Learning Probability Distributions over Georeferenced Distribution Grid Models

Domenico Tomaselli
Paul Sturberg
Michael Metzger
Siemens AG
Munich, Germany
{domenico.tomaselli, paul.stursberg, michael.metzger}@siemens.com

Florian Steinke
Energy Information Networks & Systems
Technical University of Darmstadt
Darmstadt, Germany
florian.steinke@eins.tu-darmstadt.de

***Abstract*—When grid planners design updates for existing infrastructure in power grids, they frequently encounter a lack of trustworthy and readily-usable digital grid models. This is especially the case at the low-voltage (LV) level. While the location of secondary substations and end-consumers is often known, the topology is less certain and cannot be uniquely estimated. This work proposes a probabilistic framework to efficiently sample possible georeferenced grid topologies. A parametric probability distribution assigns an exact probability to each possible, georeferenced grid topology using characteristic features. The parameters of the probability distribution can be learned from known exemplary grid topologies. A Markov chain Monte Carlo (MCMC) algorithm is then designed to sample from the learned distribution with low computational complexity, thereby enabling efficient statistical inference. The described steps are demonstrated for the probabilistic modeling of a LV distribution grid in Schutterwald, Germany.**

***Index Terms*—Grid Topology Identification, Exponential Family, MCMC, Statistical Inference**

I. Introduction

A. Motivation

The increasing prevalence of distribution grids featuring distributed renewable energy sources, electric vehicle (EV) charging units, and heat pumps often necessitates infrastructure upgrades [1]. Considering the long historical record of several distribution grids, grid planners frequently encounter a lack of trustworthy and readily-usable, digital models. This is especially the case at the LV level. As a result, appropriate grid models for the existing infrastructure must be reconstructed, often relying on sparse information. While knowledge about the secondary substations and the end-consumers might typically be available from asset management and billing systems, the topology of existing LV distribution grids is often less certain [2]. Moreover, LV distribution lines are frequently located underground, making available topology information difficult and expensive to verify [1].

Reconstructing an appropriate grid topology in this setting is a strongly under-determined problem and obtaining a single, trustworthy solution cannot be expected. An alternative, promising strategy involves adopting a probabilistic approach, explicitly representing the remaining uncertainty using a probability distribution over the possible grid topologies. This concept was first introduced in a recent work [2], showcasing its practical utility, e.g., for the overload analysis of secondary substations in response to the increasing adoption of residential EV charging units. While in [2] we used a fixed and ad-hoc randomized growth model for estimating the probability distribution, this work focuses instead on establishing a precise probabilistic model for georeferenced distribution grid topologies, learning the model's parameters based on a training dataset, and demonstrating how to perform efficient statistical inference using a specifically tailored MCMC algorithm. Efficient statistical inference enhances the practical applicability of the proposed probabilistic approach to also perform grid operation tasks, e.g., grid topology detection [3] and state estimation [4].

B. Related work

The literature on grid topology generators is extensive and can be classified into two categories. Generators for a given region of interest typically use geospatial information, e.g., the street configuration and the location of the end-consumers [5], [6], or voltage fingerprints extracted from advanced metering infrastructures [1], [3] or phasor measurement units [3]. Alternatively, there are generators that produce an ensemble of distinct grid topologies and demonstrate a realistic degree distribution [7], matching also additional topological [8], [9], [10] and electrical characteristics [8] of real-world distribution grids. The work in [11] bridges the gap between the two categories of grid topology generators by establishing a georeferenced grid topology over time that accurately reflects the degree distribution observed in real-world distribution grids.

Random distribution grid models are a subset of general random graph models that are, e.g., also used for modeling social networks [12] or biological molecules [13]. Note that this work investigates models where the graph topology is uncertain, and not the more common case where quantities defined on edges or links of a known graph topology are uncertain, see, e.g., the works on power flow calculations with graph neural networks [14] or probabilistic graphical models [15], [16]. Random

Submitted to the 23rd Power Systems Computation Conference (PSCC 2024).

graph models can be defined algorithmically, e.g, via uniform edge probabilities or preferential attachment rules [17], [18]. Such models are, however, typically not trainable to closely match a given set of examples.

Trainable models can be constructed within the framework of the exponential family [19], [20]. This parametric family of distributions offers significant flexibility and robust modeling capabilities, as it allows to select from a wide array of distinct distributions for each individual feature. Moreover, these distributions can be represented in an intuitive way, which is a key aspect in gaining the confidence of grid planners when using them to model existing grid topologies. In this work, we thus employ this approach and, to the best of our knowledge, adapt it for the first time to model distinct georeferenced grid topologies. Another viable way for constructing trainable probabilistic models for graphs involves using generative neural networks, see, e.g, [21], [10]. These models, however, typically demand a large number of training examples and are comparatively less intuitive.

C. Contributions

In this work, we address the problem of learning a well-defined probability distribution over feasible georeferenced grid topologies and leveraging the learned distribution for efficient statistical inference. To this end, we establish an exponential family framework that precisely assigns probability values to possible grid topologies. This is achieved by using characteristic features of LV distribution grids, e.g., the load factor of the secondary substations or the maximum feeder length. We introduce a Maximum Likelihood Estimation (MLE) approach leveraging known exemplary distribution grid topologies to estimate the parameters of the established exponential family model. Then, we propose a novel MCMC algorithm for sampling high-probability grid topologies. This algorithm is key for performing efficient statistical inference, i.e., calculating moments or quantiles of derived quantities of interest, e.g., the load of a secondary substation, across the sampled grid topologies. The MCMC algorithm is specifically tailored towards georeferenced grid topologies of LV distribution grids and has low computational complexity. The proposed framework, whose workflow is shown in Fig. 1, is finally demonstrated through simulations performed on a LV distribution grid in Schutterwald, Germany.

The remainder of the paper is structured as follows. The proposed framework, its training and inference procedures are introduced in Section II. The simulation experiments are presented and discussed in Section III. Section IV concludes the work.

II. Methodology

The proposed framework comprises three key components. Following the definition of the notation, first an exponential family approach for modeling probability distributions over georeferenced grid topologies within a region of interest is outlined. Second, a procedure for parameter estimation within this probabilistic approach is presented. Third, efficient statistical inference is enabled by introducing a novel MCMC algorithm for sampling georeferenced grid topologies with low computational complexity.

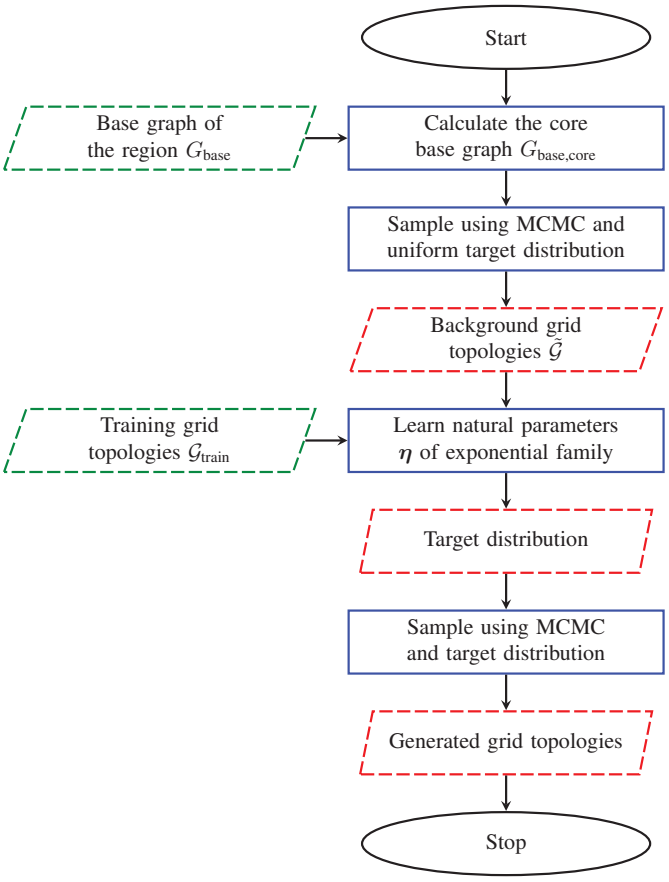

Fig. 1. Flowchart describing the procedural workflow of the proposed approach, emphasizing its inputs (- - -), outputs (- - -) and key processes (——).

A. Notation

A region of interest and a corresponding grid topology are schematically shown in Fig. 2 (a). Formally, the region of interest is represented as a quadruple $G_{\text{base}} = (V_{\text{base}}, E_{\text{base}}, P_l, V^s_{\text{base}})$, in the following referred to as base graph. G_{base} is assumed to be georeferenced such that, e.g., the lengths between nodes can be calculated. The N vertices in set V_{base} and the M edges in set $E_{\text{base}} \subseteq V_{\text{base}} \times V_{\text{base}}$ describe the underlying street configuration. The map $P_l : V_{\text{base}} \to \mathbb{R}$ returns the design load for each node and $V^l_{\text{base}} = \{v \in V_{\text{base}} | P_l(v) > 0\}$ denotes the set of end-consumer nodes. $V^s_{\text{base}} \subset V_{\text{base}}$ denotes the set of K secondary substation nodes with nameplate capacity C_k, $k = \{1, \dots, K\}$.

Let $\mathcal{G}$ denote the set of all possible grid topologies G within the region of interest. A grid topology is represented as a graph $G = (V, E)$ where $V \subseteq V_{\text{base}}$ is the set of electrical nodes (i.e., buses, secondary substations and end-consumers) and $E \subseteq E_{\text{base}}$ the set of electrical lines. According to this definition, G then follows the layout of G_{base} and is georeferenced via

G_{base}. For $k \in \{1, \dots, K\}$ let $T_k = (V_{T_k}, E_{T_k})$, $V_{T_k} \subseteq V$, $E_{T_k} \subseteq E$, denote the *subtopology* k, i.e., the subgraph of G that is connected to the secondary substation $v_k \in V^s_{\text{base}}$.

B. Probability distribution over grid topologies

Since distribution grids supply all end-consumers and are typically operated using a radial configuration [22], not every possible grid topology can be considered as practically feasible. Precisely, a grid topology G is considered feasible iff it satisfies the following three conditions: first, every end-consumer node is supplied, i.e., $V^l_{\text{base}} \subseteq V$. Second, all nodes in V are connected via E to exactly one secondary substation. This implies that each node in V is part of one subtopology T_k and that all T_k are distinct from each other. Third, each subtopology T_k is a tree, i.e., acyclic. We denote the set of feasible grid topologies by $\mathcal{G}_f \subseteq \mathcal{G}$ and by $\delta^{G_{\text{base}}}_{\mathcal{G}_f}(G)$ its indicator function, i.e., a map returning one if $G \in \mathcal{G}_f$ and zero else.

In order to assign numeric probability values to a feasible grid topology G, we characterize each $T_k \in G$, $k = \{1, \dots, K\}$, by a set of features $\mathbf{x}(T_k; G_{\text{base}}) \in \mathbb{R}^n$. These features are selected to be relevant for the planning of distribution grids. Specifically in this work, $x_1(T_k; G_{\text{base}})$ is the load factor of the secondary substation, i.e.,

$$x_1(T_k; G_{\text{base}}) = \sum_{v \in V_{T_k}} \frac{P_l(v)}{C_k}, \tag{1}$$

and $x_2(T_k; G_{\text{base}})$ is the maximum length of a feeder in T_k, i.e., the maximum distance from the secondary substation to any end-consumer in V_{T_k}. Note that other features can be used to characterize a grid topology, see, e.g., [23]. Load factor and maximum feeder length provide, however, a comprehensive set of features, with the former characterizing the grid assets and the latter the grid layout.

We then define the probability density over G using an exponential family model [24]. Specifically, we model the unnormalized density $\tilde{p}(G; G_{\text{base}}, \boldsymbol{\eta})$ as

$$\tilde{p}(G; G_{\text{base}}, \boldsymbol{\eta}) = \delta^{G_{\text{base}}}_{\mathcal{G}_f}(G) \prod_{T_k \in G} e^{\boldsymbol{\eta}^\intercal \boldsymbol{\tau}(\mathbf{x}(T_k; G_{\text{base}}))} \tag{2}$$

where $\boldsymbol{\tau} : \mathbb{R}^n \to \mathbb{R}^m$ is a sufficient statistics function characterizing the probability model associated with each feature and $\boldsymbol{\eta}$ are the corresponding natural parameters. For instance, a Gaussian model of a single feature x with mean μ and variance σ^2 can be formulated using the sufficient statistics $\boldsymbol{\tau}(x) = (x, x^2)^\intercal$ and the natural parameters $\boldsymbol{\eta} = (\mu/\sigma^2, -1/2\sigma^2)$. Likewise, other distributions can be modeled using this framework, given the sufficient statistics, e.g., for the Gamma distribution used in this work $\boldsymbol{\tau}(x) = (\log x, x)^\intercal$. The full, normalized probability density over G is then defined as

$$p(G; G_{\text{base}}, \boldsymbol{\eta}) = \frac{1}{Z(\boldsymbol{\eta}, G_{\text{base}})} \tilde{p}(G; G_{\text{base}}, \boldsymbol{\eta}), \tag{3}$$

where $Z(\boldsymbol{\eta}, G_{\text{base}})$ is the normalization factor

$$Z(\boldsymbol{\eta}, G_{\text{base}}) = \sum_{G \in \mathcal{G}} \tilde{p}(G; G_{\text{base}}, \boldsymbol{\eta}). \tag{4}$$

C. Model parameter learning

The natural parameters $\boldsymbol{\eta}$ of the exponential family model can be estimated through MLE, leveraging a set of known, exemplary georeferenced grid topologies. Given a training dataset $\mathcal{G}_{\text{train}} = \{(G^{(i)}_{\text{base}}, G^{(i)})\}_{i=1,\dots,D}$ comprising D feasible grid topologies, the objective is to determine $\boldsymbol{\eta}$ that maximizes the likelihood of $\mathcal{G}_{\text{train}}$, i.e.,

$$\max_{\boldsymbol{\eta}} \prod_{i=1}^{D} p(G^{(i)}; G^{(i)}_{\text{base}}, \boldsymbol{\eta}). \tag{5}$$

Since the negative log-likelihood of exponential family models exhibits strict convexity [24], the optimal parameters can be uniquely and robustly estimated via gradient-based optimization techniques. Specifically, the gradient of the negative log-likelihood $\mathcal{L}(\mathcal{G}_{\text{train}}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ is defined as

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \mathcal{L}(\mathcal{G}_{\text{train}}, \boldsymbol{\eta}) = &- \sum_{i=1}^{D} \sum_{T_k \in G^{(i)}} \boldsymbol{\tau}(\boldsymbol{x}(T_k; G^{(i)}_{\text{base}})) \\ &+ D \, \nabla_{\boldsymbol{\eta}} \log \sum_{G \in \mathcal{G}} \delta^{G_{\text{base}}}_{\mathcal{G}_f}(G) \, e^{\boldsymbol{\eta}^T \sum_{T_k \in G} \boldsymbol{\tau}(\boldsymbol{x}(T_k; G_{\text{base}}))}. \end{aligned} \tag{6}$$

Summing over all $G \in \mathcal{G}$ to calculate the normalization factor $Z(\boldsymbol{\eta}, G_{\text{base}})$ or its derivative is theoretically possible but computationally prohibitive, primarily due to the fact that this sum involves $2^{(M+N)}$ elements. In this work, we thus approximate $Z(\boldsymbol{\eta}, G_{\text{base}})$ by summing over a significantly smaller and computationally tractable set of feasible random grid topologies $\tilde{\mathcal{G}} \subseteq \mathcal{G}$, which we refer to as the background grid topologies. If $\tilde{\mathcal{G}}$ is representative, then omitting the remaining samples from the sum only changes $p(G; G_{\text{base}}, \boldsymbol{\eta})$ by a constant scaling factor. This scaling factor, however, has no impact on the optimal $\boldsymbol{\eta}$, due to the logarithm. Moreover, the MCMC algorithm described next is independent of this scaling factor.

D. Statistical inference

Given the trained exponential family model, efficient statistical inference can be performed to derive meaningful planning insights for a region of interest, e.g., overload probability of the secondary substations [2], voltage criticality assessment at or behind the secondary substations, or average grid losses. Formally, these inference tasks involve calculating the means, variances, or quantiles of values derived from the uncertain grid topologies. These tasks can be solved computationally through sampling and the efficiency of the solution depends on the rapid generation of different high-probability grid topologies from the exponential family model.

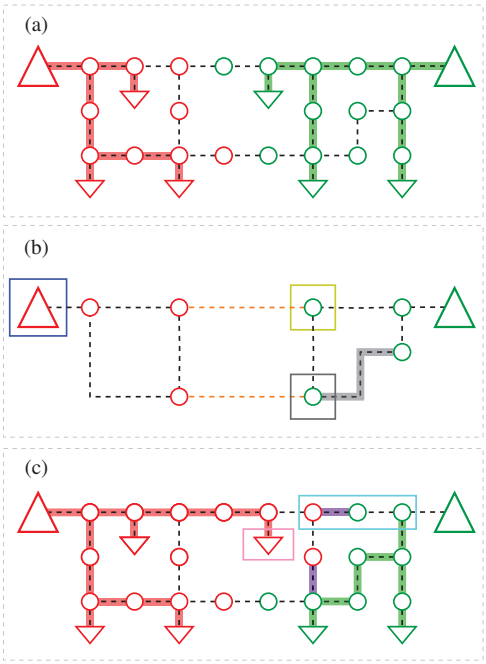

Fig. 2. Schematic of a georeferenced grid topology G_t and the novel, distribution grid-adapted generation mechanism of a Markov chain candidate G'_{t+1}. (a) The shown grid topology G_t follows the base graph G_{base} ( ) and is the union of two tree-like subtopologies (,) that connect all end-consumers (,) to exactly one secondary substation (,). (b) The core base graph $G_{\text{base,core}}$ is a reduced version of G_{base}. For the selected secondary substation k ( ), there are two active edges ( ), from which a core node v_t ( ) is sampled and reassigned to k. If the edge ( ) did not exist, the core node ( ) would be isolated when performing the reassignment and the generation procedure would be aborted. (c) To obtain the complete new grid topology candidate G'_{t+1}, subtrees ( ) assigned to nodes with the new secondary substation label and segments with the same new secondary substation label at both ends are completely reassigned. Segments ( ) with different new secondary substation label are split at random locations ( ). Finally, a minimum spanning tree is computed for each subtopology to eliminate possible loops.

1) Metropolis-Hastings: Theoretically, one can sample a random grid topology G by independently determining for each edge or node within the base graph G_{base} whether it is a part of G. However, using this procedure would probably result in a grid topology G that is infeasible. Even if a sampled G were feasible, it would most likely remain a low-probability occurrence within the learned distribution. Rather than using such a Monte Carlo (MC) approach, we introduce a more efficient MCMC algorithm. This algorithm constructs a Markov chain of high-probability grid topologies from the learned distribution $p(G; G_{\text{base}}, \boldsymbol{\eta})$, in the following referred to as target distribution. Each sample depends on its preceding one and inherits some of its high-probability characteristics. Yet the chain must also exhibit sufficient difference to rapidly explore the space of all high-probability grid topologies and obtain representative samples in finite time. Moreover, the generation of new samples must be computationally efficient to enable the use of larger sample sizes and thereby make the inference more statistically robust.

A common MCMC approach, also employed in this work, is the Metropolis-Hastings (M-H) algorithm [25]. We denote the grid topology at step t of a Markov chain as $G_t = (V_t, E_t)$. To generate the next grid topology G_{t+1} from the target distribution $p(G; G_{\text{base}}, \boldsymbol{\eta})$, the M-H algorithm generates a proposal grid topology G'_{t+1} from the proposal distribution $q(G'_{t+1}; G_t, G_{\text{base}})$ and accepts it as G_{t+1} using the acceptance ratio

$$\alpha = \frac{\tilde{p}(G'_{t+1}; G_{\text{base}}, \boldsymbol{\eta})}{\tilde{p}(G_t; G_{\text{base}}, \boldsymbol{\eta})} \frac{q(G_t; G'_{t+1}, G_{\text{base}})}{q(G'_{t+1}; G_t, G_{\text{base}})}, \tag{7}$$

otherwise G_t is retained as G_{t+1}. As long as the Markov chain is ergodic, i.e., every grid topology with non-zero probability is attainable with non-zero probability, the distribution of the samples in this chain matches $p(G_t; G_{\text{base}}, \boldsymbol{\eta})$ for $t \to \infty$ [25].

2) Specialized proposal distribution: When generating a new candidate grid topology G'_{t+1} from its predecessor G_t, G'_{t+1} must inherit many high-probability characteristics from G_t, while exhibiting sufficient difference. Moreover, both $q(G'_{t+1}; G_t, G_{\text{base}})$ and $q(G_t; G'_{t+1}, G_{\text{base}})$ must be analytically computable. These requirements can be met by using a novel and computationally efficient candidate generation mechanism, custom-designed for grids topologies and sketched in Fig. 2 (b)–(c). Reducing G_{base} into a *core* base graph $G_{\text{base,core}} = (V_{\text{base,core}}, E_{\text{base,core}})$, $V_{\text{base,core}} \subseteq V_{\text{base}}$ is a key concept for efficiently generating G'_{t+1}, see Fig. 2 (b). This reduction enables efficient operations on complete *subtrees* and *segments* within G_{base}, eliminating the need to handle individual nodes separately. Constructing $G_{\text{base,core}}$ is reminiscent of the Kron reduction technique often applied to complex power grids [26] and involves the following steps. First, all terminal nodes, i.e., nodes $v \in V_{\text{base}}$ with $\deg(v) = 1$ and $v \notin V^s_{\text{base}}$, are iteratively assigned to their neighbor until no such nodes remain. All nodes in V_{base} assigned to one of the remaining nodes form its subtree within G_{base}, see Fig. 2 (c). Second, all nodes with a degree of two are iteratively eliminated, while preserving the physical distance. During this procedure, the neighbors of a removed node are reconnected through a newly generated edge, and the physical distance is adjusted accordingly. The original edges are then assigned to the newly established connection. All edges in E_{base} assigned to a remaining connection $(u, v) \in E_{\text{base,core}}$ form a *segment* $S(u, v) \subseteq E_{\text{base}}$ within G_{base}, see Fig. 2 (c).

Let $a_t : V_{\text{base}} \to \{0, 1, \ldots, K\}$ encode the assignment of base graph nodes to the secondary substations of G_t, where $a_t(v) = 0$ indicates that node $v \in V_{\text{base}}$ is not part of V_t. Moreover, let the *active* edges of subtopology k in the core base graph $G_{\text{base,core}}$ be defined as $\tilde{E}^t_{\text{base,core}}(k) = \{(u, v) \in E_{\text{base,core}} | a_t(u) = k \wedge a_t(v) \neq k\}$.

The first step for generating a new candidate grid topology G'_{t+1} operates at the level of the reduced base graph $G_{\text{base,core}}$. Specifically, a subtopology T_k of G_t is selected uniformly at random. Then, a core node $v_t \in V_{T_{k'}}$ is switched from subtopology $T_{k'}$ to T_k by drawing from $\tilde{E}^t_{\text{base,core}}(k)$, also uniformly at random, see Fig. 2 (b). If the switch of v_t from subtopology $T_{k'}$ to subtopology T_k leads to isolated core nodes in $G_{\text{base,core}}$, i.e., $v' \in V_{\text{base,core}}$ with $a_t(v') = k'$ is not connected the secondary substation k', then the candidate

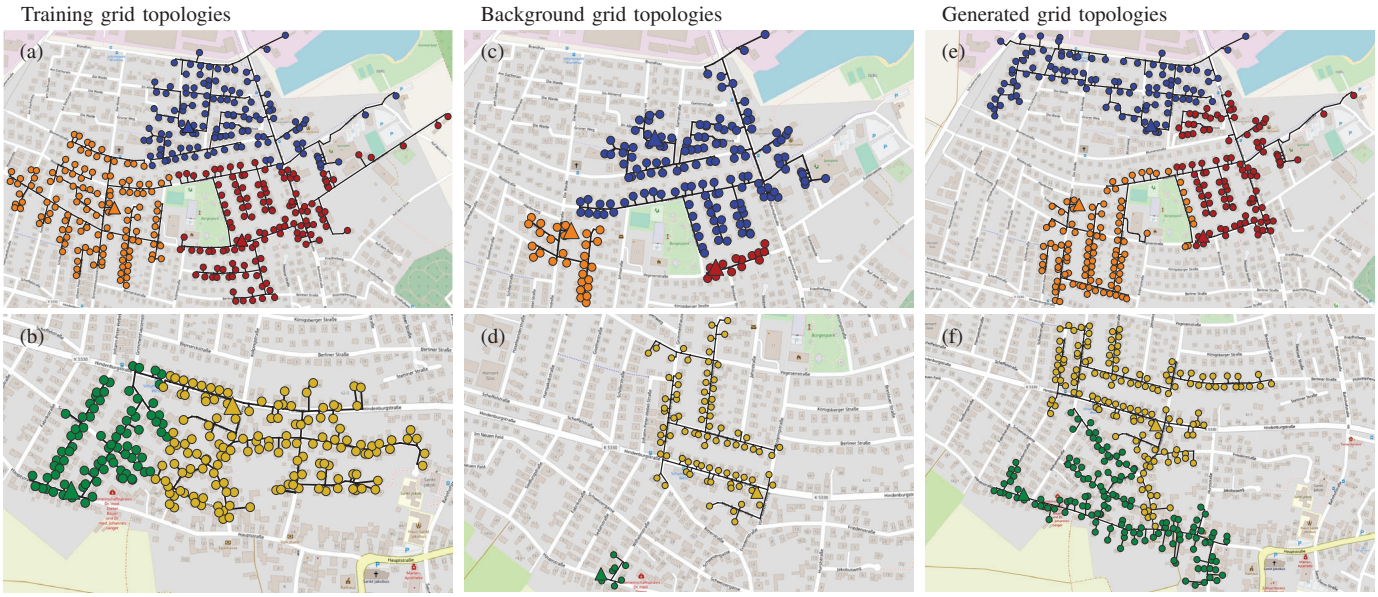

Fig. 3. Exemplary grid topologies for five different secondary substations (▲,▲,▲,▲,▲) in the considered Schutterwald region. (a)–(b) Training grid topologies. (c)–(d) Background grid topologies. (e)–(f) Generated grid topologies using the proposed MCMC algorithm.

generation is aborted and the Markov chain progresses with the previous topology[1], see Fig. 2 (b).

The second step for generating G'_{t+1} involves deriving the complete proposal within the base graph G_{base}. To this end, the following nodes are reassigned from secondary substation k' to secondary substation k. All nodes in segments $S(u, v_t)$ with $a_t(u) = k$ along with all their subtrees are reassigned to k, see Fig. 3 (c). Segments $S(u, v_t)$ with $a_t(u) \neq k$ are split at a location drawn uniformly at random. The subsegments connected to v_t are reassigned to k, while the other subsegments retain their secondary substation label k', see Fig. 3 (c). Finally, a minimum spanning tree algorithm is applied to ensure that $T_{k'}$ and T_k are acyclic. The selected nodes and edges then form the new subtopologies $T_{k'}$ and T_k for G'_{t+1}.

We calculate the numeric value of the proposal probability distribution $q(G'_{t+1}; G_t, G_{\text{base}})$ as

$$\frac{1}{K} \frac{1}{|\tilde{E}^t_{\text{base,core}}(k, v_t)|} \prod_{(u,v_t)\in E^t_{\text{base}}(k)} \frac{1}{|S(u, v_t)|}, \tag{8}$$

where $\tilde{E}^t_{\text{base,core}}(k, v_t) = \{(u, v_t) \in \tilde{E}^t_{\text{base,core}}(k) | a_t(u) = k\}$ and $E^t_{\text{base}}(k) = \{(u, v_t) \in E_{\text{base,core}} | a_t(u) \neq k\}$. The numeric value of the reverse proposal $q(G_t; G'_{t+1}, G_{\text{base}})$ is analogously calculated as

$$\frac{1}{K} \frac{1}{|\tilde{E}^{t+1}_{\text{base,core}}(k', v_t)|} \prod_{(u,v_t)\in E^{t+1}_{\text{base}}(k')} \frac{1}{|S(u, v_t)|}. \tag{9}$$

Note that the ergodicity of the procedure is inherently guaranteed by its construction.

[1]This is because the reverse probability of switching two core nodes in this case would be zero, making the acceptance ratio in Eq. (7) zero as well.

III. Simulation Experiments

The following simulation experiments comprise three parts. First, the parameter learning for the exponential family model is presented. Second, the performance of the MCMC algorithm is examined. Third, the improved statistical inference of the proposed framework when compared to a MC sampling technique is demonstrated, e.g., the one introduced in our previous work [2].

A. Setup & Implementation

A residential region in Schutterwald, Germany, is considered with 13 secondary substations and 1751 end-consumers. The geocoordinates and nameplate capacity of the secondary substations are extracted from [27]. The street configuration and the location and living area of the end-consumers are obtained open-source from OpenStreetMap.

The dataset for training the exponential family model should comprise grid topologies from real-world distribution grids. Since access to these grid topologies is, however, often restricted [10], a synthetic dataset is generated as substitute using the following procedure. First, 2000 random grid topologies are generated for the specified Schutterwald region using the algorithm from [2]. Next, we assume that the load factor of secondary substations of real-world distribution grids ranges from 0.3 to 1.3 and the maximum feeder distance does not exceed 1 kilometer. A generated grid topology is included in the training dataset iff all of its subtopologies meet these assumptions, see, e.g., Fig. 3 (a)–(b).

When yielding the background grid topologies $\tilde{\mathcal{G}}$, we refrained from using the algorithm from [2]. This is motivated by its tendency of lacking representativeness and exhibiting a bias towards certain grid characteristics. Instead, we establish $\tilde{\mathcal{G}}$ by generating 5000 grid topologies with the introduced

MCMC algorithm and a uniform target distribution. Using this approach yielded grid topologies with more evenly distributed grid characteristics, see, e.g., Fig. 3 (c)–(d). Note that the background dataset should be diverse, comprising a spectrum of grid topologies of different designs and regions to ensure its representativeness. Since the open-source availability of the secondary substation locations is, however, limited, we considered the same region in Schutterwald, Germany to establish $\tilde{\mathcal{G}}$ in this work.

The proposed framework is implemented in Python using the TensorFlow and NetworkX libraries. All experiments were run using a standard laptop with an Intel i7-12800H CPU. Note that when generating a new candidate grid topology G'_{t+1}, only the subtopologies $T_{k'}$ and T_k are changed. To enhance run time efficiency and eliminate repetitive computations, we thus calculate the features $\mathbf{x}(T_k; G_{\text{base}})$ at each step of the Markov chain only for $T_{k'}$ and T_k. Meanwhile, for the remaining subtopologies, we use a caching mechanism to store and then reuse the previously calculated features.

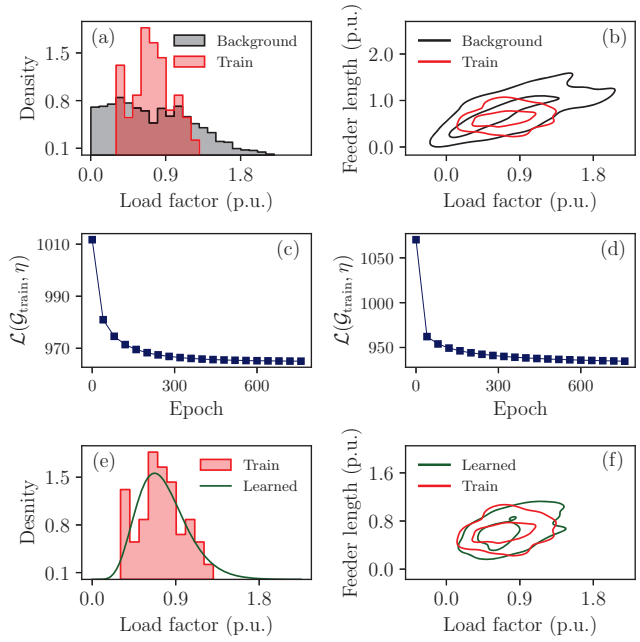

Fig. 4. (a) Histograms of the load factor for the training and background dataset. (b) Kernel density estimate of the joint distribution of the two considered features. (c) Evolution of the negative log-likelihood $\mathcal{L}(\mathcal{G}_{\text{train}}, \boldsymbol{\eta})$ across 1000 training epochs for training an exponential family model with the load factor feature only or (d) with both considered features. (e) Learned and training distribution for an exponential family model using the load factor only or (f) both considered features.

B. Parameter learning

We first investigate the process of parameter learning for the exponential family model. Fig. 4 (a)–(b) shows the feature distributions of the training and background grid topologies. As it can be seen, the training grid topologies can be differentiated from the background based on the proposed features. Note that due to the underlying street configuration and fixed total load within the distribution grid, feasible grid topologies have a bounded load factor and maximum feeder length. Fig. 4 (c)–(d) shows the robust training of the exponential family framework using either only the load factor feature or both proposed features, modeled with a Gamma distribution. The resulting learned distributions align well with the distribution of the training dataset, see Fig. 4 (e)–(f). In the following, we examine exponential family models exclusively based on the load factor to allow for one-dimensional distribution plots.

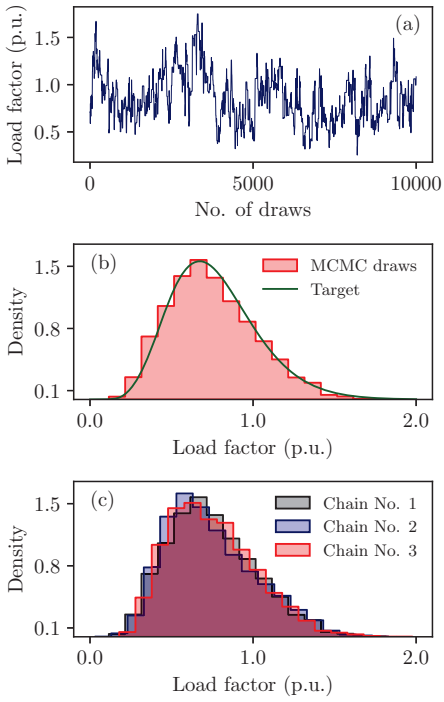

Fig. 5. (a) Trace plot depicting the load factor of a secondary substation from a Markov chain comprising 10000 draws. (b) Load factor distribution of a Markov chain comprising 10000 draws in relation to the target distribution $p(G; G_{\text{base}}, \boldsymbol{\eta})$. (c) Load factor distribution of three different Markov chains comprising 10000 draws.

C. Markov chain convergence diagnostics

In Fig. 5, we examine the performance of the proposed MCMC algorithm using three diagnostic tools. Fig. 5 (a) shows the load factor trace of a secondary substation from a Markov chain comprising 10000 draws. As it can be seen, the Markov chain exhibits robust mixing, avoiding any tendency to remain stagnant in a particular region within the feature space. In Fig. 5 (b), we show the distribution of a Markov chain comprising 10000 draws in comparison to the target distribution $p(G; G_{\text{base}}, \boldsymbol{\eta})$. As evident, the Markov chain faithfully represents $p(G; G_{\text{base}}, \boldsymbol{\eta})$. To further substantiate this finding, three distinct Markov chains were independently constructed, each initialized with a different starting grid topology and each comprising 10000 draws. As shown in Fig. 5 (c), all three chains consistently yield a similar distribution, indicating convergence of the Markov chains to their equilibrium distribution.

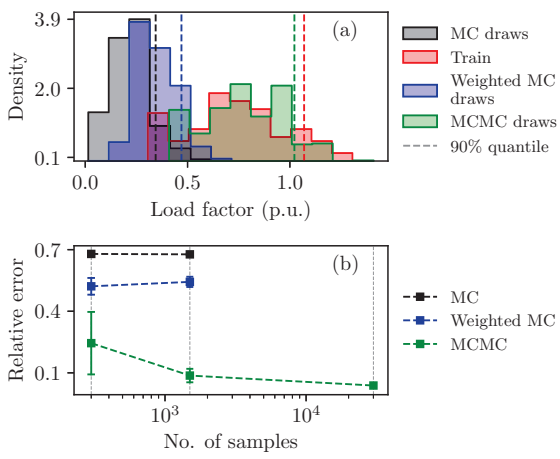

Fig. 6. (a) Histograms and 90% quantile estimates of the load factor for one exemplary secondary substation, produced with the proposed MCMC approach and two baseline MC methods (see Sec. III-D, 1500 samples), as well as equivalent values for the training dataset. (b) Relative error of the 90% quantile estimates for the selected secondary substation when compared to the 90% quantile of the training dataset. The results are presented as the mean and standard deviation across three independent repetitions of the experiment. The 30000 samples experiment was not considered for both MC approaches due to computational constraints.

D. Efficient statistical inference

In the following, we explore two possible advantages of the proposed framework in contrast to a MC approach, i.e., enhanced sample quality and shorter run time.

We use the algorithm from [2] as our MC baseline. Attempting to sample grid topologies by independently drawing edges and nodes from E_{base} and V_{base} typically yields infeasible grid topologies, making this approach computationally prohibitive. In contrast, the procedure from [2] always produces feasible grid topologies, yet the distribution of these grid topologies is unknown. We thus consider for comparison both the original samples and a weighted approach, using $\tilde{p}(G; G_{\text{base}}, \boldsymbol{\eta})$ to better align with the learned target distribution. The presented MCMC results are based on samples from three independent chains, with a non-considered burn-in of 100 samples.

In Fig. 6, we investigate the enhanced sample quality achieved by the proposed framework. Fig 6 (a) shows the load factor distributions and the 90% quantile estimates of one exemplary secondary substation within the region of interest, produced with the different considered methods and compared to the benchmark training dataset. As it can be seen, the distribution and estimated 90% quantile of the MCMC draws align closer with the benchmark training distribution and its 90% quantile, in contrast to the distribution and 90% quantile estimate of the both the weighted and unweighted MC draws. Moreover, while weighting the MC draws brings them closer to the benchmark distribution, the unclear distribution resulting from the algorithm proposed in [2] remains.

These findings are further substantiated in Fig. 6 (b), which shows the relative deviation of the different estimation procedures against the benchmark. As the number of samples increases, the proposed MCMC method converges towards the benchmark, reducing both estimation error and variance observed across the three independent repetitions of the experiment. In contrast, both MC approaches consistently produce inaccurate estimates.

TABLE I
RUN TIMES OF GRID TOPOLOGY GENERATION FOR THE CONSIDERED SCHUTTERWALD REGION.

No. of grid topologies	Run time	
	MC baseline [2]	MCMC algorithm
1	17 sec	0.8 sec
2000	567 min	27 min

Table I presents the computation times for one or more samples using the MCMC algorithm and the MC baseline. The run time required for generating a new grid topology using the MCMC algorithm is smaller by a factor of approx. 21. Moreover, the reported run time of the MCMC algorithm includes the time for feature calculation, while the reported run time of the MC baseline does not factor in this calculation since it occurs after the simulation has concluded. The computational advantage of the proposed MCMC approach can be understood as follows. While the MC baseline scales with the number of nodes V_{base} in the base graph G_{base}, the mechanism for generating a new grid topology within the MCMC algorithm uses the core base graph $G_{\text{base,core}}$ and thus scales with the number of edges $E_{\text{base,core}}$ in $G_{\text{base,core}}$, and $|E_{\text{base,core}}| \ll |V_{\text{base}}|$.

IV. CONCLUSION

In this work, we introduced a novel framework to learn a well-defined probability distribution over feasible georeferenced grid topologies. By leveraging this learned distribution, our approach can then perform efficient statistical inference using a specifically tailored MCMC algorithm.

There are several avenues of future work. We are actively engaged on expanding the proposed approach to include available grid measurements. Moreover, we are integrating real-world topology data to train the proposed algorithm. Since in this work we considered the same region for both parameter learning and statistical inference, it would be also interesting to use different grid topologies from various regions and assess their impact on the performance of the proposed algorithm. Finally, exploring additional features for characterizing grid topologies as well as alternative statistical models within the exponential family for modeling these features are also worth investigating.

In summary, there seems to be significant promise in learning probability distributions over georeferenced grid topologies and leveraging these distributions for statistical inference.

REFERENCES

[1] Y. Liao, Y. Weng, G. Liu, Z. Zhao, C.-W. Tan, and R. Rajagopal, "Unbalanced multi-phase distribution grid topology estimation and bus phase identification," *IET Smart Grid*, vol. 2, pp. 557–570, 2019.

[2] D. Tomaselli, P. Stursberg, M. Metzger, and F. Steinke, "Representing topology uncertainty for distribution grid expansion planning," in *CIRED 2023*. CIRED, 2023.
[3] D. Deka, V. Kekatos, and G. Cavraro, "Learning distribution grid topologies: A tutorial," *IEEE Transactions on Smart Grid*, vol. 30, no. 1, pp. 999–1014, 2023.
[4] K. Moffat and C. Tomlin, "The multiple model adaptive power system state estimator," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 3525–3530.
[5] C. Mateo, F. Postigo, F. de Cuadra, T. G. San Roman, T. Elgindy, and B. Palmintier, "Building large-scale us synthetic electric distribution system models," *IEEE Transactions on Smart Grid*, vol. 11, pp. 5301–5313, 2020.
[6] H. K. Çakmak, L. Janecke, and V. Hagenmeyer, "Automated generation of large-scale distribution grid models based on open data and open source software using an optimization approach," *arXiv preprint arXiv:2202.13692*, 2022.
[7] P. Schultz, J. Heitzig, and J. Kurths, "A random growth model for power grids and other spatially embedded infrastructure networks," *The European Physical Journal Special Topics*, vol. 223, pp. 2593–2610, 2014.
[8] Z. Wang, A. Scaglione, and R. J. Thomas, "Generating statistically correct random topologies for testing smart grid communication and control networks," *IEEE Transactions on Smart Grid*, vol. 1, pp. 28–39, 2010.
[9] S. Ma, Y. Yu, and L. Zhao, "Dual-stage constructed random graph algorithm to generate random graphs featuring the same topological characteristics with power grids," *Journal of Modern Power Systems and Clean Energy*, vol. 5, pp. 683–695, 2017.
[10] M. Liang, Y. Meng, J. Wang, D. L. Lubkeman, and N. Lu, "Feedergan: Synthetic feeder generation via deep graph adversarial nets," *IEEE Transactions on Smart Grid*, vol. 12, pp. 1163–1173, 2020.
[11] 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. 8, no. 6, pp. 2794–2802, 2016.
[12] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, "Recent developments in exponential random graph (p*) models for social networks," *Social Networks*, vol. 29, pp. 192–215, 2007.
[13] N. De Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs," *arXiv preprint arXiv:1805.11973*, 2018.
[14] B. Donon, B. Donnot, I. Guyon, and A. Marot, "Graph neural solver for power systems," in *2019 International Joint Conference on Neural Networks*. IEEE, 2019, pp. 1–8.
[15] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
[16] E. Kellerer and F. Steinke, "Scalable economic dispatch for smart distribution networks," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1739–1746, 2014.
[17] P. Erdös and A. Rényi, "On random graphs I," *Publ. Math.*, vol. 6, pp. 290–297, 1959.
[18] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, 2002.
[19] P. W. Holland and S. Leinhardt, "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, vol. 76, pp. 33–50, 1981.
[20] D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris, "ergm: A package to fit, simulate and diagnose exponential-family models for networks," *Journal of Statistical Software*, vol. 24, 2008.
[21] A. Bojchevski, O. Shchur, D. Zügner, and S. Günnemann, "Netgan: Generating graphs via random walks," in *International Conference on Machine Learning*. PMLR, 2018, pp. 610–619.
[22] D. Deka, M. Chertkov, and S. Backhaus, "Topology estimation using graphical models in multi-phase power distribution grids," *IEEE Transactions on Power Systems*, vol. 35, pp. 1663–1673, 2019.
[23] G. Prettico, F. Gangale, A. Mengolini, A. Lucas, and G. Fulli, "Distribution system operators observatory," *European Commission Joint Research Centre: Ispra, Italy*, 2016.
[24] M. Schweinberger, P. N. Krivitsky, C. T. Butts, and J. R. Stewart, "Exponential-family models of random graphs: inference in finite, super and infinite population scenarios," *Stat. Sci.*, vol. 35, pp. 627–662, 2020.
[25] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, pp. 327–335, 1995.
[26] F. Dorfler and F. Bullo, "Kron reduction of graphs with applications to electrical networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 1, pp. 150–163, 2012.
[27] L. Thurner, A. Scheidler, F. Schäfer, J.-H. Menke, J. Dollichon, and M. Braun, "pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems," *IEEE Transactions on Power Systems*, vol. 33, pp. 6510–6521, 2018.