sides of $g_{\ell_{\text{max}}}(h)$, 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 10b plots two different error functions. The first shows the percentage of



(a) Maximum cable length at each hop distance an Exponential fit to the data

(b) Analysis of the Exponential fit to maximum length

Figure 10: Clipping function for length assignment.

cables that are longer than the value returned by (3.16) at each hop distance,

$$g_1(h) = \frac{\sum_{m=1}^{M} \mathbb{1}\left(m.\ell > g_{\ell_{\text{max}}}(m.h) \cap m.h = h\right)}{\sum_{m=1}^{M} \mathbb{1}(m.h = h)}. \tag{3.17}$$

The second shows the maximum percent error in length with respect to (3.16),

$$g_2(h) = \frac{\max_m \mathbb{1}(m.h = h)m.\ell - g_{\ell_{\text{max}}}(h)}{g_{\ell_{\text{max}}}(h)}. \tag{3.18}$$

These two tests reveal that when the percent error is large, Equation (3.18), the percent of cables that are *longer* than the maximum $g_{\ell_{\text{max}}}(h)$, is negligible, $g_1(h) \approx 0$. Alternatively, as the value of Equation (3.17) increases, meaning there are more cables that are longer than the maximum returned by $g_{\ell_{\text{max}}}(h)$, the percent error in length is negligible, $g_2(h) \approx 0$. Therefore, we conclude that (3.16) is a good bounding function for the length assignment.

3.6.4 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.

3.7 Analysis Not Directly Used in Algorithm

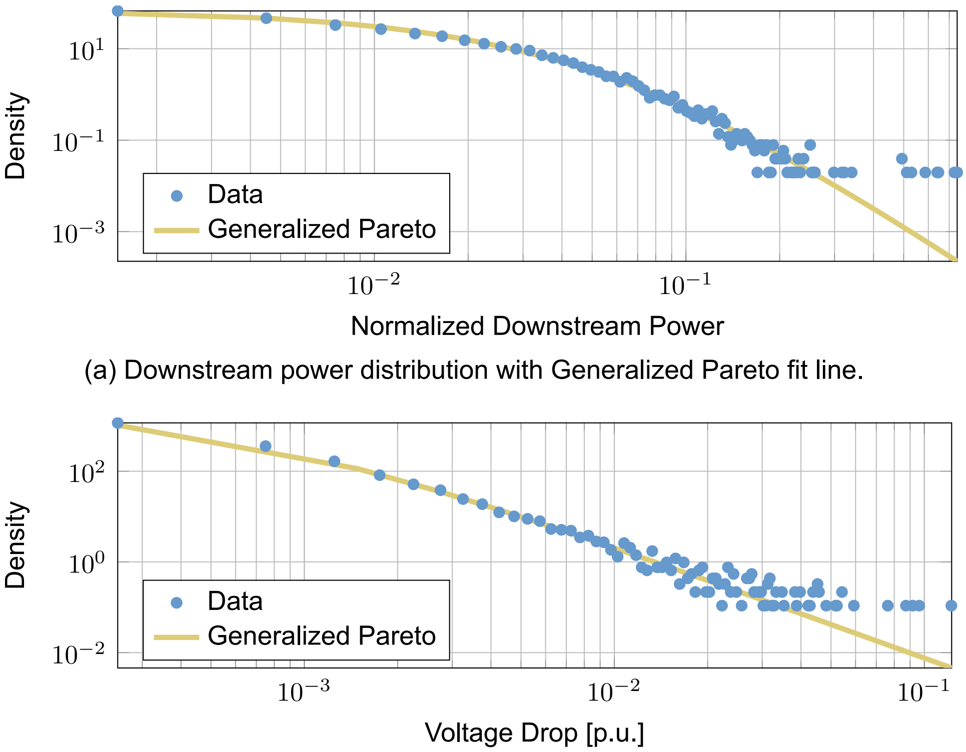
In addition to the previously introduced distributions, which, as will be shown in the next chapter, are used in synthesis, two additional distributions are considered. The natural emergence of the same trends observed in the data further validate the algorithm's ability to synthesize realistic distribution 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 11a). 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 as was done to estimate the current on a branch in Section 3.4. Since the feeder is radial, downstream power is simply,

$$n_i.P_{\text{downstream}} = \sum_{n_j \leftarrow n_i} n_j.P, \tag{3.19}$$

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 11a, 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 does not lend easily lend itself to be used in the synthesis 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



(a) Downstream power distribution with Generalized Pareto fit line.

(b) Per unit voltage drop distribution with Generalized Pareto fit line.

Figure 11: Two additional distributions, not explicitly used in synthesis, that are used to validate the effectiveness of the generation algorithm.

using the estimated current and impedance of branch m:

$$m.\Delta V = \frac{\|m.I_{\text{est}}\| \cdot \|m.Z\|}{m.V_{\text{nom}}}, \tag{3.20}$$

where $m.I_{\text{est}}$ is as in (3.11) and $m.Z$ is calculated using the per distance cable data and length, $m.\ell$.

Both the downstream power, Figure 11a, and the voltage drop, Figure 11b, distributions are fit by a Generalized Pareto distribution:

$$f(x; k, \sigma, \theta) = \frac{1}{\sigma}\left(1 + k \cdot \frac{x - \theta}{\sigma}\right)^{-1-\frac{1}{k}}, \tag{3.21}$$

where, $x > \theta$, and $k > 0$. The KL-Divergence of both fits is reported in Table 2.

4 Radial Feeder Algorithm

The statistical laws and limiting distributions from the previous section are put together to create the synthesis algorithm. Figure 12 provides an overview of the algorithm. The sections that follow mirror those from the analysis chapter and describe how the analysis is exploited in synthesis.

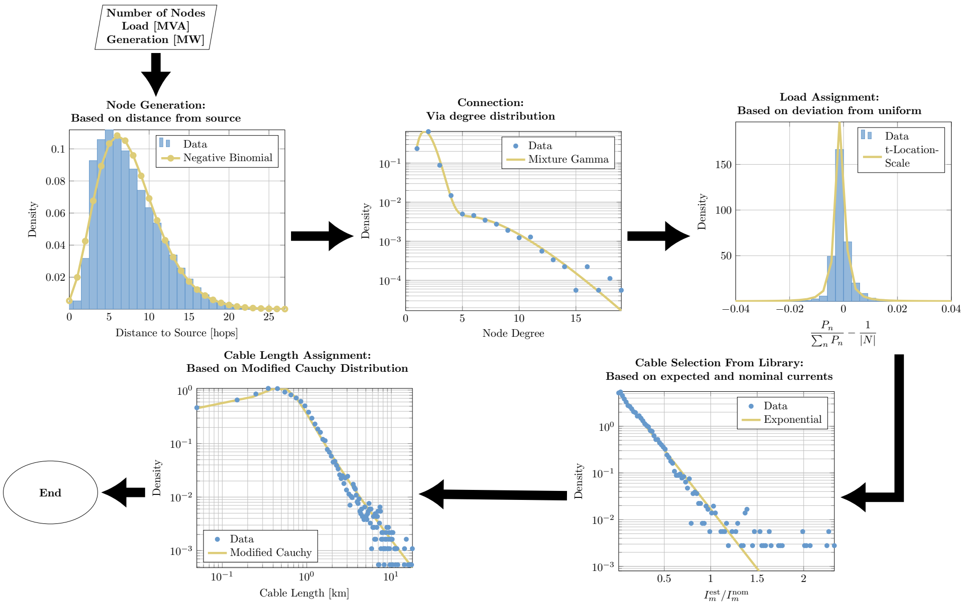


Figure 12: Flowchart for the radial feeder synthesis algorithm

4.1 Node Generation

Algorithm 1 uses samples from (3.1) 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 4. This greatly simplifies further manipulations, allowing to focus on real power.

Algorithm 1

```
procedure Generate Nodes(Power Factor cdf, Negative Binomial distribution)
    The first node is by design the source at $n.h = 0$
    The second node is by design the only node at $n.h = 1$
    for $n = 3, 4, \dots, N$ do
        $n.$power factor $\leftarrow$ power factor from input cdf
        $n.h \leftarrow$ sample from the Negative Binomial distribution
    Adjust hop distances so there are no gaps
```

Algorithm 2

```
procedure Connect Nodes(Mixture Gamma, $g_{d_{\text{max}}}(h)$)
    for all Nodes where $n.h = \text{max}_n\, n.h$ 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 $n = 1, 2, \dots, N$ do
        if No degree assigned then
            repeat
                $d_{\text{tmp}} \leftarrow$ sample from the mixture Gamma distribution
            until $d_{\text{tmp}} \leq \lceil g_{d_{\text{max}}}(n.h) \rceil$
            $n.d^{*} \leftarrow d_{\text{tmp}}$
    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:
                $n.\text{predecessor} \leftarrow \min_{p} p.d - p.d^{*}$
```

4.2 Connection Via Degree Distribution

Once the nodes have been created, Algorithm 2 is used to connect them into a tree rooted at the root (MV bus) and by extension the source (HV bus). 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 $g_{d_{\text{max}}}(h)$, described in Section 3.6.1. In this way excessive degrees further down the feeder are avoided.

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, $p.d^*$: $\min_p p.d - p.d^*$.

4.3 Node properties

Intermediate and injection nodes are treated as special due to their small number. For this reason they are handled first, after which, all the a remaining nodes are assigned load.

4.3.1 Intermediate Nodes

Algorithm 3

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

Intermediate nodes are marked first, so that load will not be assigned to them by the subsequent procedures. Algorithm 3 sets the number of zero load nodes, $N_{\text{intermediate}}$, by sampling a Beta distribution for the fraction of intermediate nodes.

Next, the HV source is designated as having zero load. For each of the remaining intermediate nodes, a sample is chosen from a mixture Poisson distribution, to determine at what hop distance the node should be.

4.3.2 Power Injections

Since the algorithm only produces the sum total of load and generation at a node, several nodes are picked in Algorithm 4, based on observations from the data, to have a net negative load. The number of injection nodes, N_{inj}, is determined using a ratio sampled from a Beta distribution and the hop distance for each injection node is then selected by sampling a mixture Normal distribution. Finally, real power injection is assigned by solving (3.8) for $n.P_{inj}$, where the while loop in the procedure is simply used to avoid sign reversals.

The Algorithm 4 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.

Algorithm 4

```
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 $1/N_{\text{inj}} + \epsilon > 0$
        $n.P \leftarrow -P_{\text{inj,total}} \left(1/N_{\text{inj}} + \epsilon\right)$
        $n.Q \leftarrow n.P \cdot \tan\left[\cos^{-1}(n.\text{power factor})\right]$
```

4.3.3 Load

After both no-load and injections have been assigned, the remaining nodes are assigned load in Algorithm 5. All the procedure does is solve (3.9) for $n.P$, generating ϵ by sampling the t-Location-Scale distribution. Once more, the while loop is used to avoid sign reversals.

Note that following all the assignments, a normalization is applied so that the actual

Algorithm 5

```
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 $1/N + \epsilon > 0$
            $n.P \leftarrow P_{\text{total}} \left(1/N + \epsilon\right)$
            $n.Q \leftarrow n.P \cdot \tan \left[\cos^{-1}(n.\text{power factor})\right]$
```

amount of load assigned and the input match.

4.4 Cable Selection

Algorithm 6

```
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 $I_{\text{nom}}$ 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 $I_{\text{nom}}$ attached to *upstream* node
            Pick cable from library with closest $I_{\text{nom}}$
```

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 Algorithm 6, 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 it 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 connected to the downstream node[1], in line with the finding that roughly two-thirds of nodes have only one type of cable incident upon them. 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.

Note that Algorithm 6 does not explicitly show the threshold from Section 3.6.2 for 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.

4.5 Conductor Length

Algorithm 7

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

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. Algorithm 7, thus simply assigns length by sampling from (3.14). 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

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

function $g_{\ell_{\mathsf{max}}}(h)$. The modified Cauchy distribution is sampled for each cable m, until the result falls below $g_{\ell_{\mathsf{max}}}(m.h)$.

5 Results

5.1 Individual Inspection

A visual test of the algorithm is done by using data from one of the real feeders to generate some samples. Figure 13 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`.

5.2 Ensemble Testing

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. Figure 14 shows a flow chart for the process used to create the ensemble of synthetic feeders. The scatter plots and KDE slices shown in the figure are from the real data, and we can see that in fact the inputs are similarly distributed.

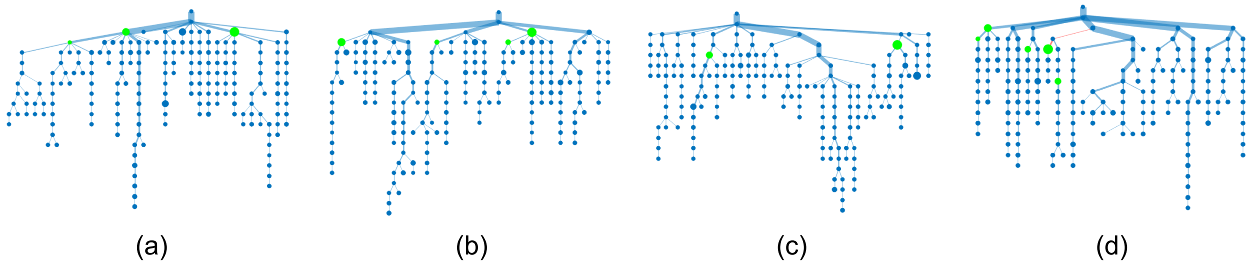


Figure 13: Three samples generated with the following inputs: N: 195, Load: 23 MVA, and Generation: 3 MVA. The width of each line represents the relative real power flow magnitude. Edges with reverse flow are marked in red. The size of each node represents the relative magnitude of real load/injection. 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`.

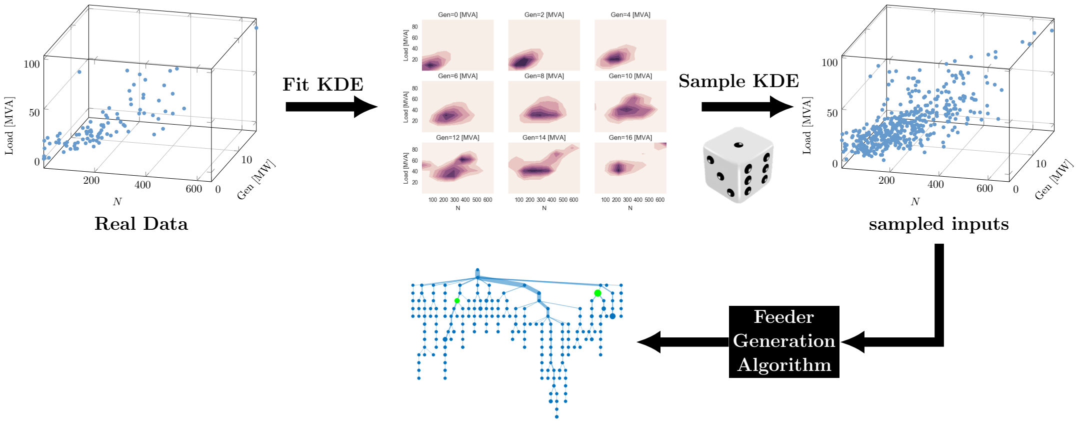
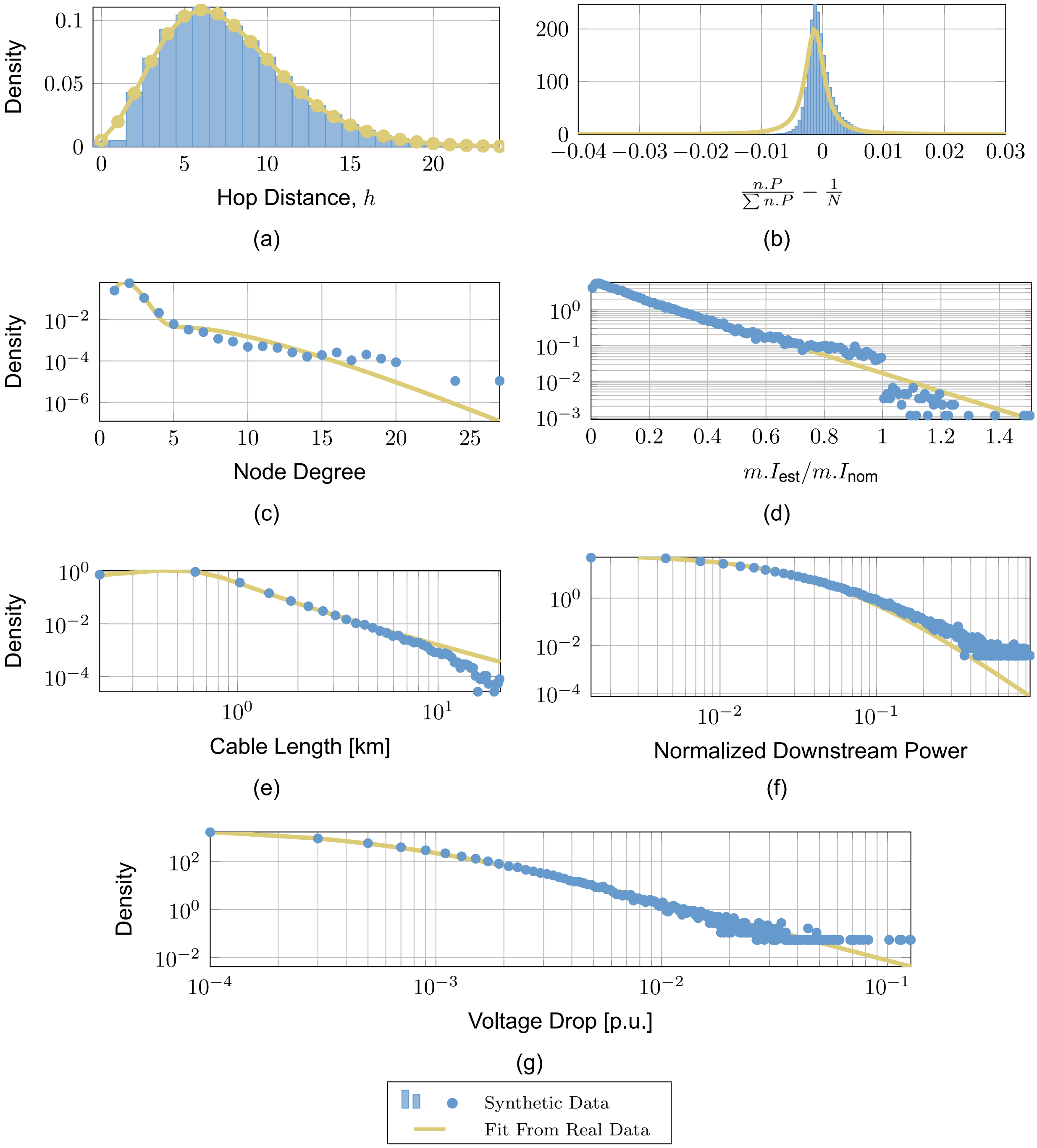


Figure 14: Procedure used to generate synthetic samples

Using the cumulative dataset, the distributions identified in Chapter 3 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 applied during synthesis. Therefore, we also consider how well the algorithm produces the distributions introduced in Section 3.7, which are not considered in the synthesis process. 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.

Figure 15 plots the various distributions from the synthetic data along with the original functions fit to the *real data*. Visual inspection suggests fairly good matches, including for the emergent downstream power, Figure 15f, and the voltage drop, Figure 15g, distributions. The KL-Divergence for each sample is reported in Table 5 and the relatively low values further help to indicate a good match.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

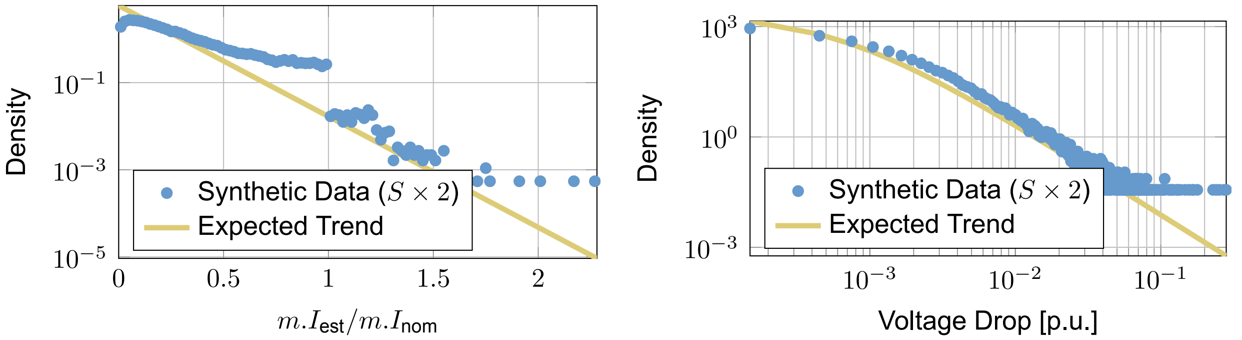
Figure 15: Results from the generated synthetic samples following the procedure illustrated in Figure 14.

Table 5: KL-Divergences Results Comparison

Property	Distribution	Real Cumulative D_{KL}	Synthetic Samples
Hop Distance	Negative Binomial	0.0173	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	0.1329
Degree Distribution	Bimodal Gamma	0.0211	0.0147
$I_{\text{est}}/I_{\text{nom}}$	Exponential	0.0098	0.0102
Cable Length	Modified Cauchy	0.0247	0.0108
Downstream Power	Generalized Pareto	0.0111	0.0243
Voltage Drop	Generalize Pareto	0.0917	0.0216

Overload Testing

Next, we consider the effect of the inputs on the output statistics. A second set of feeders is created from the same input data, except that the load is doubled. Two illuminating results are 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 15. Correspondingly, the KL divergence between the empirical distribution and expected trends has increased by roughly an order of magnitude.



(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

Figure 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.

6 Combining Feeders

Up to this point we have shown how radial feeders are created. In this chapter we address the question of how these can be joined to create a full distribution system.

6.1 Feeder Allocation

The feeder allocation problem groups feeders based on power per HV bus criteria. Figure 17c shows that there are either 1,2, or 3 feeders connected to a given source. This is consistent with Table 1.2 in [13] that lists 2 as a common value of transformers for a substation. Figure 17 also shows that the Rayleigh distribution is a reasonable fit at both the 1 and 2 feeder levels, to the total power delivered by the HV bus. The Rayleigh distributions is,

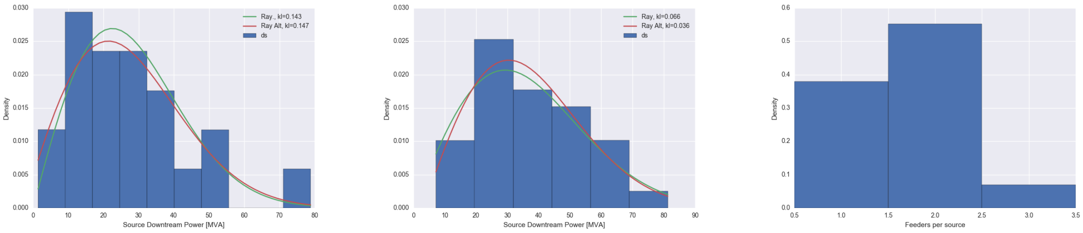
$$f(x;\mu,\sigma) = \frac{x-\mu}{\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)} \tag{6.1}$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$. Note that normally μ is not a parameter of the distribution and it is taken as 0.

Algorithm 8

```
procedure Feeder Allocation($\mathcal{P}$)
    repeat
        $M \leftarrow \{1,2,3\}$ Sampled from empirical distribution
        $C \leftarrow$ Rayleigh$(\mu_M, \sigma_M)$
        Group $M$ feeders from $\mathcal{P}$ as returned by optimization in (6.2)
    until $\mathcal{P} = \emptyset$
```

Algorithm 8 describes how feeder allocation is done on a set of feeders $\mathcal{P}$. The size of the group, M, is drawn from the empirical distribution in Figure 17c. Then the desired power, C, is sampled from the corresponding Rayleigh distribution, where the parameters for $M = 1,2$ are known (cf. Figure 17), and those for $M = 3$ are linearly extrapolated. The list of remaining feeders and inputs C,M are then given to the optimization problem in (6.2), which selects the optimal feeder group. The process is repeated until there are no feeders left unallocated.



(a) Sources with 1 feeder (b) Sources with 2 feeder (c) Feeders Per Source

Figure 17: Grouping statistics for feeders.

The optimization performed in Algorithm 8 simply tries to pick the combination of M feeders that whose power consumption most closely matches C.

$$\underset{t,\boldsymbol{u}}{\text{Minimize}} \quad t \tag{6.2a}$$

$$\text{Subject to} \quad \sum_i u_i = M \tag{6.2b}$$

$$\left| C - \sum_i P_i u_i \right| \leq t \tag{6.2c}$$

$$t \geq 0 \tag{6.2d}$$

$$\boldsymbol{u} \in \{0, 1\} \tag{6.2e}$$

6.2 Adding Normally Open Branches

The task remaining for completing the system topology is to add normally open branches (NOB). In doing so, the following quantities are considered:

1. H^{hop} hop distance assortativity: the two dimensional distribution of the hop distance for end nodes of NOBs. For example, there would be a $+1$ added to $H^{\text{hop}}_{5,5}$ for every NOB where both end nodes have a distance of 5 hops to the source node.

2. H^{deg} degree assortativity: the two dimensional distribution of the node degree for end nodes of NOBs. For example, there would be a $+1$ added to $H^{\text{deg}}_{1,1}$ for every NOB where both end nodes have a degree of 1 *in their feeder*. It is important to

note that the degree is now treated as a *fixed* number.

3. H_f fraction of normally open branches that are inter-feeder, i.e., both end nodes belong to the same feeder. That is $N_f = N_{\text{NOB inter-feeder}}/N_{\text{total NOB}}$.

4. H_s fraction of normally open branches that share the same source. These can be nodes on the same feeder *or* nodes on feeders that were grouped in Algorithm **??**.

5. H_{frac} average ratio of normally open to normally closed branches per feeder.

This problem greatly suffers from dimensionality, since it contains roughly N^2 binary variables, with N the total number of nodes in the system. If we consider something like the Netherlands dataset with $N \approx 20 \times 10^3$, that is around 400 million binary variables. To combat the dimensionality, we split the problem into parts. First inter-feeder NOBs are added and then intra-feeder ones, where in the intra-feeder step only a small subset of “neighbor” feeders is considered.

6.2.1 Inter feeder normally open branches

For each feeder the set of possible branches consists of those connected nodes that are not at hop 0 or 1 and whose upstream branch has the same nominal current. The nominal current requirement comes from observations in the data as well as some basic engineering judgment. The open branches are there to allow for reconfiguration, therefore, they are likely to handle load similar to the edges they are incident upon.

From this set, $\boldsymbol{u}$, the optimal selection is made based on optimization problem (6.5). The fraction of desired inter-feeder edges is

$$H_{\text{frac}} \times H_f \tag{6.3}$$

Therefore, the number of NOBs to add, N, for a given feeder with M closed edges is,

$$N = M \times H_{\text{frac}} \times H_f. \tag{6.4}$$

The optimization in (6.5) attempts to match the hop and degree distributions on an entry by entry basis through constraints (6.5b) and (6.5c), while getting as close to the desired number of NOBs through constraint (6.5d). Constraint (6.5e) only allows one NOB edge to be incident on any given node. Finally, weights are available in vector $\boldsymbol{w}$ to allow more or less influence to the different requirements.

$$
\begin{aligned}
&\underset{\boldsymbol{t},\boldsymbol{u}}{\text{Minimize}} \quad && \boldsymbol{w}^T\boldsymbol{t} && && \text{(6.5a)}\\
&\text{Subect to} && \left| H_{i,j}^{\text{hop}} - \frac{1}{N} \sum_{\substack{f=i\\t=j}} u_{f,t} \right| \leq t_{\text{hop}} && \forall i,j && \text{(6.5b)}\\
& && \left| H_{i,j}^{\text{deg}} - \frac{1}{N} \sum_{\substack{f=i\\t=j}} u_{f,t} \right| \leq t_{\text{deg}} && \forall i,j && \text{(6.5c)}\\
& && \left| N - \sum u \right| \leq t_N && && \text{(6.5d)}\\
& && \sum_j u_{i,j} \leq 1 && \forall i && \text{(6.5e)}\\
& && \boldsymbol{t} \geq 0 && && \text{(6.5f)}\\
& && \boldsymbol{u} \in \{0,1\} && && \text{(6.5g)}
\end{aligned}
$$

where $\boldsymbol{t} = \begin{bmatrix} t_{\text{hop}} & t_{\text{deg}} & t_N \end{bmatrix}^T$.

6.2.2 Intra feeder normally open branches

To complete the NOB assignment, branches are now added between feeders. For a given feeder with M closed branches and the input data stated in the beginning, the following values can be calculated:

$$
\begin{aligned}
N_{\text{tot}} &= M \times H_{\text{frac}} && \text{total number of desired NOBs}\\
N_s &= H_s \times N_{\text{tot}} && \text{number of NOBs with the same source}\\
N_f &= H_f \times N_{\text{tot}} && \text{number of NOBs on the same feeder}\\
N_{\text{same}} &= N_s - N_f && \text{number of NOBs with same source but different feeder}\\
N_{\text{diff}} &= N_{\text{tot}} - N_s && \text{number of NOBs with different sources}\\
N_{\text{add}} &= N_{\text{same}} + N_{\text{diff}} && \text{the number of new NOBs to add}
\end{aligned}
$$

For the intra-feeder problem the potential edges that qualify need to have *exactly one* node in the feeder in question, the hop distance must be greater than 1 and the upstream nominal must be the same for both end nodes much like in the inter-feeder problem. Additionally, only feeders that are $\pm D_{\text{max}}$ are considered, where in the implementation D_{max} is taken to be 4. This means that if feeder 10 is considered, the only nodes under consideration are those from feeders 6 through 14. Feeder distance is calculated with a modulus so that the highest numbered feeder is considered one away from feeder 0. We use η_n to represent the sets of edges connecting a given feeder to feeders that are distance n away. For example, if the edge node i belongs to feeder 4 and node j belongs to feeder 6 then $(i, j) \in \eta_2$, indicating that the edge connects two feeders that are a distance of 2 apart.

Constraints (6.6b), (6.6c), and (6.6f) work exactly the same way in this optimization problem as in the inter-feeder one. Constraints (6.6d) and (6.6e) do the same thing as (6.5d) except that in this problem there are two separate numbers we are trying to obtain. Finally, the number of edges added to different feeders is added to the constraint so we are able to steer a bit how these connections take place.

$$\underset{\boldsymbol{t},\boldsymbol{u}}{\text{Minimize}} \quad \boldsymbol{w}^T\boldsymbol{t} + \sum_{n=1}^{D_{\text{max}}} \left(w_n \sum_{(i,j)\in\eta_n} u_{i,j} \right) \tag{6.6a}$$

$$\text{Subect to} \quad \left| H_{i,j}^{\text{hop}} - \frac{1}{N_{\text{add}}} \sum_{\substack{f=i\\t=j}} u_{f,t} \right| \leq t_{\text{hop}} \qquad \forall i, j \tag{6.6b}$$

$$\left| H_{i,j}^{\text{deg}} - \frac{1}{N_{\text{add}}} \sum_{\substack{f=i\\t=j}} u_{f,t} \right| \leq t_{\text{deg}} \qquad \forall i, j \tag{6.6c}$$

$$\left| N_{\text{same}} - \sum_{s(i)=s(j)} u_{i,j} \right| \leq t_{\text{same}} \tag{6.6d}$$

$$\left| N_{\text{diff}} - \sum_{s(i)\neq s(j)} u_{i,j} \right| \leq t_{\text{diff}} \tag{6.6e}$$

$$\sum_j u_{i,j} \leq 1 \qquad \forall i \tag{6.6f}$$

$$\boldsymbol{t} \geq 0 \tag{6.6g}$$

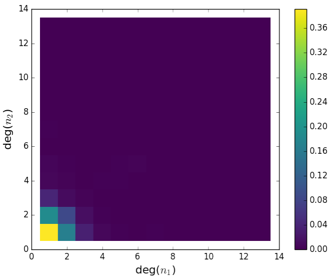
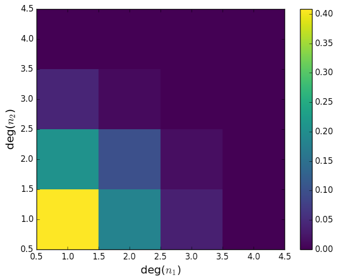
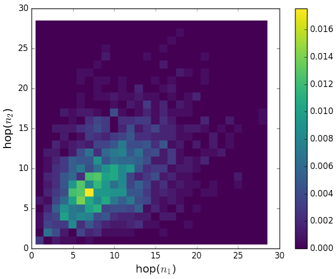
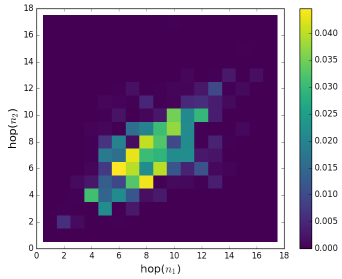
$$\boldsymbol{u} \in \{0, 1\} \tag{6.6h}$$

where $\boldsymbol{t} = \begin{bmatrix} t_{\text{hop}} & t_{\text{deg}} & t_{\text{same}} & t_{\text{diff}} \end{bmatrix}^T$.

6.3 Results

The above procedures were applied to the feeders generated in Chapter 5. Table 6 shows a comparison of the input data described at the beginning of the chapter for the real data and the resulting synthetic system. The results demonstrate that from the perspective of these metrics a valid, large scale, distribution system was created.

Table 6: Real vs. Synthetic statistics for connecting feeders

Feature	Real	Synthetic
Percent of NOB with both end nodes belonging to same feeder	77%	77%
Percent of NOB with both end nodes supplied by same HV bus	84%	83%
Average $\dfrac{\text{NOB}}{\text{closed branches}}$ per feeder	10%	15%
Node degree assortativity of NOB		
Hop assortativity of NOB		

7 Conclusion and Outlook

This report has detailed the steps from analysis through implementation of an algorithm to generate large and realistic MV distribution systems. Results show that the generated feeders closely agree with the statistics found in real feeders, supporting the claim of realism. The approach taken focuses on combining statistics seen in real data. A benefit of this approach is its modularity. Different statistics can be trivially plugged in to maintain the construction logic but drastically alter the final result. Additionally, empirical distributions could be employed in cases where particularly close adherence to a single system is desired.

It is our belief that at least initially DC systems will not deviate greatly from the structure of current AC systems. In fact, even many of the cables considered for DC systems are actually AC cables. For this reason the algorithm is applicable to both DC and AC applications provided that the right component libraries (cables and transformers for AC, cables and converters for DC) are provided.

The benefit these test systems will provide will depend greatly on how they are dovetailed into other simulation applications. For this reason we decided that a stand-alone GUI for this algorithm would not be useful and are instead focusing future efforts on pairing the output of the synthesis algorithm with various simulation platforms.

Over 800 synthetic test feeders were discussed just in this report (427×2 since the overloaded set was also created). The ease of generating large number of test cases promises to be very valuable for testing as this algorithm is paired with simulation platforms. Monte Carlo simulations are used in many fields where many uncertainties are chained together and are not easy to characterize analytically. The sort of large numbers available by automated test case generation open the door for such testing regimes. The evaluation process for new applications can therefore move from being a binary—works or does not work—to how well does it work.

Bibliography

[1] J. Northcote-Green and R. G. Wilson, *Control and automation of electrical power distribution systems*. CRC Press, 2006, vol. 28.

[2] G. T. Heydt, “The next generation of power distribution systems”, *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 225–235, 2010, issn: 1949-3053. doi: 10.1109/TSG.2010.2080328.

[3] M. Pau, P. A. Pegoraro, and S. Sulis, “Efficient branch-current-based distribution system state estimation including synchronized measurements”, *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 9, pp. 2419–2429, 2013, issn: 0018-9456. doi: 10.1109/TIM.2013.2272397.

[4] S. Hossain, H. Zhu, and T. Overbye, “Distribution fault location using wide-area voltage magnitude measurements”, in *North American Power Symposium (NAPS), 2013*, 2013, pp. 1–5. doi: 10.1109/NAPS.2013.6666936.

[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 *2015 International Conference on Renewable Energy Research and Applications (ICRERA)*, 2015, pp. 606–611. doi: 10.1109/ICRERA.2015.7418485.

[6] H. Zhu and H. J. Liu, “Fast local voltage control under limited reactive power: Optimality and stability analysis”, *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 3794–3803, 2016, issn: 0885-8950. doi: 10.1109/TPWRS.2015.2504419.

[7] 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, 2010, issn: 1949-3053. doi: 10.1109/TSG.2010.2044814.

[8] K. Gegner, A. Birchfield, T. Xu, K. Shetye, and T. Overbye, “A methodology for the creation of geographically realistic synthetic power flow models”, in *Power and Energy Conference at Illinois (PECI), Urbana, IL*, 2016.

[9] (). Ieee pes distribution test feeders, [Online]. Available: http://ewh.ieee.org/soc/pes/dsacom/testfeeders/index.html.

[10] W. H. Kersting, “Radial distribution test feeders”, in *Power Engineering Society Winter Meeting, 2001. IEEE*, vol. 2, 2001, 908–912 vol.2. doi: 10.1109/PESW.2001.916993.

[11] R. Arritt and R. Dugan, “The ieee 8500-node test feeder”, in *Transmission and Distribution Conference and Exposition, 2010 IEEE PES*, 2010, pp. 1–6. doi: 10.1109/TDC.2010.5484381.

[12] Schneider KP, Y Chen, DP Chassin, RG Pratt, DW Engel, and SE Thompson, “Modern grid initiative distribution taxonomy final report”, Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. PNNL-18035, 2008.

[13] T. A. Short, *Electric Power Distribution Handbook*. 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, 1998.

[15] P. Hines, S. Blumsack, E. Cotilla Sanchez, and C. Barrows, “The topological and electrical structure of power grids”, in *43rd Hawaii International Conference on System Sciences (HICSS)*, 2010, pp. 1–10. doi: 10.1109/HICSS.2010.398.

[16] 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, issn: 0378-7796. doi: http://dx.doi.org/10.1016/j.epsr.2005.05.013.

[17] R. Albert, I. Albert, and G. L. Nakarado, “Structural vulnerability of the north american power grid”, *Phys. Rev. E*, vol. 69, p. 025 103, 2 2004. doi: 10.1103/PhysRevE.69.025103.

[18] G. A. Pagani and M. Aiello, “The power grid as a complex network: a survey”, *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 11, pp. 2688 –2700, 2013, issn: 0378-4371.

[19] ——, “Towards decentralization: a topological investigation of the medium and low voltage grids”, *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 538–547, 2011, issn: 1949-3053. doi: 10.1109/TSG.2011.2147810.

[20] 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. 12 159, 2013. doi: 10.1073/pnas.1309151110. eprint: http://www.pnas.org/content/110/30/12159.full.pdf.

[21] 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.

[22] 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)*, 2015, pp. 223–230. doi: 10.1109/ALLERTON.2015.7447008.

[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 Transactions on Power Systems*, vol. PP, no. 99, pp. 1–1, 2016, issn: 0885-8950. doi: 10.1109/TPWRS.2016.2586460.

[24] R. Kadavil, T. M. Hansen, and S. Suryanarayanan, "An algorithmic approach for creating diverse stochastic feeder datasets for power systems co-simulations", in *2016 IEEE Power and Energy Society General Meeting*, 2016.

[25] E. Schweitzer, K. Togawa, T. Schloesser, and A. Monti, "A matlab gui for the generation of distribution grid models", in *ETG-Fachbericht-International ETG Congress 2015*, VDE VERLAG GmbH, 2015.

[26] B. Cloteaux, "Limits in modeling power grid topology", in *Network Science Workshop (NSW), 2013 IEEE 2nd*, 2013, pp. 16–22. doi: 10.1109/NSW.2013.6609189.

[27] D. Deka, S. Vishwanath, and R. Baldick, "Analytical models for power networks:the case of the western us and ercot grids", *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2016, issn: 1949-3053. doi: 10.1109/TSG.2016.2540439.

[28] 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*, 2011, pp. 1–8. doi: 10.1109/PES.2011.6039661.

[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 Transactions on Power Delivery*, vol. 21, no. 2, pp. 995–1005, 2006, issn: 0885-8977. doi: 10.1109/TPWRD.2005.858779.

[30] J. Kepner and J. Gilbert, *Graph Algorithms in the Language of Linear Algebra*, ser. Software, Environments, Tools. Society for Industrial and Applied Mathematics, 2011, isbn: 9780898719918.

[31] S. Kullback and R. A. Leibler, “On information and sufficiency”, *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951. doi: 10.1214/aoms/1177729694.

[32] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, ser. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009, isbn: 9783642033117.

[33] T. G. Lewis, *Network Science: Theory and Applications*. Wiley Publishing, 2009, isbn: 0470331887, 9780470331880.

[34] Z. Wang, A. Scaglione, and R. Thomas, “The node degree distribution in power grid and its topology robustness under random and selective node removals”, in *2010 IEEE International Conference on Communications Workshops (ICC)*, 2010, pp. 1–5. doi: 10.1109/ICCW.2010.5503926.

[35] 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, 2012, issn: 1932-8184. doi: 10.1109/JSYST.2012.2183033.

[36] D. Deka and S. Vishwanath, “Generative growth model for power grids”, in *2013 International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, 2013, pp. 591–598. doi: 10.1109/SITIS.2013.97.

[37] W. H. Kersting, *Distribution system modeling and analysis*. 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 *2013 IEEE Grenoble PowerTech (POWERTECH)*, 2013, pp. 1–6.

List of Figures

List of Tables