FRAMEWORKS FOR REALISTIC MODELING AND ANALYSIS OF POWER GRIDS

Rounak Meyur

Charlottesville, Virginia

Bachelor of Technology, National Institute of Technology, Trichy, India, 2016

Master of Science, Virginia Polytechnic & State University, Blacksburg, 2019

A Dissertation submitted to the Graduate Faculty

of the University of Virginia in Candidacy for the Degree of

Doctor of Philosophy

Department of Electrical and Computer Engineering

University of Virginia

December 2022

Anil Vullikanti, Chair

Madhav Marathe

Henning Mortveit

Jundong Li

Virgilio Centeno

Arun Phadke

Frameworks for Realistic Modeling and Analysis of Power Grids

Rounak Meyur

(ABSTRACT)

The power grid is going through significant changes with the introduction of renewable energy sources and incorporation of smart grid technologies. These rapid advancements necessitate new models and analyses to keep up with the various emergent phenomena they induce. At the same time, these need to resemble the actual power system model and dynamics. In this dissertation, I propose two frameworks – (i) for constructing synthetic power distribution networks for a given geographic region which closely resembles the actual physical counterpart, and (ii) for performing cascading failure analysis in the power grid when subjected to a severe disturbance, such that it resembles the actual power grid events as closely as possible. For the first framework, I use openly available information about interdependent road and building infrastructures and incorporate engineering and economic constraints to construct the distribution networks. The networks synthesized by this framework represent realistic power distribution systems (as compared to standard IEEE test cases) that can be used by network scientists to analyze complex events in power grids. The second framework for cascading failure analysis uses a realistic representation of the underlying power grid, including the topology, the control and protection components, and a dynamic stability analysis that goes beyond traditional work consisting of structural and linear flow analysis. The proposed framework can be used to assess vulnerability of the power grid to any disturbance like a physical attack, cyber attack or any severe weather event. In particular, I consider the case of a targeted physical attack on the power grid of

Washington DC. The results show that realistic representations and analysis can lead to fundamentally new insights that are not possible by using simplified models.

Dedication

Dedicated to my parents, teachers and friends.

Acknowledgments

I cannot be content with any acknowledgement I put together for thanking my mentor and advisor, Dr. Madhav Marathe. Apart from his obvious role in the technical works presented in this dissertation, I am indebted to him for imparting profound research acumen, ethics, inclusiveness, and tender yet strong mentoring skills. I next wish to thank my PhD committee members for guiding me through the work and motivating me with several possible research directions to proceed with. Apart from this dissertation, the confidence that I have gained in power systems comes from the extended discussions with Dr. Virgilio Centeno and Dr. Arun Phadke of Virginia Tech. I am extremely thankful to them for their support and contributions. I am also thankful for the personal and academic support from the close-knit group of students and faculty at the Biocomplexity Institute (BII), and continuous financial support from the National Science Foundation (NSF). I am particularly thankful for to NSF grants EAGER 1927791 and CINES 1916805 for supporting me over the course of my PhD. I would also like to thank Dr. Arindam Fadikar and Dr. Gursharn Kaur for motivating me with interesting statistical discussions regarding my work. I did three summer internships with the Pacific Northwest National Laboratory (PNNL), where I was fortunate to work closely with Dr. Mahantesh Halappanavar who got me connected with relevant members to help me with my work. Particularly, I am very thankful for the jubilant collaboration with Dr Bala Krishnamoorthy from Washington State University, whom I came to know during one of my internships. While grad-school is tough, it is surmountable merely because of the friends. I would like to thank them all for being a supportive part of good and less good times of my journey so far.

Contents

List of Figures

List of Tables

List of Abbreviations

$R_{\odot}$ The radius of the earth (6378 km)

AC Alternating Current

ADMM Alternating Direction Method of Multipliers

ANSI American National Standards Institute

CB Circuit Breaker

CIGRE Conseil International des Grands R éseaux Electriques

CPS Cyber Physical System

CT Current Transformer

DC Direct Current

DCBR Direction Comparison Blocking Relay

DER Distributed Energy Resource

DNP Distribution Network Protocol

DOCR Directional Overcurrent Relay

EIA Energy Information Administration

ESRI Environmental Systems Research Institute

EV Electric Vehicle

GANN Generative Adversarial Neural Network

GIS Geographic Information System

HMI Human Machine Interface

HV High Voltage

ICCP Inter Control Center Protocol

ICT Information and Communication Technology

IEC International Electrotechnical Commission

IEEE Institute of Electrical and Electronics Engineers

ILP Integer Linear Program

LAN Local Area Network

LDF Linearized Distribution Flow

LP Linear Program

LV Low Voltage

MILP Mixed Integer Linear Program

MIQP Mixed Integer Quadratic Program

MV Medium Voltage

NASEM National Academies of Sciences, Engineering and Medicine

NC Normally Closed

NERC North American Electric Reliability Corporation

NO Normally Open

OSM OpenStreetMaps

PDR Percentage Differential Relay

PLC Power Line Communication

PSRC Power System Relaying and Control Committee

PT Potential Transformer

PV Photovoltaic

QP Quadratic Program

REVS Reliability-aware EV charge Scheduling

RNM Reference Network Model

SCADA Supervisory Control and Data Acquisition

SOC State of Charge

TO Transmission Operator

TOU Time of Use

US NRC United States Nuclear Regulatory Commission

Chapter 1

Introduction

The power grid is a vital infrastructure providing functional and operational support to other civic infrastructures like health, education, transportation and housing. Hence, its reliable operation is of utmost importance for the smooth functioning of the economy of a nation. The traditional power grid consists of three parts - (i) generation, (ii) transmission and (iii) distribution systems. The generation is located near the available resources required for generating power. The generated power is transmitted over long distances through high and extra high voltage transmission lines. Thereafter, the power is distributed to the residential and commercial consumers through low and medium voltage distribution lines. A schematic of a typical power grid is shown in Fig. 1.1.

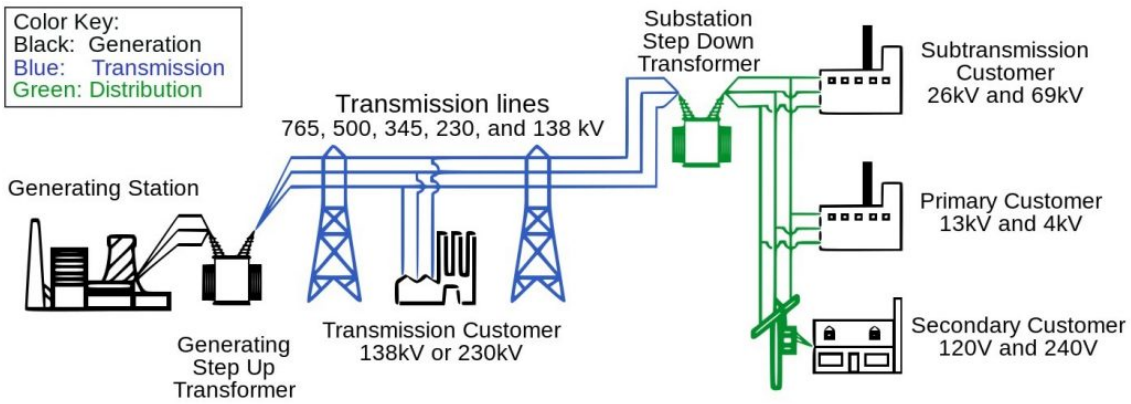

Figure 1.1: Schematic of a power grid consisting of generation, transmission and distribution systems.

With the advent of distributed energy resources (DERs) (through rooftop solar photo-

voltaics (PV), energy storage, and small scale generators), the residential consumers are becoming involved in generating power in the grid. The introduction of electric vehicles (EVs) has also led to a paradigm shift in the energy consumption of traditional end users of electricity. The consumers of electricity, who were once considered as passive entities, are presently playing an active role in the power grid. This necessitates accurate models of consumers and power networks to address questions pertaining to reliability of the power grid. Additionally, power grid is exposed to catastrophic weather events and adversarial attack, which makes it susceptible to cascading failures. Therefore, we require frameworks to analyze impact of such events in order to assess the vulnerability of the power grid.

1.1 Outline of the dissertation

The dissertation is divided into two major parts focused towards (i) realistic modeling and (ii) realistic analysis of power grids.

Chapters 2-4 consider the aspect of realistic modeling of power distribution grids. Chapter 2 provides a general framework for creating an ensemble of realistic distribution networks for a geographic region. Chapter 3 provides methods for comparing created synthetic networks with actual power distribution network counterparts. Finally Chapter 4 provides case studies where the created synthetic networks are used to address common problems associated with distribution system reliability.

Chapter 5 deals with realistic analysis of power system. The analysis is focused towards studying cascading failures in the power grid. The power grid is modeled as a multi-layered network where control and protection systems play a key role in causing cascading outages.

Chapter 2

Realistic Modeling of Power Distribution Networks

Publications

1. Rounak Meyur, Anil Vullikanti, Samarth Swarup, Henning S. Mortveit, Virgilio Centeno, Arun Phadke, H. Vincent Poor, and Madhav Marathe, "Ensembles of realistic power distribution networks" in Proceedings of the National Academy of Sciences, Vol. 119 No. 42, Oct 2022.

2. Rounak Meyur, Madhav Marathe, Anil Vullikanti, Samarth Swarup, Henning S. Mortveit, Virgilio Centeno, and Arun Phadke, "Creating realistic power distribution networks using interdependent road infrastructure", in IEEE International Conference on Big Data, Dec 2020 (pp. 1226–1235).

2.1 Introduction

A reliable power grid constitutes the backbone of a nation's economy providing vital support to various sectors of society and other civil infrastructures. Power distribution networks are created in a bottom-up fashion connecting small clusters of residential loads

to distribution substations, thereby electrifying the entire society. These bear a structural resemblance to other common networked infrastructures such as transportation, communication, water, and gas networks, and are often interdependent in their operations (Byeon et al. 2020). One may use these resemblances and interdependencies to infer one network from available data about these other networks.

Over the past decade, power engineers have aimed to enhance resilience of power systems through incorporation of distributed energy resources (DERs), by deploying advanced metering and monitoring infrastructures (Richler 2020), and by performing system vulnerability and criticality assessments thus reinforcing cybersecurity (Onyeji, Bazilian, and Bronk 2014). Furthermore, spatiotemporally variable consumer load demands, such as electric vehicles (EVs), along with an evolving trend towards a distributed operation of the power grid, have posed new challenges to the system planners and operators (Quiroga, Sauma, and Pozo 2019). Network scientists have emphasized the importance of realistic power network data for accurate analysis as opposed to stylized statistical models (Brummitt, P. D. H. Hines, et al. 2013; Z. Wang, Scaglione, and Thomas 2010; Soltan and Zussman 2016). In order to address these challenges, there is a pressing need for openly available data containing realistic grid topologies along with available geographic information. For example, in the context of power grid expansion planning, the current grid information in conjunction with geographical knowledge of wind maps and solar trajectories can aid in optimized power grid expansion while introducing DERs in the grid (You et al. 2016). Similarly, for system vulnerability analysis, a geographic correlation of grid information with cyclone/hurricane paths can help us identify critical sections in the network and raise preparedness levels for natural disasters (Bernstein et al. 2014). Further, a detailed knowledge about individual residential load usage and consumer behavior can help address policy-level questions. Examples of such problems include identifying the impact of EV adoption

and DER penetration on the current power grid infrastructure as the society moves towards net zero-emission (Popovich et al. 2021; Gaete-Morales et al. 2021).

Simulation-based frameworks capable of performing spatiotemporally-resolved simulations can be utilized to analyze the impact of such evolving trends and analyze system vulnerability. Such assessments are useful to system planners aiming to make decisions about infrastructure development and to operators, while handling emergency system conditions. A common drawback of this simulation-based approach is that it requires detailed information regarding the power network and associated components such as locations and capacities of generation, load demands, and line parameters (P. Hines, Cotilla-Sanchez, and Blumsack 2010; Fan et al. 2021; Decker et al. 2010; Biswas, Bernabeu, and Picarelli 2020). Furthermore, since the majority of grid infrastructure advancements are being done at the low voltage distribution level, a high-resolution analysis of the power distribution systems is important. This necessitates a comprehensive knowledge of customer energy-use profiles, customer behavior, and most importantly, the distribution network topology which connects them. Most such data are, at best, partially available, but more typically are not available at all due to their proprietary nature (Postigo et al. 2017). The lack of such openly available detailed real-world data has been identified as a significant hurdle for conducting research in smart grid technology (NASEM 2016). In recent years there has been an increased interest in generating synthetic power network data to address this issue. The synthetic data are not the real-world data; rather, they are generated by mathematical models operating on openly available information, and are designed to ensure the generated data are similar to the real-world data, thus allowing them to be used as a proxy for the actual data. Some examples of synthetic power grid data include synthetic transmission networks (Gegner et al. 2016; Birchfield et al. 2017; Trpovski, Recalde, and Hamacher 2018; Kadavil, Hansen, and Suryanarayanan 2016), synthetic distribution net-

works (Domingo et al. 2011; Gonzalez-Sotres et al. 2013; Schweitzer et al. 2017; Meyur, Marathe, et al. 2020) and synthetic residential customer energy usage data (Thorve et al. 2018; Tong, Nagpure, and Ramaswami 2021; Klemenjak et al. 2020).

In this chapter, we focus on constructing a modular framework for generating synthetic *power distribution networks*, that is, networks connecting individual residential customers to the distribution substations. We present a first principles approach where we generate an *optimal synthetic distribution network* connecting all residences in a given geographic region to the HV substations through medium voltage (MV) and low voltage (LV) networks. We use the example of Montgomery county of southwest Virginia (US) to create the synthetic power distribution networks, consider all residences and HV substations within the state boundary, and connect them through the synthetic distribution network.

In this context, a critical question arises: are the created networks the only feasible networks connecting the residences and substations? To tackle this problem, we present a methodology for generating an *ensemble* of feasible synthetic power distribution networks for a given region. In the literature related to modeling real-world networks, statistical physics has been used to learn significant structural patterns from an ensemble of networks (Cimini et al. 2019) and thereby to help in network reconstruction from incomplete data. In recent years, statistical aspects of the power networks have drawn the attention of the scientific community for similar reasons. A dataset spanning 70 years for the electric power grid of Hungary has been studied (Hartmann and Sugár 2021) for small-world and scale-free properties. Due to a lack of real world power distribution data, ensembles of distribution networks, which have significant resemblance to actual networks, can suffice for a detailed statistical analysis.

Our contributions include: (i) a holistic modular framework to create synthetic power distribution networks which satisfy structural and power-engineering constraints along with

accurate representation of residential load demand profiles. (ii) a method to create an ensemble of networks by generating multiple feasible networks for a given region. (iii) an open dataset consisting of ensembles of distribution networks for Montgomery county of southwest Virginia (US). This dataset is unique in terms of both size and details. The geographically embedded networks, along with the detailed residential customer usage data, become suitable tools for system-wide planning studies and for addressing policy-level questions.

2.2 Related works

Over the past decade, researchers have developed methods to create synthetic power networks of different sizes. These attempts include interdependent networks created using simple logical models (D'Souza, Brummitt, and Leicht 2014) to models created based on statistical attributes of actual networks (Soltan and Zussman 2016). The simple logical network models allowed network scientists to carry out a linear analysis which reveals big picture results for large networks. However, neglecting rich system data (line parameters and residence load demand) often makes these simple models less useful in certain kinds of applications (Brummitt, P. D. H. Hines, et al. 2013). In recent years, a substantial amount of work has gone into creating synthetic high voltage (HV) transmission networks (Gegner et al. 2016; Birchfield et al. 2017), or combinations of transmission and distribution networks (Atat et al. 2019; H. Li et al. 2020). The primary focus of these papers is to model the transmission grid with a high level of resemblance to the actual grid. These works provide little importance to distribution power networks.

For a long time, power system researchers used the standard test systems published by the Institute of Electrical and Electronics Engineers (IEEE) (*IEEE Test Feeders* 2014; *EPRI*

Test Circuits 2019) to carry out experiments and validate control/optimization algorithms. These test systems were considered a replica of the actual power grid and therefore used as a test-bed before deploying a methodology on the actual power system. These models either consider the residential customers as passive consumers of electricity or assume an aggregated version of multiple consumers as the system load demand. These networks enabled accurate system analysis as long as the role of consumers was not considered. However, with the advent of distributed energy resources (DERs) and electric vehicles (EVs), residential customer behavior has become an integral part of power system analysis (Bistline and Blanford 2021; Sepulveda et al. 2021; Joshi et al. 2021). Therefore, modern power network models need to include detailed residential consumer attributes for accurate representation of the actual power grid.

The literature dealing with the creation of synthetic distribution networks (also referred to as digital twin in the recent literature) can be broadly classified into three categories (Georgilakis and N. D. Hatziargyriou 2015; Resener et al. 2018). The first category comprises of methods that use one or more heuristics to synthesize the networks. The second category of methods uses mathematical programming to encode various physical and structural constraints that real-world networks usually obey. The third category comprises of methods to train a machine learning algorithm and then use this trained algorithm to generate synthetic networks.

In the first category, researchers have proposed several heuristic methods (Domingo et al. 2011; Gonzalez-Sotres et al. 2013; Mateo, Postigo, F. d. Cuadra, et al. 2020; Saha et al. 2019; Kadavil, Hansen, and Suryanarayanan 2016; Bidel, Schelo, and Hamacher 2021) to create random feeder networks with or without using the road network information. The heuristics are primarily used in the initial steps to assign substation locations and construct a feeder network. Typically the heuristics has three broad steps: (*i*) clustering residential

loads to assign substation locations at the cluster centroids (Gonzalez-Sotres et al. 2013), (*ii*) distributing an aggregated residential load to certain load points in a hierarchical fashion (Kadavil, Hansen, and Suryanarayanan 2016) and (*iii*) connecting the resulting load points and substations using minimum spanning tree algorithms (Domingo et al. 2011; Gonzalez-Sotres et al. 2013; Saha et al. 2019). Some of these methods construct an imaginary road network (Domingo et al. 2011; Gonzalez-Sotres et al. 2013), while others have used openly available information (Saha et al. 2019). Most of these papers do not use individual residence locations and have populated the created synthetic networks with random residences or aggregated loads to zip-code centers. Some of these works used a top-down approach where feeder networks are generated and followed by populating with random loads (Saha et al. 2019; Kadavil, Hansen, and Suryanarayanan 2016). None of these methods include the power engineering constraints in the network synthesis phase, rather cables and shunt compensators are chosen in the successive phases such that power engineering constraints are satisfied.

The RNM (Domingo et al. 2011; Gonzalez-Sotres et al. 2013) is an important heuristic-based planning tool for efficient investment options in distribution grid planning. It uses OpenStreetMap and relevant geographic data to create distribution networks in a given region. The comparison of networks generated by the RNM with actual power distribution networks show that the real and the synthetic networks are quite similar (Krishnan et al. 2020). However, the methods used to compare such networks were somewhat adhoc. In contrast, Schweitzer et al. (2017) performed a statistical fit of the distributions of network attributes and performed a numerical comparison, yielding a more rigorous approach to measuring network similarity.

The RNM framework uses four independent layers (namely logical, topological, electrical, and continuity of supply) to assign constraints while creating the networks. Although,

this approach is natural, the set of constraints are not mathematically well-defined to the extent that they can be reproduced. For instance, the framework uses a set of heuristics to satisfy the constraints, which does not always guarantee a feasible solution. Furthermore, several steps in the heuristic-based method involve user-defined parameters, which lead to multiple possible networks for different choices. Furthermore, these papers do not consider the creation of ensembles of networks.

The methods in the second category involve solving one or more optimization problems that often require large computation time (Mateo, Postigo, F. d. Cuadra, et al. 2020). The work by Schweitzer et al. (2017) is one of the earliest papers in the second category. The authors analyze a large dataset of actual distribution networks in the Netherlands. Next, they statistically fit the parameters of a class of stochastic network models to the data. Random networks are then sampled using the stochastic model. This method is reminiscent of "configuration models" used to construct random social networks (Bender and Canfield 1978). Although the generated synthetic networks have random interconnections, the aggregate structural measures are statistically similar to the real networks.

The mathematical programming methods are mostly restricted to distribution system expansion planning (Quiroga, Sauma, and Pozo 2019; You et al. 2016; Trpovski, Recalde, and Hamacher 2018; Byeon et al. 2020) or distribution system restructuring problem (Singh, Kekatos, and C. C. Liu 2019; Singh, Taheri, et al. 2022; Lei et al. 2020). We have followed similar works to construct the mathematical framework for our approach (Steps 1 and 2). However, one of the main differences is that in all these works the feeder nodes are known beforehand. In our approach, the feeder nodes are identified as an output of the proposed optimization framework. The development of such a framework allows us to use it in Step 3 of our algorithm to construct an ensemble of synthetic networks which are statistically similar to each other.

Authors have recently begun developing methods in the third category. Liang et al. have used an extensive dataset of actual distribution networks and trained a generative adversarial neural network (GANN) (Liang et al. 2021). Thereafter, random networks are generated from the trained generative model. Multiple synthetic networks can be generated resembling the actual distribution networks using this approach. The methods are data driven and do not take underlying spatial or engineering constraints into account. These methods usually need a large training dataset. The lack of availability of such datasets makes use of these methods challenging at present.

Recent works (Subbiah et al. 2017; Thorve et al. 2018) have provided detailed synthetic residential demand models along with household geographic footprints. We create synthetic distribution networks connecting substations to these individual residence locations. We propose a rigorous mathematical framework that provides optimality guarantees on the quality of networks created. The ensemble of synthetic networks created can potentially be used to train GANNs.

Our work differs from prior work in the following ways (see Table 2.1 for further discussion).

1. We propose an optimization framework which creates a minimum length network which satisfies standard structural and power flow constraints. Earlier papers, e.g. (Bidel, Schelo, and Hamacher 2021; Saha et al. 2019; Mateo, Postigo, F. d. Cuadra, et al. 2020; Liang et al. 2021) often do not ensure that all the constraints are satisfied while creating the synthetic networks.

2. The heuristics used in the methods proposed in (Z. Wang, Scaglione, and Thomas 2010; Saha et al. 2019; Domingo et al. 2011; Gonzalez-Sotres et al. 2013; Mateo, Postigo, F. d. Cuadra, et al. 2020; Krishnan et al. 2020) create a distribution network

as a minimum spanning tree. But the methods do not ensure that the power engineering voltage constraints are satisfied as per ANSI standards (ANSI 2020). Therefore, the synthesized networks might be topologically realistic but need not be realistic from the perspective of power engineering. The power engineering constraints are satisfied by adding shunt capacitors and voltage regulators, which improves the voltage profile.

3. Most heuristics described in previous papers (Mateo, Postigo, F. d. Cuadra, et al. 2020; Krishnan et al. 2020; Liang et al. 2021; Bidel, Schelo, and Hamacher 2021) deal with the issue of finding locations of substation feeders by clustering residences. In contrast, we start with given locations of substations (Homeland Security 2019) and residences (Thorve et al. 2018) based on real-world data and construct the network connecting these points. Further, in our primary network creation method (Step-2), we have included the optimal substation feeder selection problem effectively in our optimization framework, which has not been used in any prior work.

4. The power flow constraints used in our methodology are similar to the constraints used in distribution feeder planning problems, as in (Trpovski, Recalde, and Hamacher 2018; Byeon et al. 2020; Lei et al. 2020; Rotering et al. 2011). These constraints define the fundamental physics of distribution network operation and hence are similar to some of the published work. In addition, we include two novel components in our optimization framework: (i) Prior works included either a commodity flow model (Lei et al. 2020) to ensure tree structure, or included multiple constraints to avoid the occurrence of cycles (Singh, Kekatos, and C. C. Liu 2019). Here, we have theoretically proved that the power flow constraints are sufficient to ensure a tree structure. (ii) Prior papers on distribution network planning and expansion, which include a similar optimization framework assume that the number of feeders are

known beforehand. In our framework, we do not make such assumptions; rather, the optimal set of feeders are identified as an output of the primary network creation problem.

5. We construct an ensemble of networks, which is an important contribution of this work. We do not consider the optimal network as the sole output; rather we consider it as one random realization of the actual distribution network and propose a framework to create multiple feasible and realistic (but not optimal) networks. Using our well-defined optimization framework in Step-1 and Step-2, we are able to create an ensemble of feasible networks by solving a restricted version of the optimization problem. Prior works have not considered generating an ensemble of synthetic power distribution networks.

Table 2.1: Table showing comparison with other related works

Previous Works	**Includes**	**Excludes**
D'Souza, Brummitt, and Leicht (2014)	I	II,III,IV,V,VI,VII
Soltan and Zussman (2016)	I	II,III,IV,V,VI,VII
Gegner et al. (2016) and Birchfield et al. (2017)	I	II,III,IV,V,VI,VII
Atat et al. (2019)	I	II,III,IV,V,VI,VII
Trpovski, Recalde, and Hamacher (2018)	I,IV,V	II,III,VI,VII
Kadavil, Hansen, and Suryanarayanan (2016)	I	II,III,IV,V,VI,VII
Schweitzer et al. (2017) and Saha et al. (2019)	I,II	III,IV,V,VI,VII
Mateo, Postigo, F. d. Cuadra, et al. (2020)	I,II,V	III,IV,VI,VII
Liang et al. (2021)	I,VII	II,III,IV,V,VI
Bidel, Schelo, and Hamacher (2021)	I,II	III,IV,V,VI,VII

I	Geographic information embedding
II	High resolution residential hourly demand profile
III	Avoids usage of actual distribution networks
IV	Well defined mathematical framework with a guaranteed solution
V	Networks satisfies ANSI voltage constraints
VI	Optimal choice of feeders
VII	Ensemble of networks

2.3 Preliminaries of power distribution system

The power distribution network connects the high voltage (HV; greater than 33kV) distribution substation to low voltage (LV; 208-480V) residential consumers. In most practical networks, this is done through two sets of networks: (i) the medium voltage (MV; usually 6-11kV) primary network connects a step-down transformer at the substation to local pole-top transformers along the road network and (ii) the LV secondary network connects the residences to the local pole-top transformers. Additionally, most distribution networks are operated in a *radial* structure (with no cycles or loops) to facilitate protection coordination (Zamani, Sidhu, and Yazdani 2010). Residences are primarily connected in chains (the degree of a node is at most 2) to avoid branching and thereby maintaining a healthy voltage level. We term this configuration a *star-like tree* since the edges emerge from a single root transformer node and connect residences without further branching. The schematic of a typical power distribution networks is shown in Fig. 2.1.

2.4 Proposed framework

We use open source, publicly available information regarding several infrastructures to generate the synthetic distribution networks: (i) road network data from *Open Street Maps* (2021), (ii) geographic locations of HV (greater than 33kV) substations from data sets published by Homeland Security (2019), and (iii) residential electric power demand information developed in earlier work from Thorve et al. (2018). The details of each dataset are provided in Table 2.2. The rightmost column provides the size of each dataset for Montgomery County in southwest Virginia, USA.

Algorithm 1 summarises the steps we use in the work. The synthetic distribution networks

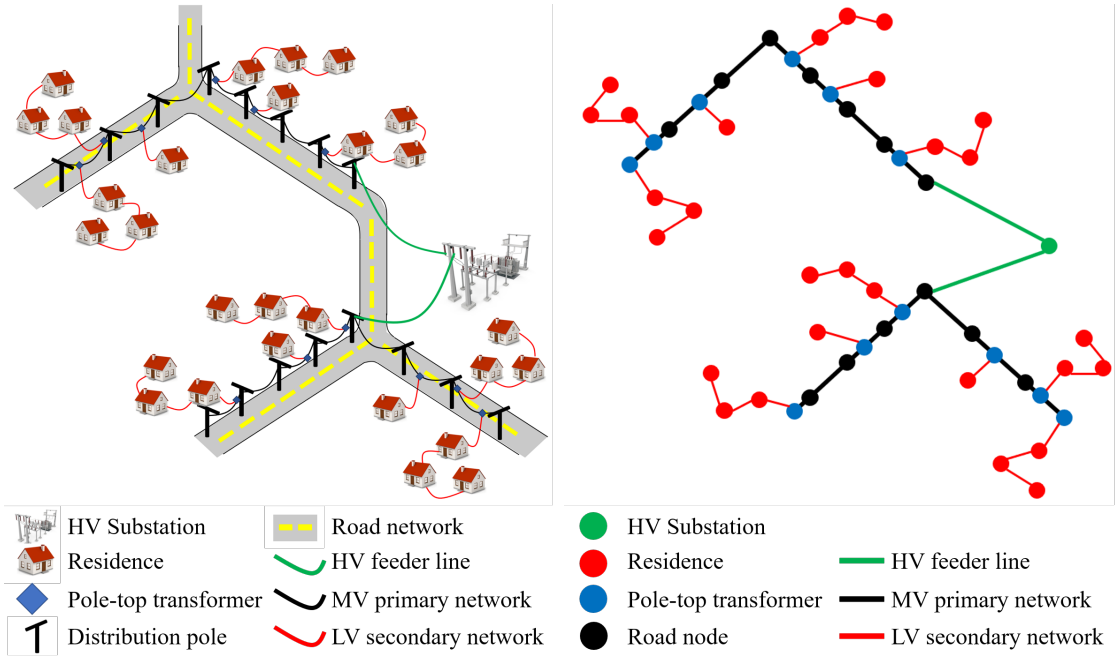

Figure 2.1: Schematic of power distribution network (left) and corresponding graph structure (right). The proposed approach first creates the secondary network (red) connecting the residences to local transformers. Thereafter, the road network (black edges) is used as a proxy to connect the transformers forming the primary network.

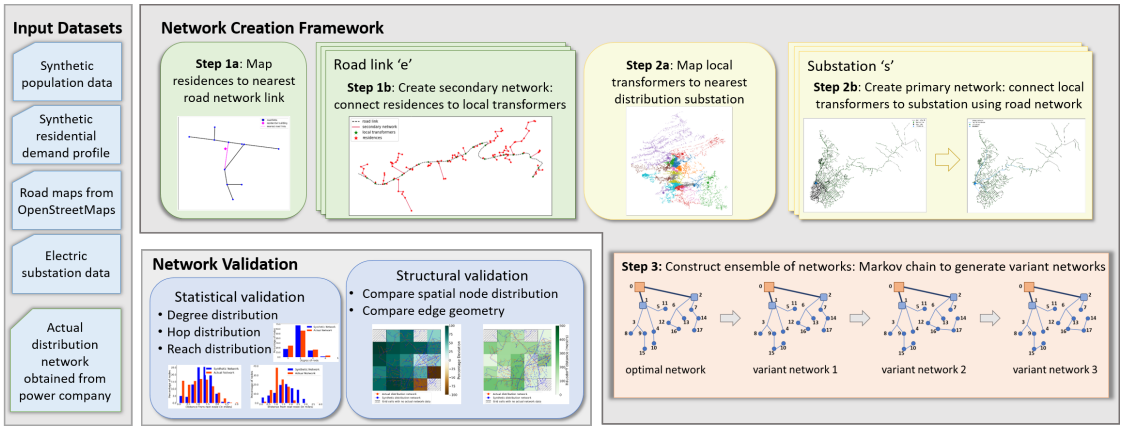

Figure 2.2: Proposed framework for constructing ensembles of realistic power distribution networks. The framework uses the input datasets and constructs an ensemble of networks using the steps detailed in Algorithm 1. The created networks are validated against actual power distribution networks.

Table 2.2: Dataset and related attributes used to generate synthetic distribution network

Dataset	**Source**	**Attributes**	**Example for Montgomery County of Virginia, USA**
Substation	Electric substation data published by Homeland Security (2019)	• substation ID • longitude • latitude	20 substations
Road network	GIS and electronic navigable maps published by *Open Street Maps* (2021)	• node ID • node longitude • node latitude • link ID • link geometry	33882 nodes 41261 edges
Residences	Synthetic population and electric load demand profiles (Thorve et al. 2018)	• residence ID • longitude • latitude • hourly load profile	35629 homes

are constructed in two steps using a bottom-up approach. First, we identify local pole-top transformers along the road network and connect the residential buildings to them to create the LV (208–480V) secondary network (Step 1). Thereafter, we use the road network as a proxy to construct the MV(6–11kV) primary network connecting the local transformers placed along roads to the substations (Step 2). To construct the ensemble of synthetic networks, we propose a Markov chain starting from the already created network to a variant network which is also a feasible distribution network (Step 3). Finally, we add attributes to nodes and edges in each network (Step 4) to create an ensemble of synthetic power distribution networks. Several aspects of the first two and the last steps are similar to the approach taken in earlier papers. The difference lies in the specifics of problem formulation and the resulting algorithmic approach. The third step that creates ensemble of networks has largely not been explored in the context of distribution networks. Fig. 2.2 shows the proposed framework for constructing and validating ensembles of realistic power distribution networks.

Algorithm 1 Create ensemble of synthetic networks

Input: Set of residences $\mathcal{H}$, set of substations $\mathcal{S}$, road network $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$, required ensemble size N

Step 1: Construct LV secondary network.
- a: Map residences to nearest road network link.
- b: Connect residences to local transformers along road link.

Step 2: Construct MV primary network.
- a: Map local transformers to nearest substation.
- b: Use road network as proxy to connect transformers to substation.

Step 3: Construct an ensemble of networks.
- a: Construct Markov Chain $\mathscr{M}$ to create a variant from an existing network.
- b: Run $\mathscr{M}$ to create N variant networks.

Step 4: Add additional attributes to nodes and edges of each network in the ensemble.
- a: Assign one of the three phases (A,B,C) to each residence.
- b: Assign a distribution line type to each edge.

Output: Ensemble of N attributed networks.

2.4.1 Step 1: constructing secondary networks

We extract residence and road network data for the geographic region. Let $\mathcal{H}$ be the set of residences and $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$ be the road network graph. We evaluate a many-to-one mapping $\mathcal{F}_M \colon \mathcal{H} \to \mathcal{E}_R$ such that each residence $h \in \mathcal{H}$ is mapped to the nearest road network link $e \in \mathcal{E}_R$. The inverse mapping $\mathcal{F}_M^{-1}$ defined by $\mathcal{F}_M^{-1}(e) = \{h \in \mathcal{H}; \mathcal{F}_M(h) = e\}$ provides the set of residences assigned to each road link $e \in \mathcal{E}_R$.

The secondary network creation problem (denoted by $\mathscr{P}_{\text{sec}}$) is defined for each road link $e \in \mathcal{E}_R$. The objective is to identify local transformers $\mathcal{V}_T(e)$ along the link and connect them to the assigned residences $\mathcal{F}_M^{-1}(e)$, thereby constructing the secondary distribution network $\mathcal{G}_S(e)$ with node set $\mathcal{V}_S(e) = \mathcal{V}_T(e) \cup \mathcal{F}_M^{-1}(e)$ and edges $\mathcal{E}_S(e)$. We impose structural constraints to connect residences in chains ensuring tree network structure so that the created networks mimic their physical counterpart.

Problem 2.1 ($\mathscr{P}_{\text{sec}}$ **construction).** *Given a road link $e \in \mathcal{E}_R$ with a set of residences $\mathcal{F}_M^{-1}(e)$ assigned to it, construct an optimal forest of trees, $\mathcal{G}_S(e)$, rooted at points (local transform-*

ers) along the link and connecting the residences.

The problem $\mathscr{P}_{\text{sec}}$ is modeled as a mixed integer linear program (MILP) which usually requires exponential computation time. We use different heuristics to reduce the number of binary variables which in turn reduce the overall time complexity. The secondary network creation process can be executed simultaneously for different road links $e \in \mathcal{E}_R$ in the geographic region. In our framework we execute the task sequentially for all edges in a county, with the entire sequence performed simultaneously for different counties. The secondary network generated for the region is

$$\mathcal{G}_S = \bigcup_{e \in \mathcal{E}_R} \mathcal{G}_S(e) = \bigcup_{e \in \mathcal{E}_R} \mathscr{P}_{\text{sec}}\left(e, \mathcal{F}_M^{-1}(e)\right) .$$

Step 1a: map residences to the nearest road network link

This section details the proposed mapping $\mathcal{F}_M \colon \mathcal{H} \to \mathcal{E}_R$ between the set of residences $\mathcal{H}$ and links $\mathcal{E}_R$ of the road network $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$. We map each residence $h \in \mathcal{H}$ to the nearest road network link $e \in \mathcal{E}_R$. Thereafter, we compute the inverse mapping $\mathcal{F}^{-1}(e)$ to identify a set of residential buildings near each road network link $e \in \mathcal{E}_R$. This information will be used in the successive steps to generate the secondary distribution network. We denote the spatial embedding of residence $h \in \mathcal{H}$ as $\mathbf{p}_h \in \mathbb{R}^2$. A road network link $e \in \mathcal{E}_R$ is denoted by $e : (u, v)$, where the nodes $u, v \in \mathcal{V}_R$ have spatial embeddings $\mathbf{p}_u, \mathbf{p}_v \in \mathbb{R}^2$ respectively.

Algorithm 2 is used to compute the nearest road network link to a given residence. First, a bounding region of suitable size is evaluated for each road network link. This is done such that any point in the region is within a radius r from any internal point of the road network link e. The spatial embedding of an interior point along link $e = (u, v)$ is the convex addition of the spatial embedding of the nodes u and v. The bounding region for link $e = (u, v) \in \mathcal{E}_R$

is denoted by $\mathscr{B}_e$.

$$\mathscr{B}_e = \left\{\mathbf{p}; ||\mathbf{p} - \mathbf{p}_e||_2 \ \leq r, \forall \mathbf{p}_e = \theta \mathbf{p}_u + (1-\theta)\mathbf{p}_v, \theta \in [0,1]\right\} \tag{2.1}$$

Similarly, a bounding region is considered for a residential building $h \in \mathcal{H}$ and it is denoted by $\mathscr{B}_h$.

$$\mathscr{B}_h = \left\{\mathbf{p} \middle| ||\mathbf{p} - \mathbf{p}_h||_2 \leq r\right\} \tag{2.2}$$

The intersections between the bounding region of the building and those for the links are stored and indexed in a *quad-tree* data structure (Raphael and Bentley 1974). We identify links $e_1, e_2, \cdots, e_k$ with bounding regions $\mathscr{B}_{e_1}, \mathscr{B}_{e_2}, \cdots, \mathscr{B}_{e_k}$ which intersect with $\mathscr{B}_h$. These k links are comparably nearer to the residential building than the others. Thus, the algorithm reduces the computational burden of evaluating the distance between all road links and residential buildings. Finally, the nearest road network link is identified by computing the minimum perpendicular distance from $\mathbf{p_h}$ as follows.

$$\mathcal{F}_M(h) = \arg \min_{e=(u,v) \in \{e_i\}_{i=1}^{k}} \frac{||(\mathbf{p}_u - \mathbf{p}_h) \times (\mathbf{p}_u - \mathbf{p}_v)||}{||\mathbf{p}_u - \mathbf{p}_v||} \tag{2.3}$$

Algorithm 2 Map residences to the nearest road network link.

Input Set of residences $\mathcal{H}$, road network $\mathcal{G}_{\mathcal{R}}(\mathcal{V}_R, \mathcal{E}_R)$, radius for bounding region r.

Step 1: **for** each link $e \in \mathcal{E}_R$ **do**
Step 2: Define bounding region $\mathscr{B}_e$ using Eq. [2.1].
Step 3: **end for**
Step 4: **for** each residence $h \in \mathcal{H}$ **do**
Step 5: Define bounding region $\mathscr{B}_h$ using Eq. [2.2].
Step 6: Find links $e_1, \cdots, e_k$ with bounding regions $\mathscr{B}_{e_1}, \cdots, \mathscr{B}_{e_k}$ which intersect with $\mathscr{B}_h$.
Step 7: Compute the nearest link to residence h using Eq. [2.3].
Step 8: **end for**

Output Mapping $\mathcal{F}_M : \mathcal{H} \rightarrow \mathcal{E}_R$.

Step 1b: connect residences to local transformers

Here we present the detailed optimization framework for the problem $\mathscr{P}_{\text{sec}}(e)$ to create the secondary distribution network $\mathcal{G}_S(e)(\mathcal{V}_S(e), \mathcal{E}_S(e))$ connecting local transformers $\mathcal{V}_T(e)$ along a given road network link $e \in \mathcal{E}_R$ to the set of residences $\mathcal{V}_H = \mathcal{F}_M^{-1}(e)$ mapped to it. We drop the dependency on e from every notation since we are discussing the same problem for each road link e.

Note that we do not have the set of local transformers $\mathcal{V}_T$ to start with. Hence, we begin with a set of probable local transformers interpolated along the link and denoted by $\mathcal{V}_{\text{prob}}$. We need to connect the set of residences $\mathcal{V}_H$ to actual transformer nodes $\mathcal{V}_T \subseteq \mathcal{V}_{\text{prob}}$ in a forest of *star-like* trees with each tree rooted at a transformer node. A candidate set of edges $\mathcal{E}_D$ between the residences and probable transformer nodes is chosen. The secondary network edges are to be selected from this candidate set. We construct an undirected graph $\mathcal{G}_D := (\mathcal{V}_D, \mathcal{E}_D)$ with node set $\mathcal{V}_D = \mathcal{V}_{\text{prob}} \cup \mathcal{V}_H$ and edge set $\mathcal{E}_D$. We provide the formal problem statement to construct the secondary distribution network.

Formal Problem Statement 1 (Secondary network creation problem)**.** *Given the undirected graph $\mathcal{G}_D(\mathcal{V}_D, \mathcal{E}_D)$, find $\mathcal{E}_S \subseteq \mathcal{E}_D$ such that the induced subgraph network $\mathcal{G}_S(\mathcal{V}_S, \mathcal{E}_S)$ with $\mathcal{V}_S = \mathcal{V}_T \bigcup \mathcal{V}_H$ is a forest of starlike trees, $\mathcal{V}_T \subseteq \mathcal{V}_{\text{prob}}$ is the set of root nodes, and the overall length of the network is minimized.*

Note that a complete graph composed of the residence and transformer nodes can always be considered as the candidate set $\mathcal{E}_D$. Here, a Delaunay triangulation (Preparata and Shamos 1985) of the residential nodes is considered to reduce the size of the problem.

Table 2.3: Sets of nodes and edges in secondary network creation problem for each road link.

Notation	Description
$\mathcal{E}_D$	Set of all candidate edges for the network
$\mathcal{E}_S$	Set of chosen edges in the network
$\mathcal{V}_D$	Set of all possible nodes in the network
$\mathcal{V}_{\text{prob}}$	Set of all probable transformer nodes along link
$\mathcal{V}_T$	Set of actual transformer nodes along link
$\mathcal{V}_H$	Set of residence nodes mapped to the link

Edge weight assignment An edge $l : (i, j) \in \mathcal{E}_D$ is assigned a weight $w(i, j)$

$$w(i,j) = \begin{cases} \infty, & \text{if } i, j \in \mathcal{V}_{\text{prob}} \\ \mathsf{dist}(i,j) + \lambda \mathsf{C}(i,j), & \text{otherwise,} \end{cases}$$

where $\mathsf{dist} : \mathcal{V}_D \times \mathcal{V}_D \to \mathbb{R}$ denotes the geodesic distance between the nodes i and j. The function $\mathsf{C}(i, j)$ penalizes the cost function if the edge l connecting nodes i and j crosses the road network link e and is defined as

$$\mathsf{C}(i,j) = \begin{cases} 0, & \text{if } i, j \text{ are on the same side of the road network link } e \\ 2, & \text{if } i, j \text{ are on the opposite side of the road network link } e \\ 1, & \text{if } i \in \mathcal{V}_{\text{prob}} \text{ or } j \in \mathcal{V}_{\text{prob}}. \end{cases}$$

λ is a weight factor to penalize multiple crossing of edges over the road links. It also penalizes multiple edges emerging from the root node. The weights are stacked in the $|\mathcal{E}_D|$-length vector $\mathbf{w}$. Note that an edge between two probable transformer nodes is assigned a weight of *infinity* which is equivalent to not considering them as candidate edges.

Edge variables We introduce binary variables $x_l \in \{0, 1\}$ for each $l \in \mathcal{E}_D$. Variable $x_l = 1$ indicates that the edge l is present in the optimal topology and $x_l = 0$ denotes otherwise.

Each edge $l := (u, v)$ is assigned a flow variable f_l (arbitrarily) directed from node u to node v. The binary variable and flows can be respectively stacked in $|\mathcal{E}_D|$-length vectors $\mathbf{x}$ and $\mathbf{f}$.

Node variables The average hourly load demand at the i^{th} residence node is denoted by p_i and is strictly positive. We stack these average hourly load demands at all residence nodes in a $|\mathcal{V}_H|$-length vector $\mathbf{p}$. We also consider the reactive power load at each household. Since the models in (Thorve et al. 2018) did not explicitly model the reactive power demand at each residence, we consider a reactive power consumption of $q_i = \gamma p_i$ for each residence i. Here, $\gamma = \tan(\phi)$ with $\cos\phi$ denoting the power factor. For all cases, we consider a power factor of 0.95 which renders $\gamma = 0.33$.

Degree constraint. Statistical surveys on distribution networks (Rotering et al. 2011; Postigo et al. 2017) show that residences along the secondary network are mostly connected in series with at most two neighbors. This is ensured by Eq. [2.4] which limits the degree of residence nodes to 2:

$$\sum_{l:(h,j)} x_l \leq 2, \quad \forall h \in \mathcal{V}_H \tag{2.4}$$

Power flow constraints. For the connected graph $\mathcal{G}_D(\mathcal{V}_D, \mathcal{E}_D)$, we define the $|\mathcal{E}_D| \times |\mathcal{V}_D|$ branch-bus incidence matrix $\mathbf{A}_{\mathcal{G}_D}$ with the entry along l^{th} row and k^{th} column as follows.

$$\mathbf{A}_{\mathcal{G}_D}(l,k) := \begin{cases} 1, & k = i \\ -1, & k = j \\ 0, & \text{otherwise} \end{cases} \qquad \forall l = (i,j) \in \mathcal{E}_D$$

Since the order of rows and columns in $\mathbf{A}_{\mathcal{G}_D}$ is arbitrary, we can partition the columns as $\mathbf{A}_{\mathcal{G}_D} = \begin{bmatrix} \mathbf{A_T} & \mathbf{A_H} \end{bmatrix}$, without loss of generality, where the partitions are the columns corresponding to transformer and residence nodes respectively. We call $\mathbf{A_H}$ the reduced branch-

bus incidence matrix.

Assuming no network losses, Eq. [2.5a] represents the power balance equations at all residence nodes. Note that the optimal network is obtained from $\mathcal{G}_D$ after removing the edges for which $x_l = 0$. Therefore, we need to enforce zero flows f_l for non-existing edges. The constraint Eq. [2.5b] performs this task along with constraining the flows f_l for existing edges to be within pre-specified capacities $\overline{f}$:

$$\mathbf{A_H}^T \mathbf{f} = \mathbf{p} \tag{2.5a}$$

$$-\overline{f}\mathbf{x} \leq \mathbf{f} \leq \overline{f}\mathbf{x} \tag{2.5b}$$

Ensuring radial topology The radiality requirement of the secondary network $\mathcal{G}_S$ can be enforced from a known graph theory property: *a forest with n nodes and m root nodes has $n-m$ edges*. In our case, $|\mathcal{V}_H| + |\mathcal{V}_T|$ nodes need to be covered in a forest of trees with $|\mathcal{V}_T|$ root nodes, which leads us to the following constraint:

$$\sum_{l \in \mathcal{E}_D} x_l = |\mathcal{V}_H| \tag{2.6}$$

However, we need to ensure that there are no disconnected cycles in the optimal network. This can be done by ensuring that the residence points are connected to a transformer node (Singh, Taheri, et al. 2022; Lei et al. 2020). In our case, this condition is satisfied by the node power flow condition in Eq. [2.5a] if all the residential nodes consume non-zero power.

Proposition 2.2. *The graph $\mathcal{G}_S(\mathcal{V}_S, \mathcal{E}_S)$ with reduced branch-bus incidence matrix $\mathbf{A_H}$ (corresponding to columns of $\mathcal{V}_H$) and node power demand vector $\mathbf{p} \in \mathbb{R}^{|\mathcal{V}_H|}$, with strictly posi-*

tive entries, has exactly $|\mathcal{V}_T|$ connected components if and only if there exists $\mathbf{f} \in \mathbb{R}^{|\mathcal{E}_D|}$ such that Eq. [2.5a] is satisfied.

Proof. Proving by contradiction, suppose $\mathcal{G}_S(\mathcal{V}_S, \mathcal{E}_S)$ has more than $|\mathcal{V}_T|$ connected components and there exists $\mathbf{f} \in \mathbb{R}^{|\mathcal{E}_D|}$ satisfying the proposed equality. Therefore, there exists a connected component $\mathcal{G}_C(\mathcal{V}_C, \mathcal{E}_C)$ which is a maximal connected subgraph with $\mathcal{V}_C \subset \mathcal{V}_H$ and $\mathcal{V}_C \bigcap \mathcal{V}_T = \emptyset$. Let $\mathbf{A}_C$ denote the bus incidence matrix of $\mathcal{G}_C$. By definition, it holds that $\mathbf{A}_C \mathbf{1} = \mathbf{0}$.

Since graph $\mathcal{G}_C(\mathcal{V}_C, \mathcal{E}_C)$ is a maximal connected subgraph of $\mathcal{G}_S$, there exists no edge (i, j) with $i \in \mathcal{V}_C$ and $j \in \mathcal{V}_{\overline{C}}$, where $\mathcal{V}_{\overline{C}} = \mathcal{V}_S \setminus \mathcal{V}_C$. Since the order of rows and columns of $\mathbf{A_H}$ are arbitrary, we can partition without loss of generality as

$$\mathbf{A_H} = \begin{bmatrix} \mathbf{A}_{\overline{C}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_C \end{bmatrix}.$$

We can partition vectors $\mathbf{f}$ and $\mathbf{p}$ conformably to $\mathbf{A_H}$ to get the following equality

$$\begin{bmatrix} \mathbf{A}_{\overline{C}} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_C \end{bmatrix}^T \begin{bmatrix} \mathbf{f}_{\overline{C}} \\ \mathbf{f}_C \end{bmatrix} = \begin{bmatrix} \mathbf{p}_{\overline{C}} \\ \mathbf{p}_C \end{bmatrix},$$

where $\mathbf{p}_C$ and $\mathbf{p}_{\overline{C}}$ are respectively the vectorized power demands for node sets $\mathcal{V}_C$ and $i \in \mathcal{V}_{\overline{C}}$. From the second block, it implies that

$$\mathbf{p}_C = \mathbf{A}_C^T \mathbf{f}_C \Rightarrow \mathbf{1}^T \mathbf{p}_C = \mathbf{1}^T \mathbf{A}_C^T \mathbf{f}_C = \mathbf{0} \tag{2.7}$$

Since all entries of $\mathbf{p}$ are strictly positive from the initial assumption, we have $\mathbf{1}^T \mathbf{p}_C \neq 0$ which contradicts Eq. [2.7] and completes the proof. □

Generating optimal network topology The optimal secondary network is obtained after solving the optimization problem.

$$\begin{aligned} &\min_{\mathbf{x}} \mathbf{w}^T \mathbf{x} \\ &\text{s.t. Eqs. [2.4], [2.5], [2.6]} \end{aligned} \tag{2.8}$$

Though the mixed integer optimization problem (MILP) in Eq. [2.8] has been relaxed, the time complexity depends on the size of the edge set $\mathcal{E}_D$. The computation time can be significantly reduced by considering the edges of a Delaunay triangulation of nodes $\mathcal{V}_H$ as the edge set instead of choosing $\binom{|\mathcal{V}_H|}{2}$ combinatorial edges between the nodes. The MILP is solved individually for each road link $e \in \mathcal{E}_R$ to construct the secondary network $\mathcal{G}_S(e)$ corresponding to the road link. The overall secondary network is a combination of such generated networks.

2.4.2 Step 2: constructing primary networks

The secondary network results in local transformer nodes $\mathcal{V}_T = \bigcup_{e \in \mathcal{E}_R} \mathcal{V}_T(e)$ along the road network links. The goal of the primary network construction is to connect these transformers to the set of substation nodes $\mathcal{S}$ using the road network as proxy. First, we define a many-to-one mapping $\mathcal{F}_V \colon \mathcal{V}_T \to \mathcal{S}$ based on a Voronoi partitioning. The details of this mapping are provided in the Appendix. We are interested in the inverse mapping $\mathcal{F}_V^{-1}(s) = \{t \in \mathcal{V}_T; \mathcal{F}_V(t) = s\}$ which assigns a group of transformers to each substation node.

The primary network creation problem (denoted by $\mathscr{P}_{\text{prim}}$) is defined for each substation node $s \in \mathcal{S}$, and the goal is to crate a minimum length primary network $\mathcal{G}_P(s)$ connecting substation node s to the mapped transformers $\mathcal{F}_V^{-1}(s)$ using road network $\mathcal{G}_R$ as proxy,

such that the following set of structural and operational constraints are valid: (i) the network should be a tree rooted at the substation, (ii) all transformer nodes are to be connected, and (iii) all nodes should have acceptable voltages (based on American National Standards Institute (ANSI) standards between 0.95 and 1.05 per unit (pu)) when the residential customers are consuming average hourly loads.

Problem 2.3 ($\mathscr{P}_{\text{prim}}$ construction)**.** *Given a substation $s \in \mathcal{S}$ with an assigned set of local transformer nodes $\mathcal{F}_V^{-1}(s)$, construct a tree network $\mathcal{G}_P(s)$ using the road network $\mathcal{G}_R$ as a proxy which connects all local transformers while ensuring acceptable node voltages by power engineering standards.*

We formulate an MILP to solve the problem $\mathscr{P}_{\text{prim}}$ which requires exponential computation time. We do not use any heuristic to reduce the computational complexity which is determined by the size of underlying road network (used as the proxy). On many occasions, we terminate the optimization program reaching an optimal solution, in order to reduce the running time. This has resulted in the constructed network being a near-optimal solution, but with an acceptable optimality gap of $0-5\%$. In our framework we execute the task of primary network creation simultaneously for all the substations in the geographic region. The created primary network $\mathcal{G}_P$ for the entire region is,

$$\mathcal{G}_P = \bigcup_{s\in\mathcal{S}} \mathcal{G}_P(s) = \bigcup_{s\in\mathcal{S}} \mathscr{P}_{\text{prim}}\left(s, \mathcal{F}_V^{-1}(s)\right) .$$

Step 2a: map local transformers to the nearest substation

The secondary network results in local transformer nodes $\mathcal{V}_T = \bigcup_{e\in\mathcal{E}_R} \mathcal{V}_T(e)$ along the road network links of the geographic region (in our case, throughout the state of Virginia). The task of primary network creation is to connect these transformer nodes to substation nodes

in the region. In order to tackle the size of the problem, we partition the task into a number of subtasks. We perform this partition such that each substation is required to connect only the transformer nodes near to it. Therefore, we produce a map $\mathcal{F}_V$ from the set of local transformers $\mathcal{V}_T$ to the set of substations $\mathcal{S}$. Since we use the road network as a proxy for creating primary networks, we map each local transformer to the nearest substation along the road network.

Algorithm 3 Map local transformer to nearest substation.

Input Set of substations $\mathcal{S}$, set of local transformers $\mathcal{V}_T$, road network $\mathcal{G}_{\mathcal{R}}(\mathcal{V}_R, \mathcal{E}_R)$.

Step 1: **for** each substation $s \in \mathcal{S}$ **do**
Step 2: Find the nearest road network node $v_s \in \mathcal{V}_R$.
Step 3: **end for**
Step 4: **for** each local transformer $t \in \mathcal{V}_T$ **do**
Step 5: Compute the nearest substation using Eq. [2.9].
Step 6: **end for**

Output Mapping $\mathcal{F}_V : \mathcal{V}_T \rightarrow \mathcal{S}$.

Algorithm 3 details the steps involved in creating this mapping. First, we find the geographically located nearest road network node $v_s \in \mathcal{V}_R$ for each substation $s \in \mathcal{S}$. Let $\mathsf{NetDist}(u, v)$ denotes the shortest path distance between nodes $u, v \in \mathcal{V}_R$ along road network $\mathcal{G}_R$. We identify the nearest substation to each local transformer using

$$\mathcal{F}_V(v) = \arg\min_{s \in \mathcal{S}} \mathsf{NetDist}(v_s, v). \tag{2.9}$$

The inverse mapping $\mathcal{F}_V^{-1}(s)$ assigns a set of local transformer nodes to each substation $s \in \mathcal{S}$. These sets of local transformer nodes are mutually exclusive and are exhaustive. We perform the primary network creation for each of these sets individually.

Step 2b: connect local transformers to the substation

Here we present the detailed optimization framework for the problem $\mathscr{P}_{\text{prim}}(s)$ to create the primary distribution network $\mathcal{G}_P(s)(\mathcal{V}_P(s), \mathcal{E}_P(s))$ connecting substation s to local transformers $\mathcal{V}_T = \mathcal{F}_V^{-1}(s)$ mapped to it. We drop the dependency on s from every notation since we are discussing the same problem for each substation s.

We start with an undirected graph $\mathcal{G}_R(\mathcal{V}, \mathcal{E}_R)$ which is the road network subgraph induced from the mapped transformer nodes. Here $\mathcal{V} = \mathcal{V}_R \bigcup \mathcal{V}_T$ comprises of transformer nodes as well as road nodes. The goal of the primary network creation problem is to select the edge set $\mathcal{E}_P \subseteq \mathcal{E}_R$ to generate the optimal primary network $\mathcal{G}_P(\mathcal{V}_P, \mathcal{E}_P)$. Note that all transformer nodes are required to be included. In contrast, road nodes are dummy points with no load and are only used to connect the local transformers. For example, we cannot have a road node as a leaf node since it does not connect transformer nodes. The node set $\mathcal{V}_P = \mathcal{V}_R^\star \bigcup \mathcal{V}_T$ comprises road and transformer nodes respectively where $\mathcal{V}_R^\star \subseteq \mathcal{V}_R$ are the selected road nodes.

In all practical distribution networks, multiple feeder lines originate from a substation and connect the local transformers. To this end, we construct the primary network $\mathcal{G}_P$ as a forest of trees with the root of each tree connected to the substation through high voltage feeder lines. These root nodes are road network nodes with no loads connected to them directly. Note that some nodes in $\mathcal{V}_R^\star$ which are the root nodes in the constructed primary network. Here, we present the formal problem statement for constructing the primary distribution network.

Formal Problem Statement 2 (Primary network creation problem)**.** *Given a connected network $\mathcal{G}_R(\mathcal{V}, \mathcal{E}_R)$ where $\mathcal{V} = \mathcal{V}_R \bigcup \mathcal{V}_T$, find $\mathcal{E}_P \subseteq \mathcal{E}_R$ such that the induced subgraph network $\mathcal{G}_P(\mathcal{V}_P, \mathcal{E}_P)$ is a forest of trees with each tree rooted at some $r \in \mathcal{V}_R^\star$ where $\mathcal{V}_P =$*

$\mathcal{V}_R^\star \bigcup \mathcal{V}_T$ *and* $\mathcal{V}_R^\star \subseteq \mathcal{V}_R$.

Binary variables. We assign a binary variable $x_e \in \{0,1\}$ for each edge $e \in \mathcal{E}_R$. Variable $x_e = 1$ indicates that the edge e is included in the network, and $x_e = 0$ otherwise. Further, we introduce binary variables $y_r, z_r \in \{0,1\}$ for each road network node $r \in \mathcal{V}_R$. Variable $y_r = 1$ indicates that road node r is part of the primary network and vice versa. A selected road node ($y_r = 1$) may be chosen to be a root node or otherwise. Binary variable $z_r = 0$ indicates that road node r is a root node, and $z_r = 1$ implies that r is not a root node. Note that a road node which is not selected ($y_r = 0$) needs to be treated as a non-root node ($z_r = 1$) and hence we have the following.

$$1 - z_r \leq y_r, \quad \forall r \in \mathcal{V}_R \tag{2.10}$$

Table 2.4: Sets of nodes and edges in primary network creation problem

Notation	Description
$\mathcal{E}_R$	Set of all candidate edges in the network
$\mathcal{V}$	Set of all nodes in the network
$\mathcal{V}_T$	Set of all transformer nodes in the network
$\mathcal{V}_R$	Set of all road nodes in the network
$\mathcal{E}_P$	Set of all chosen edges in the optimal primary network
$\mathcal{V}_P$	Set of all chosen nodes in the optimal primary network
$\mathcal{V}_R^\star$	Set of all chosen road nodes in the optimal primary network

Table 2.5: Binary variables in primary network creation problem

Notation	Description
x_e	1 if edge $e \in \mathcal{E}_R$ is included in optimal network
y_r	1 if road node $r \in \mathcal{V}_R$ is included in optimal network
z_r	0 if road node $r \in \mathcal{V}_R$ is a root node

Topology constraints. We need to ensure the following: (i) a selected non-root road node

should not be a terminal node, (ii) the network does not consist of any cycles, and (iii) the network covers all transformer nodes.

Let $e = (r, j) \in \mathcal{E}_R$ denote an edge which is incident on the road node $r \in \mathcal{V}_R$. The degree of a road node r in graph $\mathcal{G}_R$ is denoted by $\sum_{e:(r,j)} x_e$. For the first condition to hold true, we need to ensure that the degree of a selected non-root road node (with $y_r = 1, z_r = 1$) is at least 2, and the degree of an unselected road node (with $y_r = 0$) is 0. Finally, the degree of a selected road node which is also a root node (with $y_r = 1, z_r = 0$) is positive. This is ensured through the following inequalities:

$$\sum_{e=(r,j)} x_e \leq |\mathcal{E}_R| y_r, \qquad \forall r \in \mathcal{V}_R \tag{2.11a}$$

$$\sum_{e=(r,j)} x_e \geq 2(y_r + z_r - 1), \qquad \forall r \in \mathcal{V}_R \tag{2.11b}$$

$$\sum_{e=(r,j)} x_e \geq y_r, \qquad \forall r \in \mathcal{V}_R \tag{2.11c}$$

To enforce the second condition (no cycles in the network) or 'radiality' condition, we use results from graph theory. We know that a forest with n nodes and m components has $n - m$ edges. In our case, the total number of nodes is $|\mathcal{V}_T| + \sum_{r \in \mathcal{V}_R} y_r$, while the number of components is the number of root nodes, i.e, $\sum_{r \in \mathcal{V}_R} (1 - z_r)$. Therefore, the radiality constraint is given by

$$\sum_{e \in \mathcal{E}_R} x_e = |\mathcal{V}_T| + \sum_{r \in \mathcal{V}_R} y_r - \sum_{r \in \mathcal{V}_R} (1 - z_r). \tag{2.12}$$

However, this is not a sufficient condition for radiality since it does not avoid the formation of disconnected components with cycles. We can extend Proposition 2.2 and ensure radiality with the power balance equations at each node in the network.

Table 2.6: Power flow variables and parameters in primary network creation problem

Notation	Description
$V_j = v_j e^{\mathrm{j}\theta_j}$	complex bus voltage phasor at bus $j \in \mathcal{V}$; magnitude: v_j, angle: θ_j
I_e	complex current flowing through edge $e \in \mathcal{E}_R$
$S_{ij} = P_{ij} + \mathrm{j}Q_{ij}$	complex power flowing through edge $e = (i,j) \in \mathcal{E}_R$ from node i to node j; real power flow: P_{ij}, reactive power flow: Q_{ij}
$s_j = p_j + \mathrm{j}q_j$	complex power injection at bus $j \in \mathcal{V}$; real power injection: p_j, reactive power injection: q_j
$Z_e = R_e + \mathrm{j}X_e$	complex impedance of edge $e \in \mathcal{E}_R$; resistance: R_e, reactance: X_e
f_e	power flow limit in edge $e \in \mathcal{E}_R$
$\overline{s}_j$	substation feeder rating
$\underline{v}, \overline{v}$	lower and upper limits of bus voltage

Power balance constraints. The complex power injection at bus $j \in \mathcal{V}$ is denoted by $s_j = p_j + \mathrm{j}q_j$. Note that the power injections are the difference between generation and load demand at each bus. For transformer nodes, the power injection is the negative load demand; for non-root road nodes, the power injection is zero; and for root nodes, the power injection is bounded by a defined limit. It is assumed that the root nodes are connected to the substation through high voltage feeders. Therefore, the feeder capacity can be considered as the defined limit.

$$\sum_{e=(i,j)} x_e S_{ij} - \sum_{e=(j,k)} x_e S_{jk} = -s_j, \quad \forall j \in \mathcal{V}_T \tag{2.13a}$$

$$\sum_{e=(i,j)} x_e S_{ij} - \sum_{e=(j,k)} x_e S_{jk} = 0, \quad \forall j \in \mathcal{V}_R, z_j = 1 \tag{2.13b}$$

$$\left| \sum_{e=(i,j)} x_e S_{ij} - \sum_{e=(j,k)} x_e S_{jk} \right| \leq \overline{s}_j, \quad \forall j \in \mathcal{V}_R, z_j = 0 \tag{2.13c}$$

In the secondary network creation method, we discuss the residential reactive power model. We assumed a constant power factor of 0.95 for all residences. This leads to a linear relationship between the real and reactive power throughout the network with $Q_{ij} = \gamma P_{ij}$ for all edges. We can drop either of the real or reactive power terms in Eq. [2.13], since they

would lead to the same equality constraints. Such assumptions are standard for problems related to network planning (Rotering et al. 2011; Singh, Taheri, et al. 2022; Lei et al. 2020). Therefore, we can rewrite Eq. [2.13] by dropping the reactive power terms (imaginary parts) and combining Eq. [2.13b] and Eq. [2.13c] to get:

$$\sum_{e=(i,j)} P_{ij} - \sum_{e=(j,k)} P_{jk} = -p_j, \quad \forall j \in \mathcal{V}_T \tag{2.14a}$$

$$-\overline{p}_j(1-z_j) \leq \sum_{e=(i,j)} P_{ij} - \sum_{e=(j,k)} P_{jk} \leq \overline{p}_j(1-z_j), \forall j \in \mathcal{V}_R \tag{2.14b}$$

Power flow constraints. We define complex node voltage $V_j = v_j e^{\mathrm{j}\theta_j}$ (magnitude v_j and angle θ_j) for each node $j \in \mathcal{V}$ and complex power $S_{i,j} = P_{i,j} + \mathrm{j}Q_{i,j}$ for each edge $e = (i,j) \in \mathcal{E}_R$ flowing from node i to j, where $P_{i,j}$ and $Q_{i,j}$ respectively denote real and reactive power flowing along edge $e \in \mathcal{E}_R$. The current flowing through the edge is defined by I_e and the complex impedance of the edge is denoted by $Z_e = R_e + \mathrm{j}X_e$. The following equations/inequalities describe the relationship between the variables and the associated constraints. Note that the constraints are only activated for those edges which are selected in the optimal network. Therefore, we have introduced the binary variable x_e in Eq. [2.15a]. We assume that HV feeder lines from the substation to the root nodes in the optimal primary distribution network end in voltage regulators, which ensure that the root nodes have a

voltage of 1.0 per unit (pu).

$$x_e(V_i - V_j - Z_e I_e) = 0, \qquad \forall e \in \mathcal{E}_R \tag{2.15a}$$

$$S_{ij} = V_j I_e^\star, \qquad \forall e \in \mathcal{E}_R \tag{2.15b}$$

$$|S_{ij}| \leq \overline{f}_e, \qquad \forall e \in \mathcal{E}_R \tag{2.15c}$$

$$\underline{v} \leq v_j \leq \overline{v}, \qquad \forall j \in \mathcal{V} \tag{2.15d}$$

$$v_j = 1, \qquad \forall j \in \mathcal{V}, z_j = 0 \tag{2.15e}$$

Generating optimal primary network. Each edge $e = (i, j) \in \mathcal{E}_R$ is assigned a weight $w_e = w(i, j) = \mathsf{dist}(i, j)$ which is the geodesic distance between the nodes. Additionally, for every road node $j \in \mathcal{V}_R$, we compute its geodesic distance from the substation s, denoted by d_j. The optimal primary network topology is obtained by solving the optimization problem:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \sum_{e \in \mathcal{E}_R} x_e w_e + \sum_{j \in \mathcal{V}_R} (1 - z_j)\mathsf{d}_j \\ &\text{s.t. Eqs. [2.10], [2.11], [2.12], [2.14], [2.15]} \end{aligned} \tag{2.16}$$

Relaxing non-linear power flow constraints. Note that power flow constraints in Eq. [2.15] are non-convex because of the quadratic equality constraint in Eq. [2.15b] and the bilinear terms in Eq. [2.15a]. First, we consider relaxing the quadratic equality constraint. Multiplying by the complex conjugates in Eqs. [2.15a] and [2.15b], we get the following result:

$$v_j^2 = v_i^2 - 2R_e P_{ij} - 2X_e Q_{ij} + (R_e^2 + X_e^2)|I_e|^2 \tag{2.17}$$

Assuming small R_e and X_e, the current magnitudes can be eliminated. Note that Eq. [2.17]

is still non-linear in v_i and v_j; however, it is linear in squared magnitude. This relaxed model is known in the literature as the Linearized Distribution Flow (LDF) model. The squared voltage is often approximated as $|v_i|^2 \approx 2v_i - 1$ which leads to the relation $v_j = v_i - R_e P_{i,j} - X_e Q_{i,j}$ (Bolognani and Dörfler 2015). Notice that the constraint in Eq. [2.17] is required to be enforced on only those edges $e \in \mathcal{E}_R$ for which $x_e = 1$, which leads us to the relaxed version of Eq. [2.15a] as

$$x_e(v_i - v_j - R_e P_{ij} - X_e Q_{ij}) = 0\,. \tag{2.18}$$

Here Eq. [2.18] is non-convex due to the bi-linear terms in the equality. In order to deal with this non-convexity, McCormick relaxation has been used widely in several previous works (Singh, Kekatos, and C. C. Liu 2019; Singh, Taheri, et al. 2022). In general, McCormick relaxation replaces the non-convex equality constraint with its convex envelope (McCormick 1976). However, in the case of bilinear variables with at least one binary variable, this relaxation becomes exact. The convex relaxed version of Eq. [2.15] is:

$$-(1-x_e)M \le v_i - v_j - R_e P_{ij} - X_e Q_{ij} \le (1-x_e)M,\ \forall e \in \mathcal{E}_R \tag{2.19a}$$

$$-\overline{f}_e x_e \le P_{ij} \le \overline{f}_e \mathbf{x_e}, \quad \forall e \in \mathcal{E}_R \tag{2.19b}$$

$$(1-v_r) \le z_r, \quad \forall r \in \mathcal{V}_R \tag{2.19c}$$

$$\underline{v} \le v_j \le \overline{v}, \quad \forall j \in \mathcal{V} \tag{2.19d}$$

The overall relaxed optimization problem for creating the primary network topology is:

$$\begin{aligned} &\min_{\mathbf{x},\mathbf{y},\mathbf{z}} \sum_{e\in\mathcal{E}_R} x_e w_e + \sum_{j\in\mathcal{V}_R} (1-z_j) d_j \\ &\text{s.to. Eqs. } [2.10], [2.11], [2.12], [2.19], [2.14] \end{aligned} \tag{2.20}$$

2.4.3 Step 3: constructing ensembles of networks

In this section, we address the problem of creating multiple realizations of the distribution network which connect the residences to substations. Albeit the modification of user defined parameters in $\mathscr{P}_{\text{sec}}$ and $\mathscr{P}_{\text{prim}}$ can produce different realizations of synthetic networks, the procedure is computationally expensive, since optimization problems of similar order need to be solved. We propose a methodology which uses the already created (near)-optimal primary network for a region and creates an ensemble of synthetic networks by reconnecting the transformer nodes in a different manner from the (near)-optimal primary network, while maintaining the structural and power engineering operational constraints. Thereafter, we connect the residences in the same way as in the optimal secondary network. Thus, we construct an ensemble of networks where each network is a combination of a variant primary network and the optimal secondary network (solution of $\mathscr{P}_{\text{sec}}$). The variant primary networks are "feasible" (but not necessarily "optimal") solutions of $\mathscr{P}_{\text{prim}}$.

Problem 2.4. *Given a near-optimal primary network $\mathcal{G}_P^0 := (\mathcal{V}_0, \mathcal{E}_0)$ constructed using underlying road network graph $\mathcal{G}_R := (\mathcal{V}_R, \mathcal{E}_R)$, construct N variants of the primary network $\mathcal{G}_P^1, \cdots, \mathcal{G}_P^N$ by identifying respective edge sets $\mathcal{E}_1, \cdots, \mathcal{E}_N \subseteq \mathcal{E}_R$ such that the networks are feasible solutions of $\mathscr{P}_{\text{prim}}$.*

We consider the ensemble of networks generation problem for each substation s and the mapped transformer nodes $\mathcal{F}_V^{-1}(s)$. Let $\mathscr{F}_{\text{feas}}$ denote the set of feasible solutions of $\mathscr{P}_{\text{prim}}\left(s, \mathcal{F}_V^{-1}(s)\right)$.

From here on we omit the dependency on s in our notation. We design a Markov chain $\mathscr{M}$ to create variant networks with each state denoting a feasible realization of the network $\mathcal{G}_P^t \in \mathscr{F}_{\text{feas}}$. The steps involved in transitioning from the primary network $\mathcal{G}_P^t := (\mathcal{V}_t, \mathcal{E}_t)$ to $\mathcal{G}_P^{t+1} := (\mathcal{V}_{t+1}, \mathcal{E}_{t+1})$ are described below.

Let $\mathscr{F}_{\text{rstr}}(e) = \{\mathcal{G} := (\mathcal{V}, \mathcal{E}) \in \mathscr{F}_{\text{feas}} : e \notin \mathcal{E}\}$. If $\mathscr{F}_{\text{rstr}}(e) \neq \emptyset$, we select a random edge $e \in \mathcal{E}_t$ to be deleted with probability $1/|\mathcal{E}_t|$, and then pick $\mathcal{G}_P^{t+1} := (\mathcal{V}_{t+1}, \mathcal{E}_{t+1}) \in \mathscr{F}_{\text{rstr}}(e)$ uniformly at random; else $\mathcal{G}_P^{t+1} = \mathcal{G}_P^t$. The ensemble of synthetic power distribution networks for the region is

$$\mathscr{E} := \mathcal{G}_S \bigcup \{\mathcal{G}_P^t : t = 1, \ldots, N\}$$

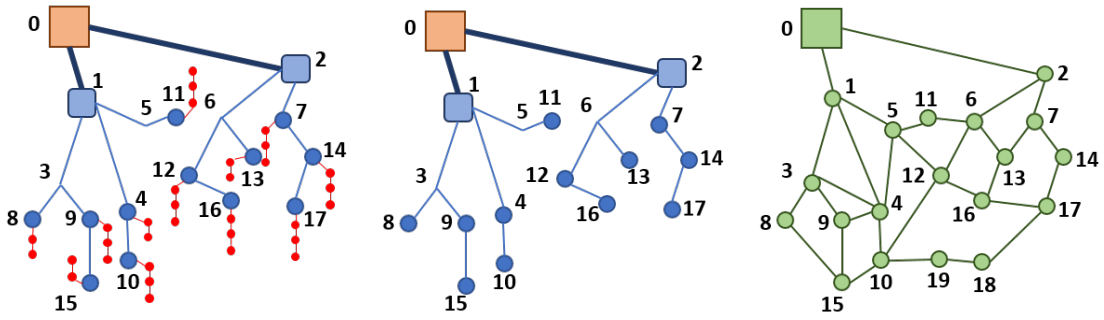

Figure 2.3: Plots summarizing the optimal distribution network constructed using Steps 1 and 2. The overall network (left figure) consists of the primary network (blue) and secondary network (red). The optimal primary network (middle figure) is constructed from the underlying road network (right figure). The edges in the primary network are a subset of the road network graph edges.

Using the optimization frameworks in Eqs. [2.8] and [2.20], we create the optimal (sometimes near-optimal) distribution network as shown in Fig. 2.3 (left). It consists of the primary network which connects the substation to the local transformers in a tree structure, and these transformers are connected to the residences in chains, forming the secondary network. Fig. 2.3 (middle) shows the near-optimal primary network denoted by $\mathcal{G}_P^0(\mathcal{V}_P^0, \mathcal{E}_P^0)$.

Note that the MILP Eq. [2.20] identified $\mathcal{G}_P^0(\mathcal{V}_P^0, \mathcal{E}_P^0)$ as the solution by choosing edges from the underlying road network $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$. Fig. 2.3 (right) shows the underlying road network corresponding to the primary network. We note that all primary network edges are selected from the road network graph. The resulting primary network is a tree, and it connects all the transformer nodes. The road network nodes $\{3,5,6\}$ are used to construct the network, whereas some other road network nodes such as $\{18,19\}$ are not included in the generated primary network.

Algorithm 4 Constructing ensemble of networks.

Input Near-optimal primary network $\mathcal{G}_P^0(\mathcal{V}_P^0, \mathcal{E}_P^0)$, optimal secondary network $\mathcal{G}_S$, underlying road network $\mathcal{G}_R(\mathcal{V}_R, \mathcal{E}_R)$, size of ensemble N.

Step 1: Initialize network count, $t \leftarrow 0$
Step 2: **while** $t < N$ **do**
Step 3: Choose a random set of edges $\mathcal{E}_D \subset \mathcal{E}_P^t$ to be deleted.
Step 4: Formulate optimization problem $\mathscr{P}_t$ by combining Eq. [2.20] with constraints in Eq. [2.21].
Step 5: **if** $\mathscr{P}_t$ is feasible **then**
Step 6: Solve $\mathscr{P}_t$ to get the variant primary network $\mathcal{G}_P^{t+1}(\mathcal{V}_P^{t+1}, \mathcal{E}_P^{t+1})$.
Step 7: Increment network count $t \leftarrow t+1$.
Step 8: **end if**
Step 9: Augment the secondary network to the variant primary network; $\mathcal{G}^t \leftarrow \mathcal{G}_S \bigcup \mathcal{G}_P^t(\mathcal{V}_P^t, \mathcal{E}_P^t)$.
Step 10: **end while**

Output Ensemble of networks $\mathcal{G}^1, \mathcal{G}^2, \cdots, \mathcal{G}^N$.

We model the variant network creation process as a Markov chain $\mathscr{M}$ where each state denotes a valid realization of the primary network. The transitions $\mathcal{G}_P^t(\mathcal{V}_P^t, \mathcal{E}_P^t) \rightarrow \mathcal{G}_P^{t+1}(\mathcal{V}_P^{t+1}, \mathcal{E}_P^{t+1})$ of $\mathscr{M}$ are specified in the following manner. Let s denote the substation node in the network $\mathcal{G}_P^t(\mathcal{V}_P^t, \mathcal{E}_P^t)$ and $\mathcal{N}(s)$ denote the neighbors of s. We choose a random set of edges $\mathcal{E}_D = \{e_d \in \mathcal{E}_P^t | e_d = (u,v), u \neq v \neq s\}$ to be deleted and thereafter solve a restricted version

of Eq. [2.20] with the following additional constraints:

$$x_e = 0, \text{ for } e \in \mathcal{E}_D; \quad x_e = 1, \forall e \in \mathcal{E}_P^t \setminus \mathcal{E}_D \tag{2.21a}$$

$$y_r = 1, \forall r \in \mathcal{V}_R \cap \mathcal{V}_P^t \tag{2.21b}$$

$$z_r = 0, \forall r \in \mathcal{N}(s); \quad z_r = 1, \forall r \in \mathcal{V}_R \setminus \mathcal{N}(s) \tag{2.21c}$$

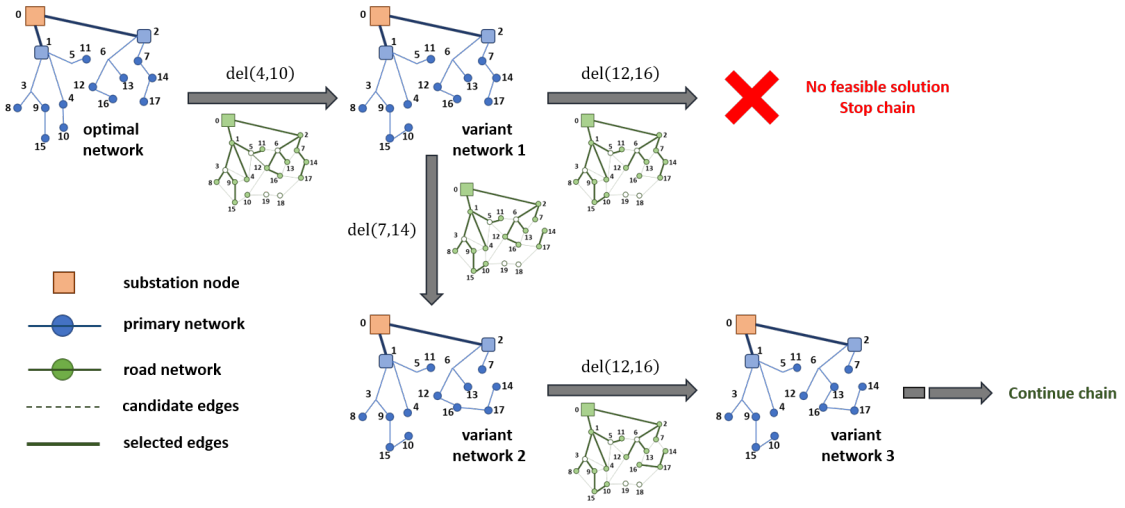

Figure 2.4: Schematic of the Markov chain process to generate an ensemble of primary networks from the optimal primary distribution network. Every transition in the Markov chain involves deletion of a random edge from the primary network followed by solving a restricted version of the primary network creation problem. If no feasible solution exists for the restricted version of the problem, the chain transitions to itself.

The restrictions reduce the order of the optimization problem Eq. [2.20] to a significant extent and thereby requires less time to reach an optimal solution. Eq. [2.21a] forces all edges in the set $\mathcal{E}_P^t \setminus \mathcal{E}_D$ to be selected in the new network $\mathcal{G}_P^{t+1}(\mathcal{V}_P^{t+1}, \mathcal{E}_P^{t+1})$. The edge set $\mathcal{E}_D$ is not selected and the optimization problem chooses edges from $\mathcal{E}_R \setminus \mathcal{E}_P^t$. Further, Eq. [2.21b] forces all road network nodes in $\mathcal{G}_P^t$ to be selected in $\mathcal{G}_P^{t+1}$; the other road nodes are left free to be selected. Finally Eq. [2.21c] keeps the feeder nodes (which are connected to substation node through HV feeders) similar in $\mathcal{G}_P^t$ and $\mathcal{G}_P^{t+1}$. Therefore, the

number of binary variables reduces from $(|\mathcal{E}_R| + 2|\mathcal{V}_R|)$ to $(|\mathcal{E}_R \setminus \mathcal{E}_t| + |\mathcal{V}_R \setminus \mathcal{V}_t|)$, which is a significant improvement. Here $|\cdot|$ denotes the cardinality of a set. Algorithm 4 lists the steps involved in constructing the ensemble of N networks. We start with the near-optimal primary network and create the variant primary networks. Finally, we augment the variant primary networks to the optimal secondary network to create the networks in the ensemble.

The schematic of the process is shown in Fig. 2.4 where each transition involves deletion of a random edge followed by solving the restricted version of the optimization problem. For example, in case of the first transition, the edge $(4, 10)$ is chosen at random to be deleted. The restricted version of the optimization problem forces all the edges in the optimal primary network except $(4, 10)$ to be selected. The remaining edges are selected from the candidate set of edges (shown in dotted green lines) so that the output result is a feasible network.

2.4.4 Step 4: post-processing of networks

Step 4a. adding node and edge attributes to networks

The final step of the synthetic network creation process is to assign *line types* to the different distribution lines in the network. We use the catalog of distribution lines (Mateo, Postigo, F. D. Cuadra, et al. 2020) to assign the line types and parameters to secondary and primary distribution lines. A simultaneity factor of 0.8 has been used to rate the transformers and distribution lines similar to the steps used in Domingo et al. (2011). This means that when all residences are consuming their respective peak hourly load, the distribution lines are loaded to at most 80% of their rating. Table 2.7 provides a list of lines used in the synthetic secondary and primary distribution networks. A summary of the percentage of these lines in the entire network of Montgomery County is also appended in the table. Table 2.8 lists

the node and edge attributes in the generated synthetic power distribution networks. We have uploaded our data to the GitHub repository (Meyur 2021).

Table 2.7: Catalog of LV and MV distribution network lines

Line Name	**Line Type**	**Resistance (ohms/1000 ft)**	**Reactance (ohms/1000 ft)**	**Current (A)**	**Voltage (kV)**	**% of lines**
OH Voluta	TRPLX #6	0.661	0.033	95	0.24	73.40
OH Periwinkle	TRPLX #4	0.416	0.031	125	0.24	8.70
OH Conch	TRPLX #2	0.261	0.030	165	0.24	7.00
OH Neritina	TRPLX 1/0	0.164	0.030	220	0.24	5.00
OH Runcina	TRPLX 2/0	0.130	0.029	265	0.24	2.20
OH Zuzara	TRPLX 4/0	0.082	0.027	350	0.24	3.70
OH Swanate	ACSR #4	0.407	0.113	140	12.47	90.80
OH Sparrow	ACSR #2	0.259	0.110	185	12.47	3.90
OH Raven	ACSR 1/0	0.163	0.104	240	12.47	2.70
OH Pigeon	ACSR 3/0	0.103	0.099	315	12.47	1.70
OH Penguin	ACSR 4/0	0.082	0.096	365	12.47	0.80

A key aspect of creating the synthetic networks is regulating the voltage at different nodes in the network. ANSI C.84 Range A limits the acceptable node voltage within $0.95 - 1.05$ p.u.; however, in certain sections, we observe undervoltage. Since such occurrences of undervoltage are common in real distribution networks as well, we relax the acceptable limits as follows: accept networks where voltage at $> 99.5\%$ of residences are within the limits set by ANSI C.84 Range A ($0.95 - 1.05$ p.u.). In the rejected cases, we perform either of the following post-processing steps: (i) increase the load tap changer (LTC) setting at the distribution setting to 1.02 p.u. (ii) add line regulators at the end of HV feeders, which are used to connect remote loads.

Step 4b. constructing three phase networks

We also created three phase synthetic distribution networks as a post-processing step. The synthetic networks created thus far are positive sequence networks, which are suitable for

Table 2.8: Node and edge attributes in created synthetic power distribution networks

Attribute Name	**Attribute Details**	**Size of Output Network for Montgomery County, Virginia (US)**
Node ID	integer ID of node	63220 nodes
Node geometry	(x, y): longitude and latitude of node	20 substation nodes
Node label	'S': substation, 'T': local transformer, 'H': residence, 'R': auxiliary node	35629 residence nodes 8812 auxiliary nodes 18759 transformer nodes
Node average load	Average hourly load demand in Watts	
Node peak load	Peak hourly demand in Watts	
Node phase	Phase assigned to node (A, B or C)	
Edge ID	integer IDs of connecting nodes	
Edge geometry	shapely LineString geometry of edge	
Edge label	'E': HV feeder line, 'P': MV primary network, 'S': LV secondary network	63200 edges 156 HV feeder lines
Edge name	Name of conductor type used for the edge	27415 primary edges
Length	Length of the edge in meters	35629 secondary edges
Resistance	Resistance of the edge conductor in p.u.	
Reactance	Reactance of the edge conductor in p.u.	
Edge phase	Phase assigned to edge (A, B, C or mixed)	

planning studies. We create networks comprising of three phases (A, B, and C) which would be useful to run studies on the operation of the network involving phase unbalances. We aim to create a balanced three phase network where the power delivered at each substation feeder is balanced in the three phases. For the sake of simplicity, we consider branching of three different phases only in secondary networks. This implies that each residence is assigned one of the three phases, and the branching occurs at the local pole top transformers. Each branch originating from a transformer feeds residences with the same phase. Thus, we have a problem of three way partitioning the set of residences, such that the total load supplied to residences in each partition is balanced.

Problem 2.5 (Three phase network creation problem)**.** *Given a set of residences $\mathcal{H}$ connected to a substation feeder with respective load demand p_h for every $h \in \mathcal{H}$, find a partition into three sets $\mathcal{H}_A, \mathcal{H}_B, \mathcal{H}_C$ such that*

$$\sum_{h \in \mathcal{H}_A} p_h = \sum_{h \in \mathcal{H}_B} p_h = \sum_{h \in \mathcal{H}_C} p_h$$

.

We define an optimization problem similar to the one proposed in Saha et al. (2019), which assigns a phase to every residence connected to a substation feeder and minimizes the deviation between the total load fed by each pair of phases. We provide the details of the optimization framework, which we solve to assign one of the three phases to each residence connected to a substation feeder. We assign binary variables $u_h^A, u_h^B, u_h^C \in \{0,1\}$ to every residence $h \in \mathcal{H}$ in order to denote the phase assigned to the residence. For example, $u_h^A = 1$ denotes that residence h is assigned phase A. We stack the binary variables to construct vectors $\mathbf{u}^A, \mathbf{u}^B, \mathbf{u}^C \in \{0,1\}^{|\mathcal{H}|}$ respectively. Since we assign exactly one of the

phases (A,B,C) to each feeder, we enforce the following constraint:

$$\mathbf{u}^A + \mathbf{u}^B + \mathbf{u}^C = \mathbf{1} \tag{2.22}$$

Let p_n denote the average hourly power demand at residence n. We stack the power demand to obtain the vector $\mathbf{p} \in \mathbb{R}^{|\mathcal{H}|}$. Hence, we can compute the total power fed by three phases at the substation feeder as follows:

$$p^A = \mathbf{p}^T \mathbf{u}^A \tag{2.23a}$$

$$p^B = \mathbf{p}^T \mathbf{u}^B \tag{2.23b}$$

$$p^C = \mathbf{p}^T \mathbf{u}^C \tag{2.23c}$$

Note that the superscript 'T' denotes the transpose of a vector. The phases are assigned to each residence by solving the optimization problem in Eq. [2.24].

$$\begin{aligned} &\min_{\mathbf{u}^A, \mathbf{u}^B, \mathbf{u}^C} |p^A - p^B| + |p^B - p^C| + |p^C - p^A| \\ &\text{s.to. Eqs. [2.22], [2.23]} \end{aligned} \tag{2.24}$$

2.5 Implementation examples

2.5.1 Mapping residences to nearest road network link.

The steps involved in mapping a given residence to the nearest road network link are shown in Fig. 2.5. The first figure shows the residence (magenta), which is required to be mapped to the nearest road link. We draw bounding regions around each road link as well as the residence. Thereafter, we identify the bounding regions which intersect with the bounding

region of the residence. The nearest road link is identified from these short-listed links.

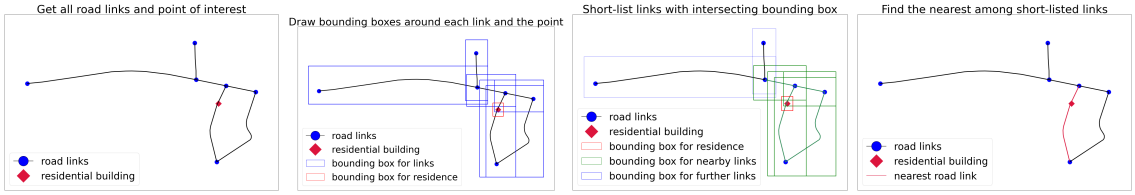

Figure 2.5: Plots showing the steps in mapping a residence to the nearest road network link. Bounding regions are drawn around each link (blue boxes) and the residence (red box). The intersections (green boxes) between these bounding regions are identified to shortlist nearby road links. The nearest road network link (red edge) is then selected from these shortlisted links.

2.5.2 Connecting residences to local transformers

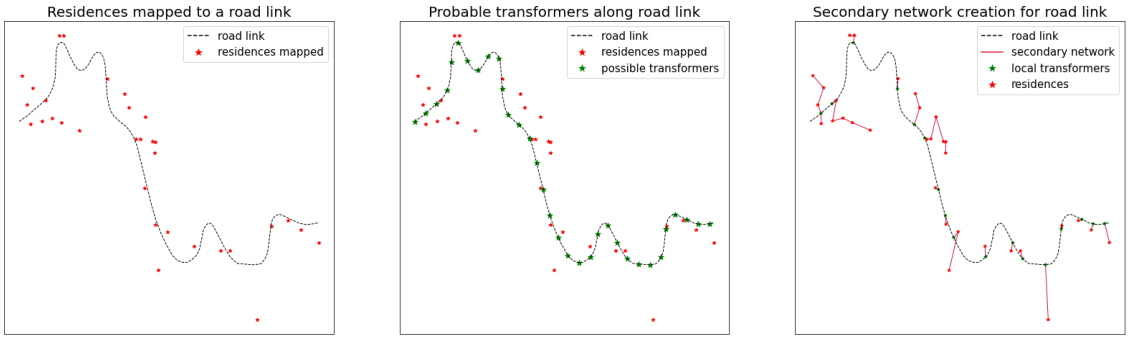

Figure 2.6: Plots showing steps in creating the secondary network for a road network link. First the probable transformer locations are identified along the link. Thereafter, the network is generated by solving the optimization problem Eq. [2.8]. The network originates from the road link and connects residences mapped to it in a forest of star-like trees rooted at the transformers.

Here we present an example of the secondary distribution network created for the state of Virginia. The steps involved in constructing the synthetic secondary network connecting residences along a road network link are shown in Fig. 2.6. The first figure shows the residences (red) mapped to the road link and the second figure shows the probable transformers (green points) along the link. A Delaunay triangulation is considered to connect the points

and obtain the set of possible edges. Thereafter, the MILP Eq. [2.8] is solved which identifies the optimal set of edges as shown in the third figure. We observe that all residences are connected in star-like trees with roots as the transformers along the road link.

2.5.3 Mapping local transformers to substations

The secondary network creation step generates local transformer nodes all over the footprint of the region under consideration, which in our case is the Montgomery County of southwest Virginia. Next, we perform the mapping of the transformer nodes to the nearest substation along the road network. We show the partitions in Fig. 2.7 where each color denotes a partition of the transformer nodes.

2.5.4 Connecting local transformers to substation

We show the steps involved in creating the primary distribution network within the boundary of Montgomery County of Virginia, USA in Fig. 2.8. We use the road network obtained from Open Street Maps as a proxy for the primary network, i.e., the edges in the primary network are chosen from the road network edges by solving the optimization problem in Eq. [2.20]. The first figure shows the road network which is used to construct the primary network. The zoomed-in inset figures show two examples of the road network subgraphs assigned to two substations. Note that this assignment is computed from the inverse mapping $\mathcal{F}_V^{-1}$ in the preceding step. The second figure shows the solution of the optimization problem where the primary network is created from the set of edges in the road network. The primary network is a tree originating from a substation and connects the local transformers through multiple feeders. Note that the road network in the first figure had multiple cycles, which are no longer present in the primary network in the second figure. Finally,

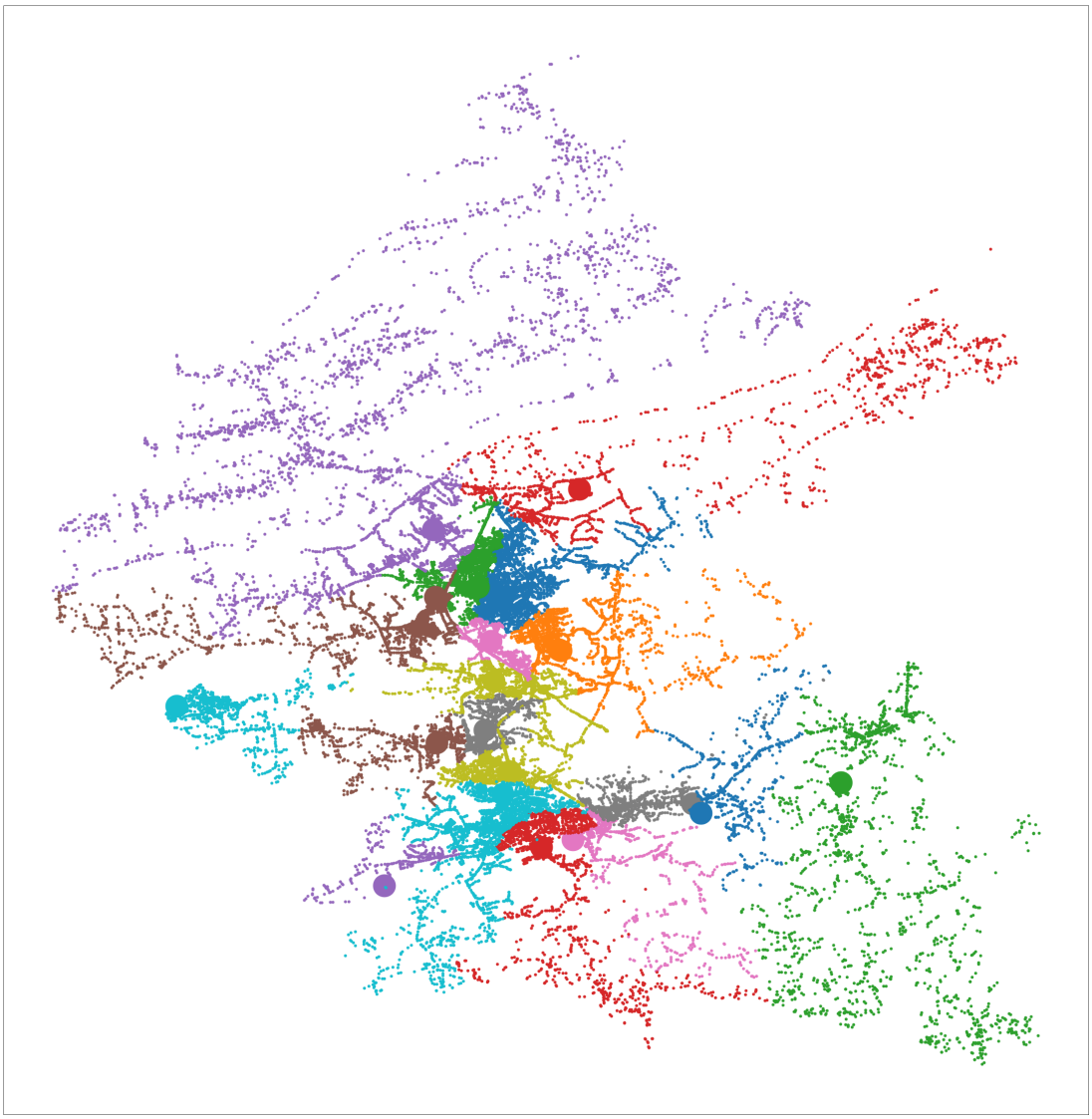

Figure 2.7: Plot showing the mapping between local transformer nodes and substations in Montgomery County of southwest Virginia. Each color depicts a particular set of transformer nodes mapped to a substation. Note that same color has been used to denote more than one mapping. The primary network creation problem (in the succeeding step) would generate a primary network connecting the substation to the mapped transformer nodes.

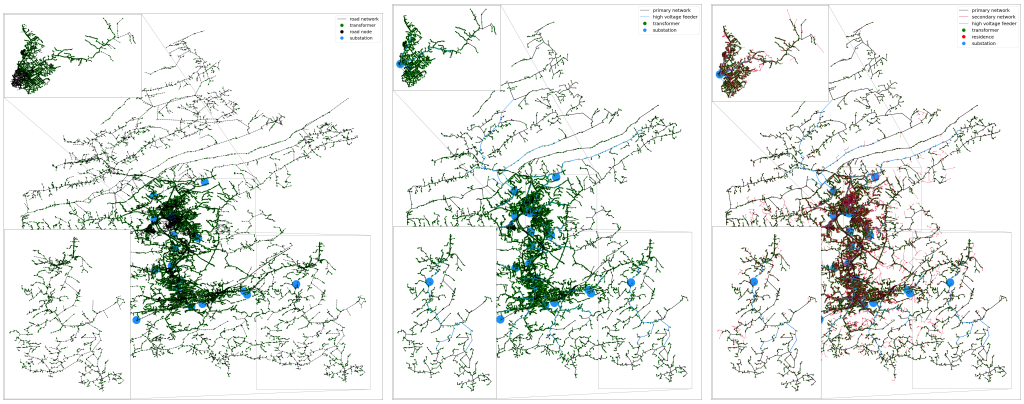

Figure 2.8: Plots showing the steps involved in creating the primary network of Montgomery County in southwest Virginia. The edges of the primary network (middle figure) are chosen from the road network (left figure) by solving the optimization problem in Eq. [2.20]. The network originates from the substations and follows the road links to connect the local transformers in a tree structure with no loops. The secondary network is appended to the local transformers to obtain the final synthetic distribution network (right figure).

the secondary network is appended to the local transformers to obtain the final distribution network, which is shown in the third figure.

2.5.5 Power flow studies in created networks

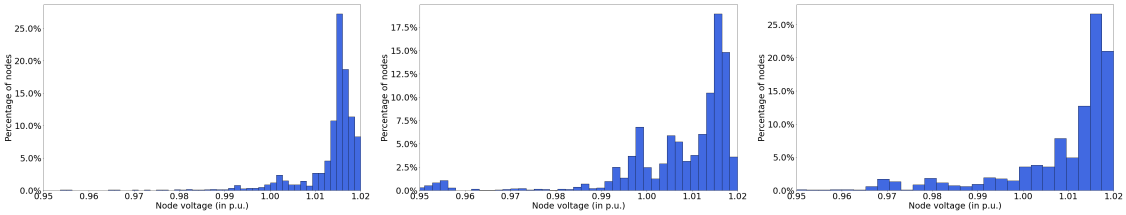

Figure 2.9: Plots showing histogram of node voltages for three synthetic distribution networks operating at peak load. We observe that majority of the nodes are within acceptable voltage limits set by the ANSI C.84 A standard. The substation voltage is set at 1.02 p.u. to avoid low voltages at the leaf residence nodes.

The three figures in Fig. 2.9 show the histogram of node voltages after performing these post-processing steps. The three plots in Fig. 2.10 show the variation of node voltage with

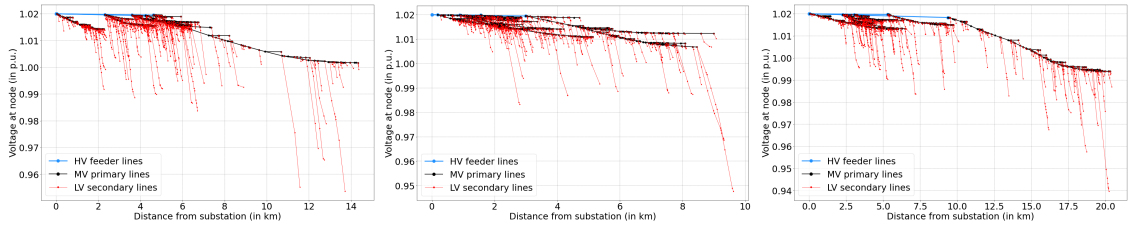

Figure 2.10: Plots showing variation of voltage with distance from substation (bottom three figures) for three synthetic distribution networks operating at peak load. The HV feeders are used to connect distant residences while maintaining a healthy voltage profile. The primary network maintains the voltage level within acceptable engineering standard for distribution; while significant voltage drop occurs at in the LV secondary network.

distance from the substation. The blue lines show the HV feeder lines, which results in the minimum voltage drop. The black lines denote the MV primary network lines, which causes voltage drop as we tend to move away from the substation. Finally, the red lines are the LV secondary lines resulting in a major voltage drop at the residences. However, in all cases, we notice that $> 99.5\%$ of customer node voltages are within the acceptable ANSI C.84 A limits.

2.5.6 Creating three phase networks

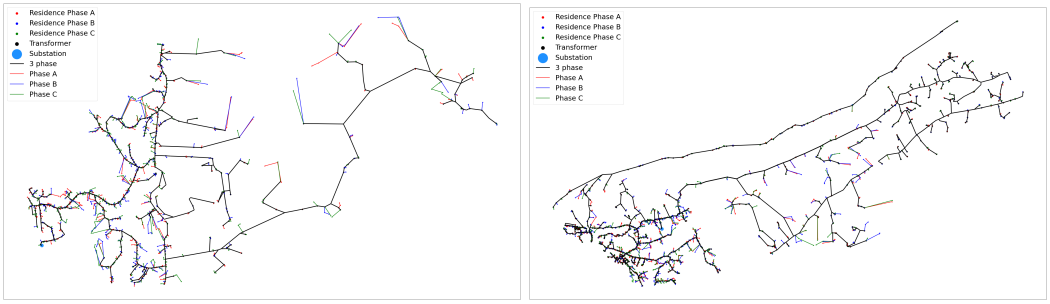

Figure 2.11: Plots showing two 3-phase networks generated from the optimal positive sequence networks.

Fig. 2.11 shows the three phase network created from an existing positive sequence network after solving the optimization problem in 2.24. Though we have assigned one of the

three phases to each residence in the network, we do not consider three phase circuits with different transformer configurations (wye and delta). Therefore, we limit the created synthetic networks with only positive sequence impedance. To this end, such networks can be useful to perform studies involving balanced loads across three phases.

2.5.7 Degree, hop and reach distribution of created networks

In this section, we compare the statistical attributes of the synthetic distribution networks created for rural and urban areas. The degree of a node in a network denotes the number of edges connected to it. The degree distribution gives an idea about the connectivity within the network. The 'hop' of a node from substation (root) node is defined as the number of edges lying between them. Hence, the "hop distribution" provides an idea about the radial layout of nodes around the root substation node. Finally, we define "reach" of a node as the length of network (in miles) connecting it to the substation. The associated "reach distribution" of a network becomes a relevant statistic in the context of networks with associated geographic attributes since it provides a distance metric to the hop distribution.

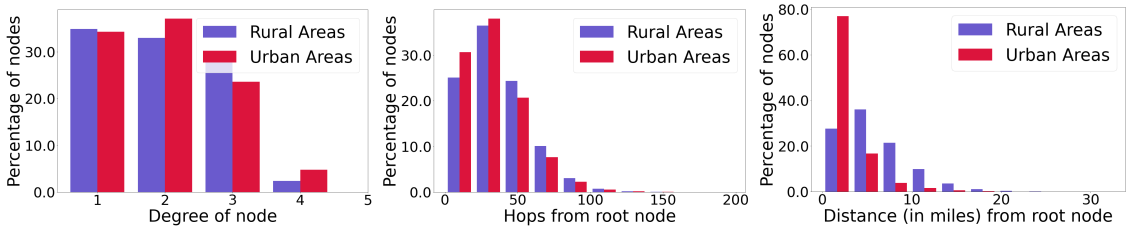

Figure 2.12: Plots showing degree distribution (left), hop distribution (middle), and reach distribution (right) in rural and urban areas. Colors depict network attributes of urban versus rural areas. The degree and hop distribution are similar for both rural and urban regions. The reach distribution of urban networks peak at small value since the distribution network nodes are more closely placed to the substation than rural areas.

Fig. 2.12 shows comparison of degree, hop and reach distributions in urban and rural distribution networks. We observe that the degree and hop distributions are fairly similar.

However, the reach distribution differs for rural and urban areas. In case of urban areas we notice that a majority of nodes are located very close to the substation whereas rural areas are often characterized by long length network edges. This observation is also consistent with the distribution of residences in urban and rural regions where rural regions have more widely spread out residences than urban areas.

2.5.8 Motifs in synthetic networks

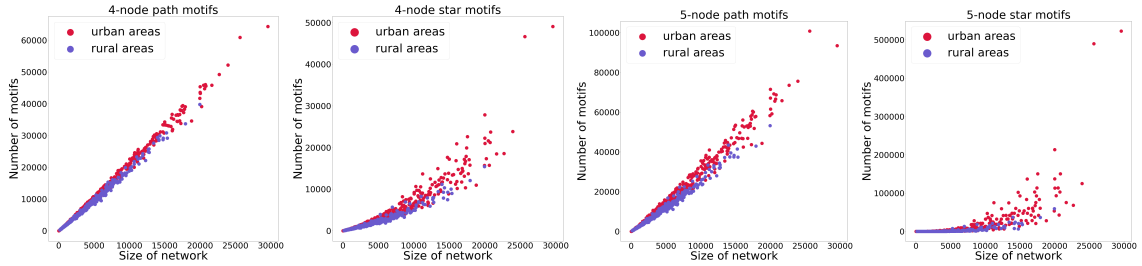

Figure 2.13: Plots showing number of 4-node paths (top left), 4-node star motifs (top right), 5-node paths (bottom left), and 5-node star motifs (bottom right) as a function of network size (measured as number of nodes in the network). Colors depict motif numbers in urban versus rural areas. Urban distribution networks have a larger number of star motifs than rural networks. In contrast, the path motif count does not differ significantly across rural and urban areas. Urban networks are often larger than rural networks as measured by number of nodes due to larger population size.

Network motifs are interesting subgraphs which build up the entire network. Network motifs have been used as a metric to understand network resilience in earlier work (Dey, Gel, and Poor 2019). We focus our attention to small size subgraphs with at most 4 nodes. Since the distribution networks are tree graphs, the relevant subgraphs are paths and stars. We define a k-path motif as a subgraph of k nodes that form a path. A k-star motif is a subgraph of k nodes which form a star, i.e., it consists of a single central node with degree $k-1$, and the remaining $k-1$ nodes are connected to the central node. Here, we are interested in four types of network motifs: (i) 4-node path, (ii) 4-node star, (iii) 5-node path, and (ii) 5-node star. Fig. 2.13 shows the number of 4-node and 5-node motifs in

the synthetic distribution networks. The two colors show the results for urban and rural networks separately. The star motifs are higher for urban networks as compared to rural networks of similar size. This can be explained from the observation in degree distribution where we notice that urban networks have higher fraction of nodes with degree 4. A single node with degree 4 results in $\binom{4}{3} = 4$ counts of $4-$node star motifs.

2.5.9 Features in ensemble of networks

We create ensembles of distribution networks for Montgomery county in southwest Virginia. The entire network within Montgomery county is composed of 19 sub-networks (each fed by a different substation). We create an ensemble of 20 networks for each sub-network and study the variation in network attributes over the ensembles. We plot the variation in degree, hop and reach distributions in Fig. 2.14. The error bar shows the extent of variation in the ensemble. Fig. 2.15 shows variation in 4-node path and star motif counts for the networks in each ensemble. The bar plots show the motif counts for each ensemble of networks and the error bars (on top of each bar) depict the variation over the ensemble. We observe that the variation of network features over each ensemble is not significant. This shows that the networks are fairly close to each other and each of them can be considered as a digital twin of the actual network. Thus our framework is capable of creating an ensemble of synthetic distribution networks which are statistically equivalent to each other. In order to create statistically different networks, the Markov chain in step 3 needs to be altered - deleting multiple random edges, instead of one.

In general, an 'ensemble' of networks consist of multiple structurally different networks which connect the same set of residences to the substation. Each synthetic network in the ensemble is a feasible network (has a tree structure and satisfies power engineering

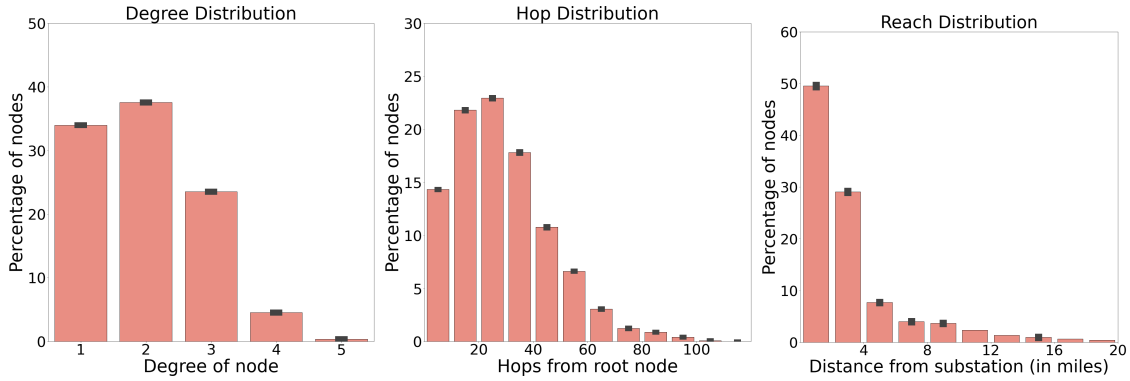

Figure 2.14: Plots showing variation in degree distribution (left), hop distribution (middle), and reach distribution (right) for the ensemble of distribution networks created for Montgomery county of southwest Virginia. The error-bars in the bar plots show the variation over the networks in the ensemble.

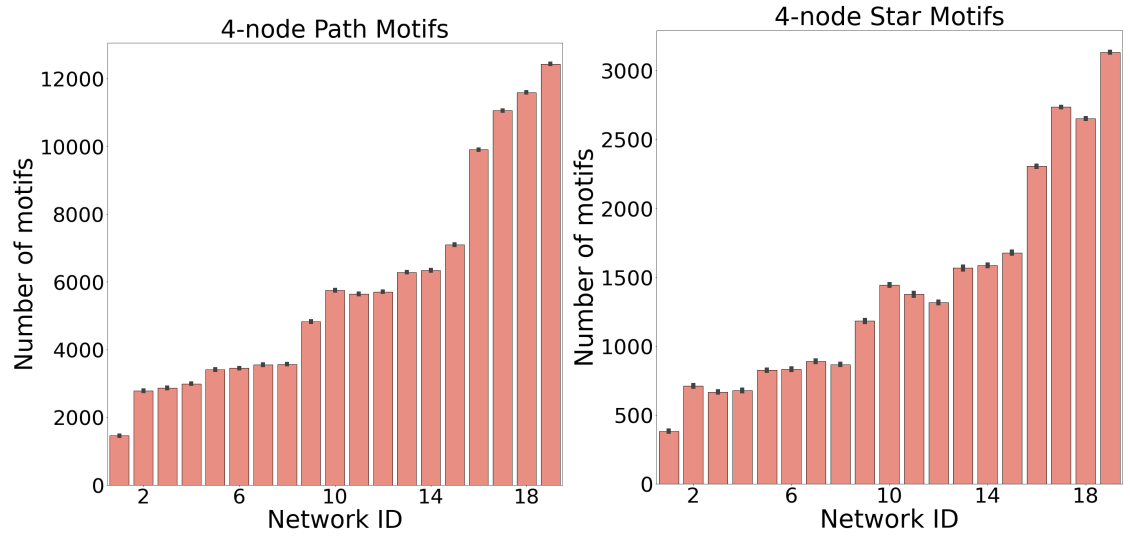

Figure 2.15: Plots showing variation in number of 4-node path motifs (left), and number of 4-node star motifs (right) for the ensembles of distribution networks created for Montgomery county of southwest Virginia. Results are shown for 19 ensembles of varying size in the county fed by different substations. Each ensemble consists of 20 networks. The error-bars in the bar plots show the variation over the networks in each ensemble.

constraints) but is not the optimal length network. Therefore, we can consider it as a single random realization of the actual network. This allows us to perform analysis on an ensemble of networks instead of a single network and thereby capture the deviation arising due to different network structure in the ensemble.

2.6 Concluding remarks

Although the synthetic power distribution network dataset produced by our framework is comprehensive, it is not without its limitations. In this work we generate networks with only positive sequence parameters. The ensemble of synthetic networks can be used as a tool for performing planning studies or addressing system-wide policy level questions. We can also perform short circuit analysis with symmetrical three phase faults.

However, distribution systems are networks of mixed phase order and mixed network configuration. They are usually three phases in the primary network and the secondary network consists of mixed single and three phase circuits. We have provided a framework in the SI to partition the residences into three phases and thereby create a three phase network. A complete three phase network requires inclusion of zero sequence line parameters and transformer configurations (wye-wye, delta-wye, wye-delta and delta-delta). Therefore, in their current version, these networks might not be suitable to be used for performing dynamic stability analysis or studying detailed transient responses to power grid contingencies.

Further, shunt compensation is used in the primaries for maintaining voltage level within engineering standards. These comprise of capacitor banks which elevate voltage level along the network. Hence, they can be optimally placed in the network to avoid severe undervoltage issues at remotely located residences. We can consider critical sections of the network

and design necessary shunt compensation to maintain a high degree of reliability of the network. Additionally, the proposed framework creates a network to connect only the residential buildings in a geographic region. In order to connect heavy load centers, networked secondaries with pad-mounted transformers are used in some large urban areas. These additions can be made to our existing synthetic networks and would be a direction for future research.

Chapter 3

Validating Synthetic Power Distribution Networks

Publications

1. Rounak Meyur, Anil Vullikanti, Samarth Swarup, Henning S. Mortveit, Virgilio Centeno, Arun Phadke, H. Vincent Poor, and Madhav Marathe, "Ensembles of realistic power distribution networks" in Proceedings of the National Academy of Sciences, Vol. 119 No. 42, Oct 2022.

2. Rounak Meyur, Madhav Marathe, Anil Vullikanti, Samarth Swarup, Henning S. Mortveit, Virgilio Centeno, and Arun Phadke, "Creating realistic power distribution networks using interdependent road infrastructure", in IEEE International Conference on Big Data, Dec 2020 (pp. 1226–1235).

3. Rounak Meyur, Lyman Kostiantyn, Bala Krishnamoorthy, and Mahantesh Halappanavar, "Structural validation of synthetic power distribution networks using the multiscale flat norm", in 14th ACM/SPEC Conference on Performance Engineering (ICPE 2023), Apr 2023, Coimbra, Portugal (*submitted and under review*).

3.1 Introduction

In the previous chapter, we discussed about the framework to generate synthetic power distribution network for a given geographic region. The aim is to create networks which resemble their actual physical counterparts. A critical question pertaining to creating synthetic power distribution networks is: how similar are the synthesized networks to the actual power distribution networks? Therefore, the generated synthetic networks require detailed validation before they can be used as a substitute for actual networks for various applications. To this end, the networks need to be compared against actual distribution networks in terms of their structural properties as well as their power engineering attributes (Krishnan et al. 2020). In this chapter, we compare the created synthetic networks for the town of Blacksburg (in Montgomery county of southwest Virginia, USA) against actual networks obtained from a power company operating in the same region. In addition to comparing standard graph attributes such as degree and hop distributions, we compute the difference in geometries between the actual and synthetic networks and provide a measure of deviation.

We obtained real-world power distribution networks for the town of Blacksburg in southwest Virginia from a distribution company to validate the created networks. This network has been incrementally built over a long period of time with a close dependency on the population growth in the region. In contrast, our proposed framework uses the current population information with no consideration of any historical data. The created synthetic networks are optimal in terms of economic and engineering perspectives. Therefore, it is expected that there would be structural differences between the networks. Furthermore, the comparison methods need to be relevant in the context of distribution networks with associated geographic attributes.

Due to its proprietary nature, the node and edge labels were redacted from the actual net-

work data. Further, the networks were shared as a set of handmade drawings, many of which had not been drawn to a well defined scale. We have subsequently digitized these drawings by overlaying them on OpenStreetMaps (*Open Street Maps* 2021) and georeferencing them to particular points of interest (ESRI 2022). Thereafter, the geometries corresponding to the actual network edges are obtained as shape files.

This chapter compares the generated synthetic networks with the actual distribution network based on various operational, statistical and structural attributes. The operational validation ensures that we observe similar node voltages and edge flows in both networks. This makes the networks suitable to be used by the scientific community to aid in their research. The methods to compare statistical attributes help us compare the overall connectivity properties of the networks. The comparison of structural attributes involving node and edge geometries enable us to validate the created synthetic networks on a much higher resolution. The results of the comparison show that the created networks bear a significant amount of resemblance to the actual networks.

We particularly focus on the structural validation aspect. The literature pertaining to frameworks for synthetic distribution network creation include certain validation results that compare the generated networks to the actual counterpart (Krishnan et al. 2020; Schweitzer et al. 2017; Bidel, Schelo, and Hamacher 2021). But the validation results are mostly limited to comparing the statistical network attributes such as degree and hop distributions and power engineering operational attributes such as node voltages and edge power flows. Since power distribution networks represent real physical systems, the created digital replicates have associated geographic embedding. Therefore, a structural comparison of synthetic network graphs to their actual counterpart becomes pertinent for power distribution networks with geographic embedding.

3.1.1 Related works

Several well defined graph structure comparison metrics such as subgraph isomorphism and edit distance have been proposed in the literature along with algorithms to compute them efficiently. Tantardini et al. (2019) compare graph network structures for the entire graph (global comparison) as well as for small portions of the graph known as motifs (local comparison). Other researchers have proposed methodologies to identify structural similarities in embedded graphs (Bai et al. 2018; Ok 2020). However, all these methods depend on one-to-one correspondence of graph nodes and edges rather than considering the node and edge geometries of the graphs. The edit distance, i.e., the minimum number of edit operations to transform one network to the other, has been widely used to compare networks having structural properties (Xu 2015; Paaβen 2022; Riba et al. 2021). Riba et al. (2021) used the Hausdorff distance between nodes in the network to compare network geometries. Majhi and Wenk (2022) modified the traditional definition of graph edit distance to be applicable in the context of "geometric graphs" or graphs embedded in a Euclidean space. Along with the usual insertion and deletion operations, the authors have proposed a cost for translation in computing the geometric edit distance between the graphs. However, the authors also show that the problem of computing this metric is $\mathscr{NP}$-hardness.

In this chapter, we attempt to compare the network geometries using Hausdorff distance after partitioning the geographic region into small rectangular grids followed by comparing the geometries for each grid. However, the Hausdorff distance metric is sensitive to outliers as it focuses on the maximum possible distance between a pair of geometries. For a pair of geometries that coincides almost entirely except for a few small portions, the Hausdorff distance metric records the discrepancy as the deviation between them, without accounting for the similarity over the majority of portions. A similar approach was used by Brovelli et al. (2017) to compare a pair of road networks in a given geographic region, and suffers from

the same drawback. This necessitates a well-defined distance metric between networks with geographic embedding (Ahmed, Fasy, and Wenk 2014).

Several comparison methods have been proposed in the context of planar graphs embedded on an Euclidean space (Cardillo et al. 2006; Morer et al. 2020). They include local and global metrics to compare road networks. The local metrics characterize the networks based on cliques and motifs, while the global metrics involve computing the *efficiency* of constructing the infrastructure network. The most efficient network is assumed to be the one with only straight line geometries connecting node pairs. Albeit useful to characterize network structures, these methods are not suitable for comparing network geometries.

3.1.2 Contributions

We propose a new distance measure to compare a pair of geometries using the *flat norm*, a notion of distance between generalized objects studied in geometric measure theory (Federer 1969; F. Morgan 2008). This distance combines the difference in length of the geometries with the area of the patches contained between them. The area of patches in between the pair of geometries accounts for the lateral displacement between them. We employ a *multiscale* version of the flat norm S. P. Morgan and Vixie 2007 that uses a scale parameter $\lambda \geq 0$ to combine the length and area components (for the sake of brevity, we refer to the multiscale flat norm simply as the flat norm). Intuitively, a smaller value of λ captures larger patches of area between the geometries while a large value of λ captures more of the (differences in) lengths of the geometries. Computing the flat norm over a range of values of λ allows us to compare the geometries at multiple scales. For computation, we use a discretized version of the flat norm defined on simplicial complexes (Sharif, Krishnamoorthy, and Vixie 2013), which are triangulations in our case. A lack of one-to-one

correspondence between edges and nodes in the pair of networks prevents us from performing one-to-one comparison of edges. Instead we can sample random regions in the area of interest and compare the pair of geometries within each region. For performing such local comparisons, we define a *normalized flat norm* where we normalize the flat norm distance between the parts of the two geometries by the sum of the lengths of the two parts in the region. Such comparison enables us to characterize the quality of the digital duplicate for the sampled region. Further, such comparisons over a sequence of sampled regions allows us to characterize the suitability of using the entire synthetic network as a duplicate of the actual network.

Our main **contributions** are the following: (i) we propose a distance measure for comparing a pair of geometries embedded on the same plane using the flat norm that accounts for deviation in length and lateral displacement between the geometries; and (ii) we perform a region-based characterization of synthetic networks by sampling random regions and comparing the pair of geometries contained within the sampled region. The proposed distance allows us to perform a global as well as local comparison between a pair of network geometries.

3.1.3 Outline of the chapter

We outline the chapter as follows: (i) first, we perform a visual comparison between the networks and point out the differences which can be observed in plain sight; (ii) next, we perform an operational validation by comparing nodes voltages and edge flows in the networks; (iii) thereafter, we compare the statistics of network attributes; (iv) and finally, we perform a structural validation of the networks, where we discuss about our proposed metric of structural validation through the multiscale flat norm.

3.2 Visual comparison

In this section, we perform a visual comparison of the generated synthetic network to the actual network covering the same geographical region. Fig. 3.1 shows the actual distribution network of the town of Blacksburg in southwest Virginia along with the synthetic network generated for the same. At first sight, the two networks are adjacent to and almost overlap each other. The inset figure confirms this hypothesis. A zoomed-in view of the figure shows that the two networks are similar to each other. While the synthetic network retraces the road network, as expected based on our assumption, the actual network is also adjacent to it. This validates our primary assumption that the primary distribution network follows the road network.

The significant structural difference between the two networks is the substation feeder to which it is connected. Fig 3.2 depicts the structural differences between the two networks. The actual network is connected to the nearest geographically located substation, where as the synthetic network is connected to a distant feeder. It is to be noted that the synthetic distribution network is generated using first principles with the assumption that it follows the road network to the maximum extent (it is easier to place distribution poles along the road network). It is seen that the nearest substation (from which the actual network is fed) has no road links connecting itself to the neighborhood. On the contrary, the other substation, though located at a further distance, has road links connecting itself to the neighborhood under consideration. Therefore, the synthetic network generation algorithm prefers the latter substation over the former.

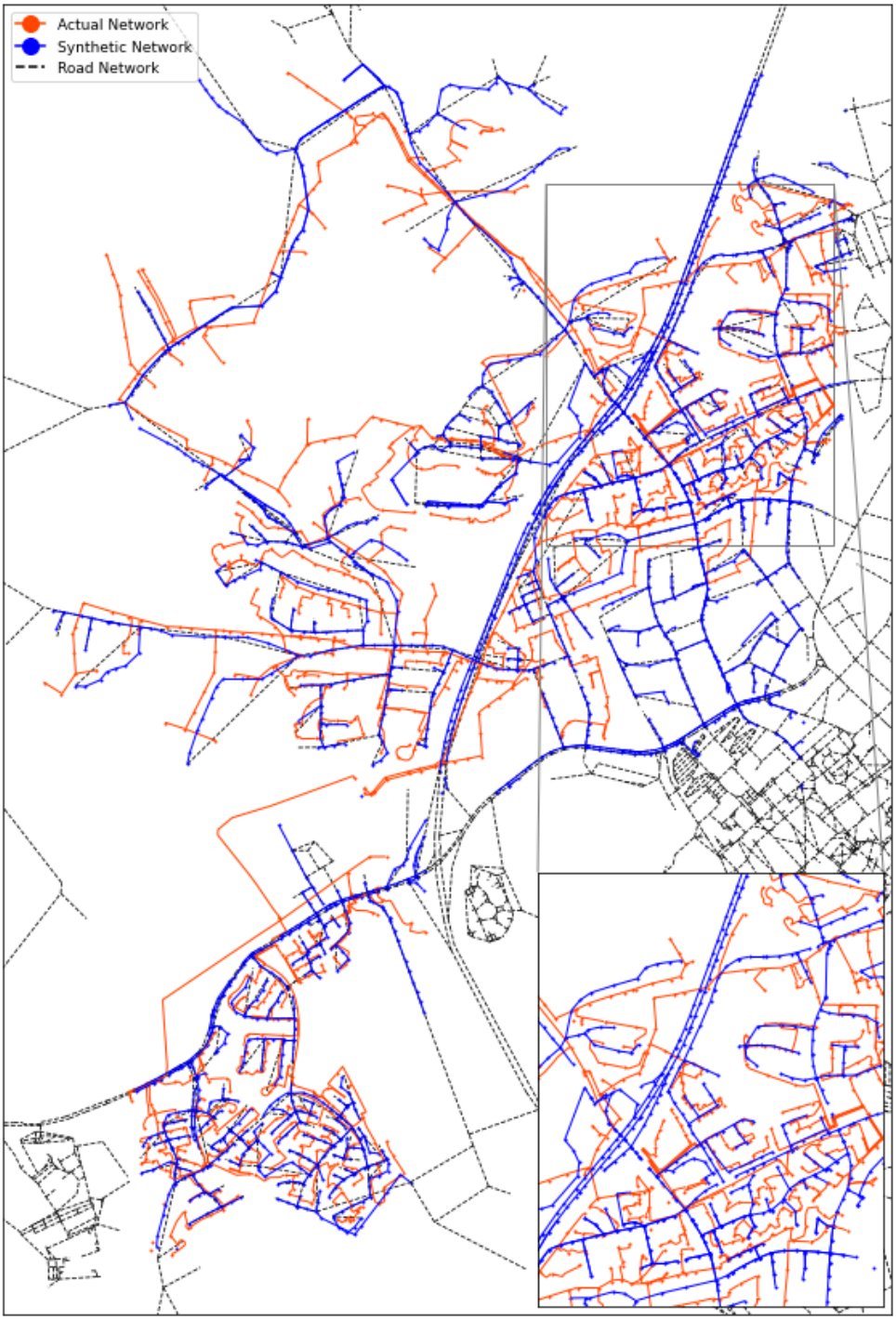

Figure 3.1: Plot comparing the structure of the actual distribution network (red) with the synthetic distribution network (blue) for the town of Blacksburg in southwest Virginia. The black dotted network denotes the underlying road network, which is used as a proxy to create the synthetic network.

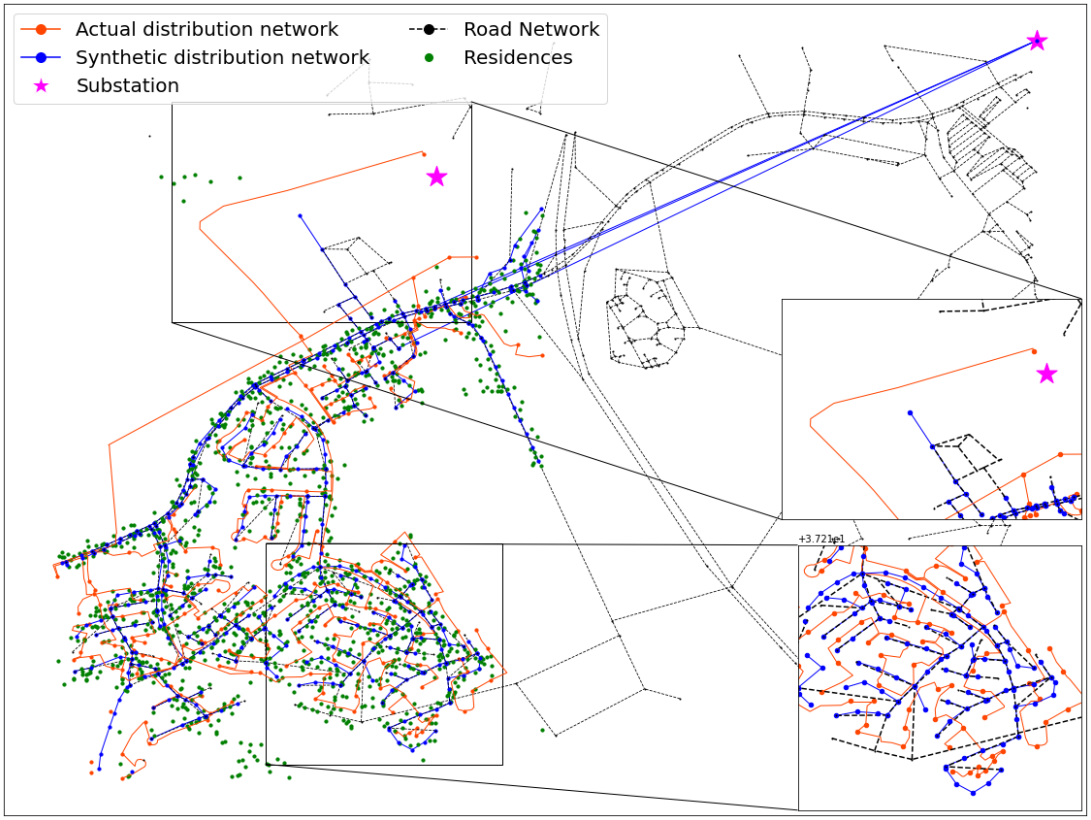

Figure 3.2: Plot explaining the visual structural difference between actual distribution network (red) and synthetic distribution network (blue) generated using first principles from the knowledge of the road network (black). A zoomed-in view of the figure shows that while the synthetic network retraces the road network, as expected based on our assumption, the actual network is also adjacent to it. The significant structural difference between the two networks is the substation feeder to which it is connected.

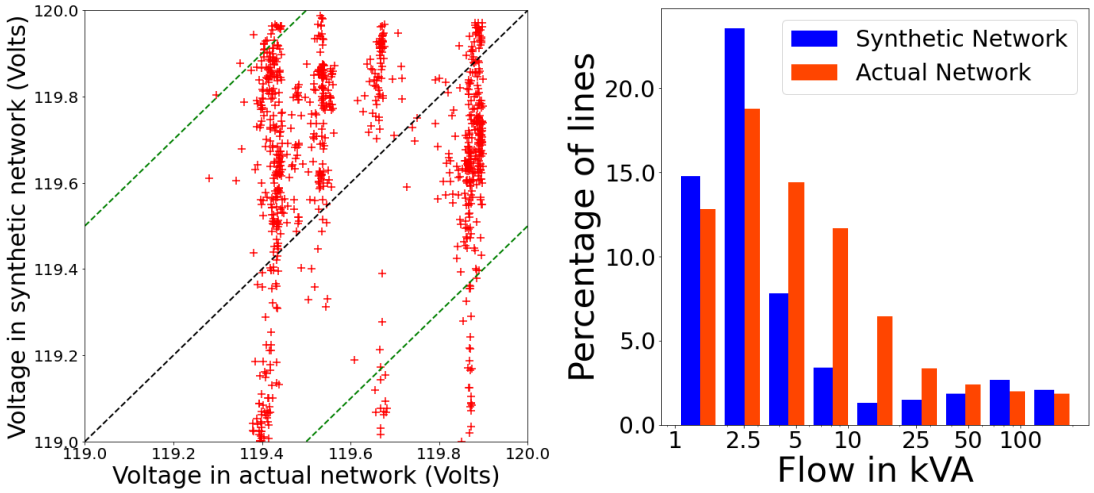

Figure 3.3: Plots comparing the residential node voltages (left) and edge power flows (right) for actual and synthetic networks. Majority of residence voltages in the synthetic network are within $\pm 0.4\%$ voltage regulation of the voltages in the actual network. The edge flows in both network follow similar distributions with a computed KL divergence of 0.15.

3.3 Operational validation

We compare voltages at the residences when they are connected to the actual and synthetic network in left plot of Fig. 3.3. We term this validation as *operational validation*, where the basic idea is that if we substitute the actual network with the synthetic network, we should see minimal voltage differences at the residences connected to either network. Here, the black dotted line denotes the identity line (exact same voltages) and green lines signify $\pm 0.4\%$ deviation from the identity line. We observe that majority of residence voltages in the synthetic network remain within this $\pm 0.4\%$ regulation. We also compare the edge flows in the two networks through the histogram in right plot of Fig. 3.3 which also bear a significant resemblance. We performed statistical fit of the flow distributions and the KL-divergence is 0.15.

3.4 Statistical validation

The created networks are expected to have similar graph attributes to the actual network. We focus on basic graph attributes such as degree and hop distributions and also the newly defined "reach" distribution. Fig. 3.4 compares the synthetic and actual network for the town of Blacksburg in southwest Virginia in terms of these statistical attributes. We use the Kullback-Leibler (KL) divergence to compare each pair of distributions. KL-divergence values for various structural measures are as follows: (i) degree distributions: 0.0208; (ii) hop distribution: 0.0323; and (iii) reach distribution: 0.0096. The small KL-divergence values indicate that the real and the synthetic networks are structurally very similar.

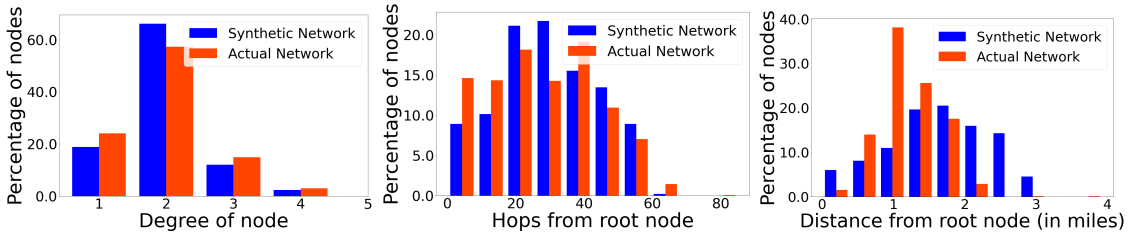

Figure 3.4: Plots comparing the degree distribution (left), hop distribution (middle), and reach distribution (right) of actual and synthetic distribution networks for town of Blacksburg in southwest Virginia. The degree and hop distributions are fairly close to each other which signifies their resemblance. The reach distribution differs between the networks because of the difference in the way each of them are created.

3.4.1 Degree distribution

The degree k_i of a node i in a graph with n nodes and internode adjacency matrix $\mathbf{A}$ with elements a_{ij} is defined by $k_i = \sum_{j=1}^{n} a_{ij}$. In general, the node degree computes the number of lines connected to it in the network. We consider three separate sections of the two networks and compare the degree distribution for each section. We observe that the two distributions match each other very closely. We observe that the degree of the nodes ranges

between 1 and 4 for either network and the degree distribution shows that majority of nodes have a degree of 2. The networks are created in a manner so that when operating with average hourly load demand, the network maintains an acceptable voltage profile. Every branch originating from a node leads to a small voltage drop caused by the load demand of the children nodes. This, in turn, causes a voltage drop in the parent nodes. Hence, for a feasible network to operate within acceptable voltage limits, it is expected that multiple branching from a single node is avoided as much as possible. This is the same observation we notice in the actual and synthetic networks, where the node degree hardly exceeds 3.

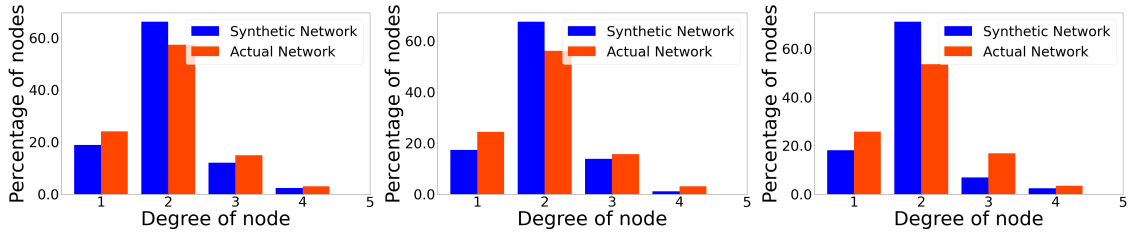

Figure 3.5: Plots comparing degree distribution of the actual and synthetic networks for three sections of region.

3.4.2 Hop distribution

A *path* $\mathcal{P}_{ij}$ in a graph between two nodes i and j is defined as a sequence of adjacent edges starting from node i and terminating at node j. A *tree* network (devoid of any cycles) has a unique path between any pair of nodes. The hops between nodes i and j in a tree network is defined as $|\mathcal{P}_{ij}|$ where $|\cdot|$ denotes the cardinality of a set. Essentially, it denotes the number of edges between the nodes i and j. Here we are interested in the number of edges between any node i and the substation node, which is the 'root' node in the synthetic networks. Therefore, we define the number of hops between any node i and substation (root) node r by $h_i = |\mathcal{P}_{ri}|$. We consider the empirical distribution for the hops h_i's of all nodes and term it as the *hop distribution* of the networks. In this analysis, we study how the nodes in each

network are distributed around the root node. We observe that the hop distributions for the synthetic and actual networks are significantly different. This is primarily because of the dissimilarity in the spatial distribution of nodes in each network which is studied later on in the later section.

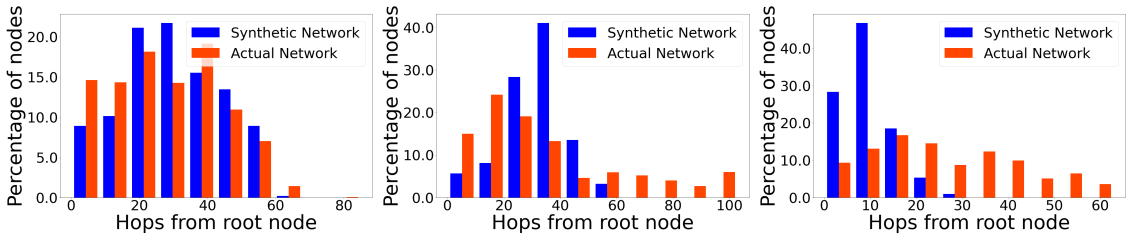

Figure 3.6: Plots comparing hop distribution of the actual and synthetic networks for three sections of the geographical region.

3.4.3 Reach distribution

Since the synthetic networks have a geographic attribute associated with them, we can use the geographic distance of each node from the substation node to study how the nodes in the network are physically present around the root node. Each edge e has an attribute w_e which denotes the length of the edge in miles. We define *reach* of a node i in the synthetic network as $l_i = \sum_{e \in \mathcal{P}_{ri}} w_e$. Essentially, the reach of a node denotes the physical distance between the node and the substation (root) node. We consider the empirical distribution for the reach l_i's of all nodes and term it as the *reach distribution* of the networks. Fig. 3.7 compares the reach distribution of synthetic and actual network for three different sections. An interesting observation is that the actual network consists of a large number of long edges. This is primarily because of the fact that the network has been built over multiple years as the population grew in the geographical location. However, the synthetic networks are generated using first principles and as an output of an optimization problem where the generated network has the minimum length (or requires minimal installation and maintenance

cost).

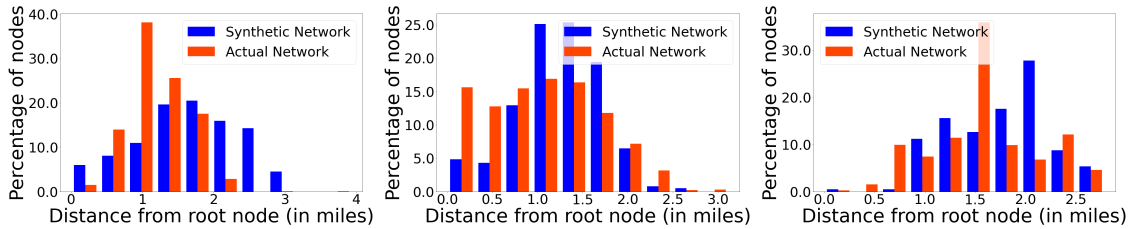

Figure 3.7: Plots comparing distribution of distance of each node to root in the actual and synthetic networks for three sections.

3.4.4 Distribution of network motifs

We also compare the 4-node and 5-node path and star motifs in actual and synthetic distribution networks. We consider three different sections in the network. Since the size of the actual and synthetic networks are very different, we compare the ratio of motif counts to the network size in Fig. 3.8. We observe that the ratios are very similar for all three sections of the networks. This validates the structural resemblance of the networks through statistical attributes.

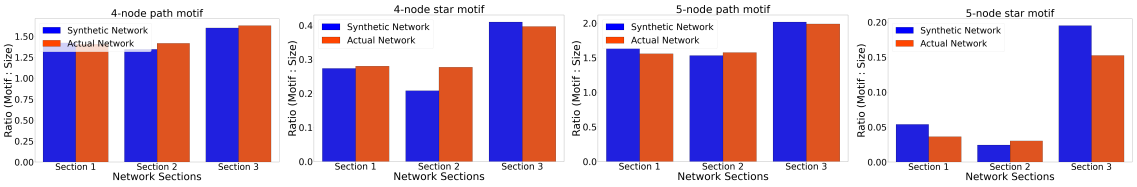

Figure 3.8: Plots comparing motif count to network size ratio of actual and synthetic distribution networks over three different sections.

3.5 Structural validation

One of the important aspects of our work is that the created synthetic networks have a geographic attribute associated with them. Therefore, we need to include network comparison methods which incorporate the geographic embedding while measuring the deviation. Here, we use a metric for geometry comparison, i.e., how the edge geometries in the networks deviate from each other. Due to unavailability of actual network information for the entire region, we propose an effective way to compare the structural attributes of the networks. We divide the entire geographic region into multiple rectangular grid cells and perform comparison in each cell separately. In this way we can omit the cells for which network data is missing.

3.5.1 Problems related to structural validation

The actual distribution network has been obtained from a power distribution company as images. The associated geographic topology is obtained by overlaying these images on maps from Open Street Maps. The distribution network in the considered region is owned by multiple power companies and we have obtained the network for only one of them. For this reason, we focus our comparison on only those portions where we have data pertaining to both networks. Further, the above-mentioned overlaying process has been performed manually and thereby has introduced some errors in assigning the geographical attributes to the actual network. To this end, the effective way to compare the structural attributes of the networks is to divide the entire geographic region into multiple rectangular grid cells and perform a comparison in each cell separately.

3.5.2 Spatial distributions of nodes

We compare the spatial distribution of nodes in the actual and synthetic networks as follows. Let the actual and synthetic networks for a region be denoted by $\mathcal{G}_{\text{act}}(\mathcal{V}_{\text{act}}, \mathcal{E}_{\text{act}})$ and $\mathcal{G}_{\text{syn}}(\mathcal{V}_{\text{syn}}, \mathcal{E}_{\text{syn}})$, and let $\mathcal{V}_{\text{act}}^{\text{CELL}} \subseteq \mathcal{V}_{\text{act}}$ and $\mathcal{V}_{\text{syn}}^{\text{CELL}} \subseteq \mathcal{V}_{\text{syn}}$ denote the set of nodes in the actual and synthetic networks lying within a rectangular grid cell. We use the following metric to compare the percentage deviation in the node distribution for grid cells where $|\mathcal{V}_{\text{act}^{\text{CELL}}}| \neq 0$:

$$\mathsf{D}_{\text{N}}^{\text{CELL}} = \frac{\frac{|\mathcal{V}_{\text{act}}^{\text{CELL}}|}{|\mathcal{V}_{\text{act}}|} - \frac{|\mathcal{V}_{\text{syn}}^{\text{CELL}}|}{|\mathcal{V}_{\text{syn}}|}}{\frac{|\mathcal{V}_{\text{act}}^{\text{CELL}}|}{|\mathcal{V}_{\text{act}}|}} \times 100$$

Fig. 3.9 shows the spatial comparison of node distribution for uniform rectangular cell partitions of two different resolutions followed by shifting the demarcations horizontally. The color code denotes the intensity of the percentage deviation in the distribution. Note that some grid cells are shaded with black dots to denote the unavailability of network data. It is easy to note that the spatial distribution of nodes is a function of the size of the rectangular grid cells as well as the location of the demarcations between cells. To address this issue, one might consider comparing networks through multiple grid sizes (varying the resolution) or by considering a single-sized rectangular grid cells followed by shifting the horizontal and vertical demarcations of the grid to account for different sized cells. The color code denotes the intensity of the percentage deviation in the distribution. Note that for some grids, the actual network data might be missing. In such cases, the comparison is not done, and the grid is shaded with black dots.

We observe that as the grid demarcations are shifted horizontally, the spatial distribution of nodes in the grids alter. Therefore, the grid resolution and grid demarcations play an important role in the structural comparison. The other alternative is to perform a quad-tree

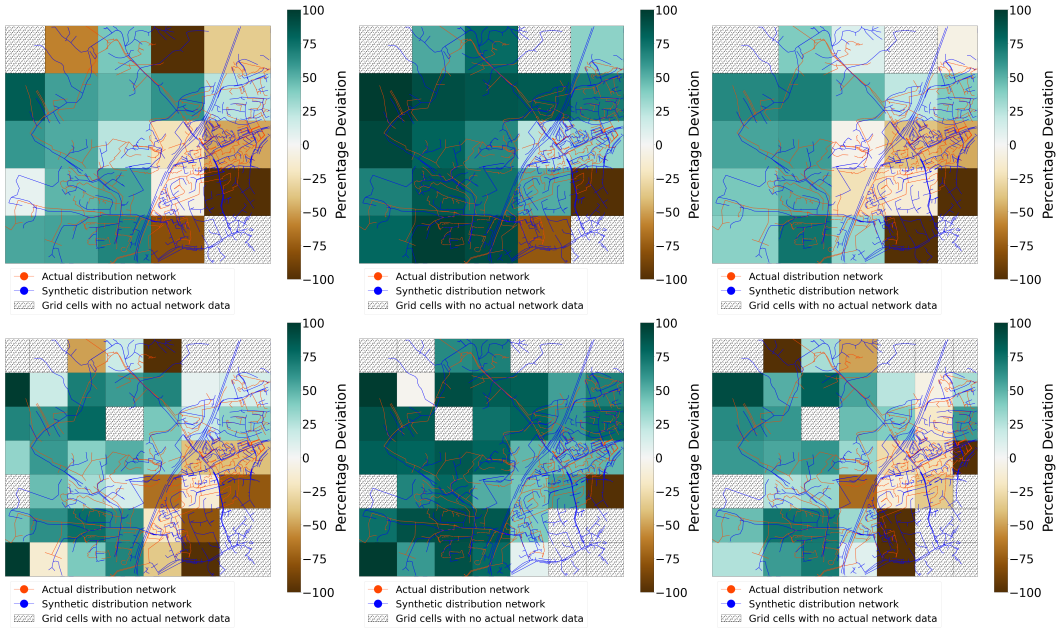

Figure 3.9: Plots comparing spatial distribution of nodes in the actual and synthetic networks. The comparison is made for rectangular grid cells where data regarding actual network is available. Two different grid resolutions are considered for the spatial distribution comparison. Also we compare the node distribution by moving the horizontal grid demarcations and observe varied results.

based partitioning of the nodes in the actual network such that each partition consists at most a particular number of nodes which is shown in Fig. 3.10. This kind of partitioning ensures that density of nodes in each grid is same throughout the network, thereby bringing some uniformity while comparing deviation in the number of nodes for each grid. However, the resolution is different for the different rectangular grids.

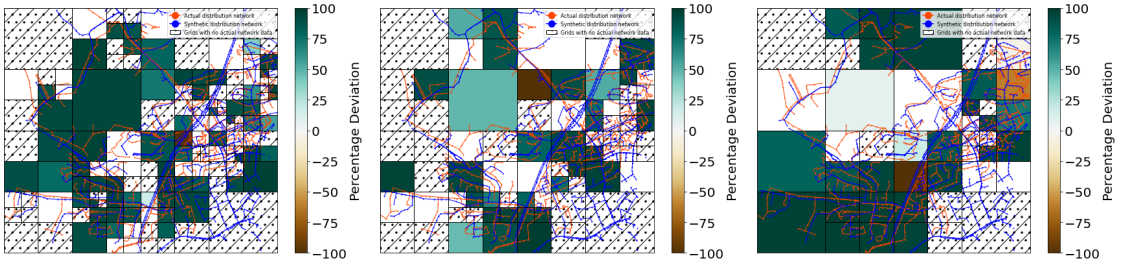

Figure 3.10: Plots comparing spatial distribution of nodes in the actual and synthetic networks. The rectangular grid cells are formed by performing a quad-tree partitioning of the nodes in the actual distribution network such that each grid has at most a given number of nodes. Color denotes the magnitude of deviation.

3.5.3 Geometry comparison

We consider a metric space $(\mathcal{M}, d)$ where $d(x,y)$ is the measure between two elements $x, y \in \mathcal{M}$. Given two non-empty subsets $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{M}$ of the metric space, the Hausdorff distance $d_{\mathrm{H}}(\mathcal{X}, \mathcal{Y})$ between the sets is defined as follows.

$$d_{\mathrm{H}}(\mathcal{X}, \mathcal{Y}) = \sup_{x \in \mathsf{X}} \inf_{y \in \mathsf{Y}} d(x,y) \tag{3.1}$$

In the context of networks associated with geographic attribute, we define the metric space $\mathcal{M}$ on which the network exists and the associated metric $d(x,y) := \mathsf{dist}(x,y)$ which is the geodesic distance between a pair of points $x, y \in \mathcal{M}$. We interpolate points along each edge

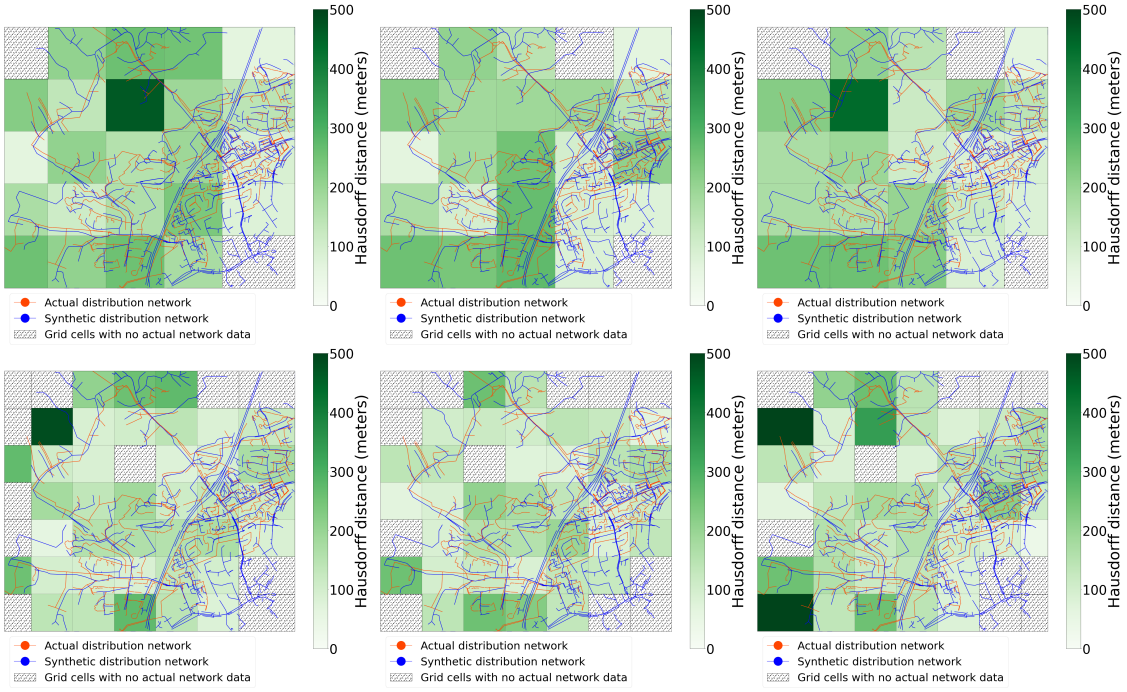

Figure 3.11: Plots showing Hausdorff distance based geometry comparison of actual and synthetic networks for the town of Blacksburg in southwest Virginia. The geometry comparison is performed for grid cells with two different resolutions: low resolution of 5×5 grid cells (top) and high resolution of 7×7 grid cells (bottom). Color in each grid cell denotes the magnitude of deviation in meters. Grid cells with no available actual network data are shaded with black dots.

of either network in order to evaluate an exact comparison of the network geometries. Let the set of interpolated points along the actual and synthetic networks be represented by $\mathcal{P}_{\mathrm{act}}$ and $\mathcal{P}_{\mathrm{syn}}$ respectively. Note that $\mathcal{P}_{\mathrm{act}}, \mathcal{P}_{\mathrm{syn}} \subseteq \mathcal{M}$. We define the Hausdorff distance between networks $\mathcal{G}_{\mathrm{act}}$ and $\mathcal{G}_{\mathrm{syn}}$ as follows.

$$\mathsf{D}_{\mathrm{H}}\left(\mathcal{G}_{\mathrm{act}}, \mathcal{G}_{\mathrm{syn}}\right) := \max_{x \in \mathcal{P}_{\mathrm{act}}} \min_{y \in \mathcal{P}_{\mathrm{syn}}} \mathsf{dist}(x, y) \tag{3.2}$$

The above metric of geometry comparison allows us to measure a degree of proximity for edge geometries which are non-overlapping, yet close to each other. Fig. 3.11 shows a Hausdorff distance-based network geometry comparison between actual and synthetic networks for uniform rectangular grid partitions of two different resolutions. This is accompanied by shifting the grid cell demarcations horizontally. We observe that as the cell demarcations are shifted horizontally, the Hausdorff distance between the edge geometries in each grid cell alters. Therefore, the grid resolution and grid demarcations play an important role in the geometry comparison as well. Note that network geometries in certain regions show a significant deviation when compared with low resolution, while comparing with a higher grid resolution shows small deviation. This shows that the networks are fairly close to each other.

3.6 Structural validation using simplicial flat norm

3.6.1 Problem

In recent years, the problem of evaluating quality of reconstructed networks has been studied for street maps. The authors have defined suitable metrics to compare algorithms and

frameworks which use GPS trajectory data to reconstruct street map graphs (Ahmed, Fasy, and Wenk 2014; Ahmed, Fasy, Hickmann, et al. 2015). The abstract problem can be stated as follows: *compute the similarity between a given pair of embedded planar graphs*. This is analogous to the well known subgraph isomorphism problem (Eppstein 1995) wherein we look for isomorphic subgraphs in a pair of given graphs. One major precursor to such a problem is that we require a one-to-one mapping between nodes and edges of the two graphs. Though such mappings are well-defined for street networks, the same cannot be inferred for power distribution networks. Since the power network dataset is proprietary, the node and edge labels are redacted from the network before it is shared. The actual network is obtained as a set of "drawings" with associated geographic embeddings. Each drawing can be considered as a collection of line segments termed a *geometry*. Hence the problem of comparing a set of power distribution networks with geographic embedding can be stated as the following: *compute the similarity between a given pair of geometries lying on a geographic plane*. Following Majhi and Wenk (2022), we use the term *geometric graph* to define network graphs embedded in a Euclidean space. Next, we define what we mean by structurally similar geometric graphs.

Definition 3.1 (Geometric graph)**.** A graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ with node set $\mathcal{V}$ and edge set $\mathcal{E}$ is said to be a geometric graph of $\mathbb{R}^d$ if the set of nodes $\mathcal{V} \subset \mathbb{R}^d$ and the edges are Euclidean straight line segments $\{\overline{uv} \mid e := (u,v) \in \mathcal{E}\}$ which intersect (possibly) at their endpoints.

Definition 3.2 (Structurally similar geometric graphs)**.** Two geometric graphs $\mathcal{G}_0(\mathcal{V}_0,\mathcal{E}_0)$ and $\mathcal{G}_1(\mathcal{V}_1,\mathcal{E}_1)$ are said to be *structurally similar* at the level of $\delta \geq 0$, termed δ-similar, if $\text{dist}(\mathcal{G}_0,\mathcal{G}_1) \leq \delta$ for the distance function dist between the two graphs.

We could consider a given network as a set of edge geometries. Hence we could consider the problem of comparing geometric graphs $\mathcal{G}_0$ and $\mathcal{G}_1$ as that of comparing the set of edge geometries $\mathcal{E}_0$ and $\mathcal{E}_1$. In this work, we propose a suitable distance that allows us to

compare between a pair of geometric graphs or a pair of geometries. We use the multiscale flat norm, which has been well explored in the field of geometric measure theory, to define such a distance between the geometries.

The other aspect of this work is to identify a suitable threshold δ for inferring the structural similarity of a pair of geometric graphs. But there is no general method to choose the threshold. Here, we perform a statistical analysis for our particular case of comparing power distribution networks. We validate the comparison results with visual inspection to conclude that the proposed metric serves its purpose to identify structurally similar geometric graphs.

3.6.2 Multiscale flat norm

We use the multiscale simplicial flat norm proposed by Sharif, Krishnamoorthy, and Vixie (2013) to compute the distance between two networks. We now introduce some background for this computation. A d-dimensional *current* T (referred to as a d-current) is a generalized d-dimensional geometric object with orientations (or direction) and multiplicities (or magnitude). An example of a 2-current is a surface with finite area (multiplicity) and a specific orientation (clockwise or counterclockwise). The boundary of T, denoted by ∂T, is a $(d-1)$-current. The *multiscale flat norm* of a d-current T at scale $\lambda \geq 0$ is defined as

$$\mathbb{F}_\lambda(T) = \min_S \{V_d(T - \partial S) + \lambda V_{d+1}(S)\}, \tag{3.3}$$

where the minimum is taken over all $(d+1)$-currents S, and V_d denotes the d-dimensional *volume*, e.g., length in 1D or area in 2D. Computing the flat norm of a 1-current (curve) T identifies the optimal 2-current (area patches) S that minimizes the sum of the length of current $T - \partial S$ and the area of patch(es) S. Fig. 3.12 shows the flat norm computation for a

generic 1D current T (blue). The 2D area patches S (magenta) are computed such that the expression in Eq. (3.3) is minimized for the chosen value of λ that ends up using most of the patch under the sharper spike on the left but only a small portion of the patch under the wider bump to the right.

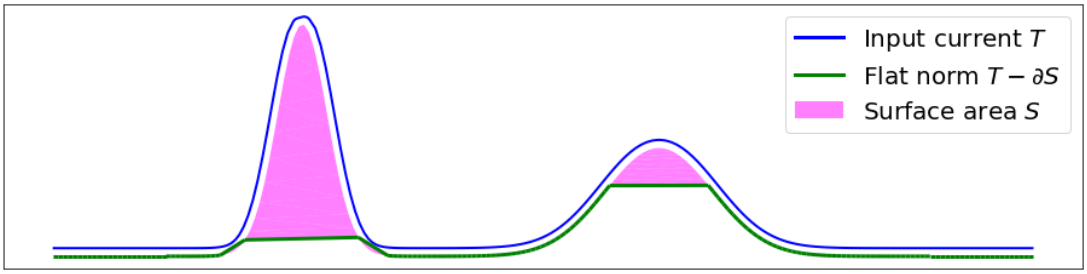

Figure 3.12: Multiscale flat norm of a 1D current T (blue). The flat norm is the sum of length of the resulting 1D current $T - \partial S$ (green) and the area of 2D patches S (magenta). We show $T - \partial S$ slightly separated for easy visualization.

The scale parameter λ can be intuitively understood as follows. Rolling a ball of radius $1/\lambda$ on the 1-current T traces the output current $T - \partial S$ and the untraced regions constitute the patches S. Hence we observe that for a large λ, the radius of the ball is very small and hence it traces major features while smoothing out (i.e., missing) only minor features (wiggles) of the input current. But for a small λ, the ball with a large radius smoothes out larger scale features (bumps) in the current. Note that for smaller λ, the cost of area patches is smaller in the minimization function and hence more patches are used for computing the flat norm. We can use the flat norm to define a natural distance between a pair of 1-currents T_1 and T_2 as follows (Sharif, Krishnamoorthy, and Vixie 2013).

$$\mathbb{F}_\lambda(T_1, T_2) = \mathbb{F}_\lambda(T_1 - T_2) \tag{3.4}$$

We show a simple example depicting the use of flat norm to compute the distance between a pair of geometries that are two line segments of equal length meeting at their midpoints

in Fig. 3.13. As the angle between the two line segments decreases from 90 to 15 degrees, the computed flat norm also decreases.

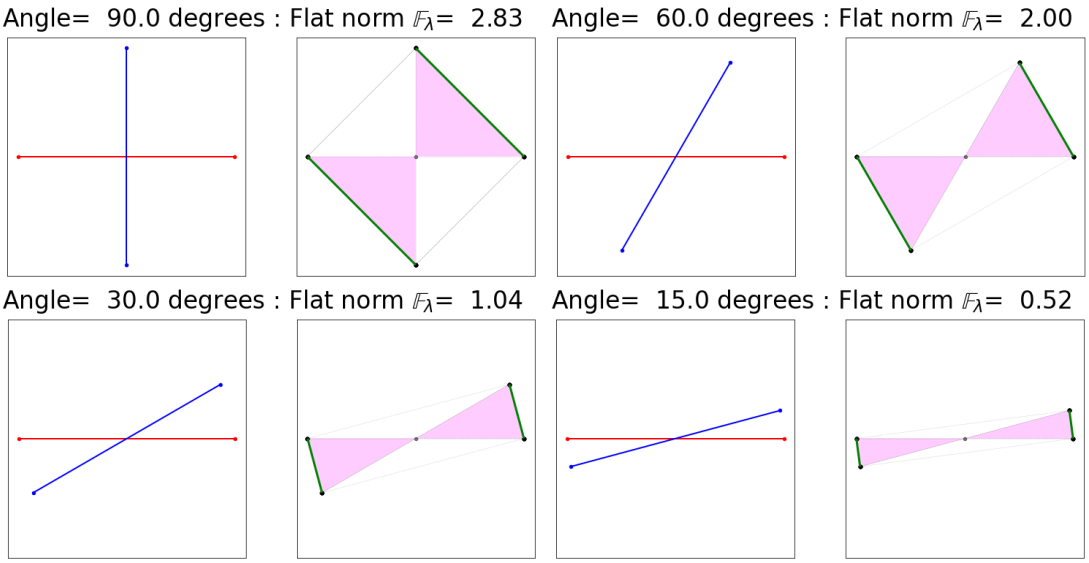

Figure 3.13: Variation in flat norm for pairs of geometries as the angle between them decreases. When the geometries are perpendicular to each other, flat norm distance is the maximum and it decreases as the angle decreases.

3.6.3 Computing the multiscale flat norm

We compute the flat norm distance between a pair of input geometries (synthetic and actual) as the flat norm of the current $T = T_1 - T_2$ where T_1 and T_2 are the currents corresponding to individual geometries. Let Σ denote the set of all line segments in the input current T. We perform a constrained triangulation of Σ to obtain a 2-dimensional finite oriented simplicial complex K. A constrained triangulation ensures that each line segment $\sigma_i \in \Sigma$ is an edge in K, and that T is an oriented 1-dimensional subcomplex of K.

Let m and n denote the numbers of edges and triangles in K. We can denote the input current T as a 1-chain $\sum_{i=1}^{m} t_i \sigma_i$ where σ_i denotes an edge in K and t_i is the corresponding multiplicity. Note that $t_i = -1$ indicates that orientation of σ_i and T are opposite, $t_i = 0$

denotes that σ_i is not contained in T, and $t_i = 1$ implies that σ_i is oriented the same way as T. Similarly, we define the set S to be the 2-chain of K and denote it by $\sum_{i=1}^{m} s_i \omega_i$ where ω_i denotes a 2-simplex in K and s_i is the corresponding multiplicity.

The boundary matrix $[\partial] \in \mathbb{Z}^{m \times n}$ captures the intersection of the 1 and 2-simplices of K. The entries of the boundary matrix $[\partial]_{ij} \in \{-1, 0, 1\}$. If edge σ_i is a face of triangle ω_j, then $[\partial]_{ij}$ is nonzero and it is zero otherwise. The entry is -1 if the orientations of σ_i and ω_j are opposite and it is $+1$ if the orientations agree.

We can respectively stack the t_i's and s_i's in m and n-length vectors $\mathbf{t} \in \mathbb{Z}^m$ and $\mathbf{s} \in \mathbb{Z}^n$. The 1-chain representing $T - \partial S$ is denoted by $\mathbf{x} \in \mathbb{Z}^m$ and is given as $\mathbf{x} = \mathbf{t} - [\partial]\mathbf{s}$. The multi-scale flat norm defined in Eq. (3.3) can be computed by solving the following optimization problem:

$$\begin{aligned} \mathbb{F}_\lambda(T) = \min_{\mathbf{s} \in \mathbb{Z}^n} & \sum_{i=1}^{m} w_i |x_i| + \lambda \left(\sum_{j=1}^{n} v_j |s_j| \right) \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{t} - [\partial]\mathbf{s}, \quad \mathbf{x} \in \mathbb{Z}^m, \end{aligned} \tag{3.5}$$

where $V_d(\tau)$ in Eq. (3.3) denotes the volume of the d-dimensional simplex τ. We denote volume of the edge σ_i as $V_1(\sigma_i) = w_i$ and set it to be the Euclidean length, and volume of a triangle ω_j as $V_2(\omega_i) = v_j$ and set it to be the area of the triangle.

In this work, we consider geometric graphs embedded on the geographic plane and are associated with longitude and latitude coordinates. We compute the Euclidean length of edge σ_i as $w_i = R\Delta\phi_i$ where $\Delta\phi_i$ is the Euclidean normed distance between the geographic coordinates of the terminals of σ_i and R is the radius of the earth. Similarly, the area of triangle τ_j is computed as $v_j = R^2\Delta\Omega_j$ where $\Delta\Omega_j$ is the solid angle subtended by the geographic coordinates of the vertices of τ_j.

Using the fact that the objective function is piece-wise linear in $\mathbf{x}$ and $\mathbf{s}$, the minimization

problem can be reformulated as an integer linear program (ILP) as follows:

$$\mathbb{F}_\lambda(T) = \min \sum_{i=1}^{m} w_i \left(x_i^+ + x_i^-\right) + \lambda \left(\sum_{j=1}^{n} v_j \left(s_j^+ + s_j^-\right) \right) \tag{3.6a}$$

$$\text{s.t.} \quad \mathbf{x}^+ - \mathbf{x}^- = \mathbf{t} - [\partial]\left(\mathbf{s}^+ - \mathbf{s}^-\right) \tag{3.6b}$$

$$\mathbf{x}^+, \mathbf{x}^- \geq 0, \quad \mathbf{s}^+, \mathbf{s}^- \geq 0 \tag{3.6c}$$

$$\mathbf{x}^+, \mathbf{x}^- \in \mathbb{Z}^m, \quad \mathbf{s}^+, \mathbf{s}^- \in \mathbb{Z}^n \tag{3.6d}$$

The linear programming relaxation of the ILP in Eq. (3.6) is obtained by ignoring the integer constraints Eq. (3.6d). We refer to this relaxed linear program (LP) as the ***flat norm LP***.

$$\mathbb{F}_\lambda(T) = \min \sum_{i=1}^{m} w_i \left(x_i^+ + x_i^-\right) + \lambda \left(\sum_{j=1}^{n} v_j \left(s_j^+ + s_j^-\right) \right) \tag{3.7a}$$

$$\text{s.t.} \quad \mathbf{x}^+ - \mathbf{x}^- = \mathbf{t} - [\partial]\left(\mathbf{s}^+ - \mathbf{s}^-\right) \tag{3.7b}$$

$$\mathbf{x}^+, \mathbf{x}^- \geq 0, \quad \mathbf{s}^+, \mathbf{s}^- \geq 0 \tag{3.7c}$$

Sharif, Krishnamoorthy, and Vixie (2013) showed that the boundary matrix $[\partial]$ is totally unimodular for our application setting. Hence the flat norm LP will solve the ILP, and hence the flat norm can be computed in polynomial time.

We demonstrate the steps involved in computing the flat norm for a pair of input geometries in Fig. 3.14. The input geometries are a collection of line segments shown in blue and red (top left). We construct the set Σ by combining all the edges of either geometry along with the bounding rectangle (top right). Thereafter, we perform a constrained triangulation to construct the 2-dimensional simplicial complex K (bottom left). Finally, we compute

the multiscale simplicial flat norm with $\lambda = 1$ (bottom right). Note that this computation captures the length deviation (shown by green edges) and the lateral displacement (shown by the magenta patches).

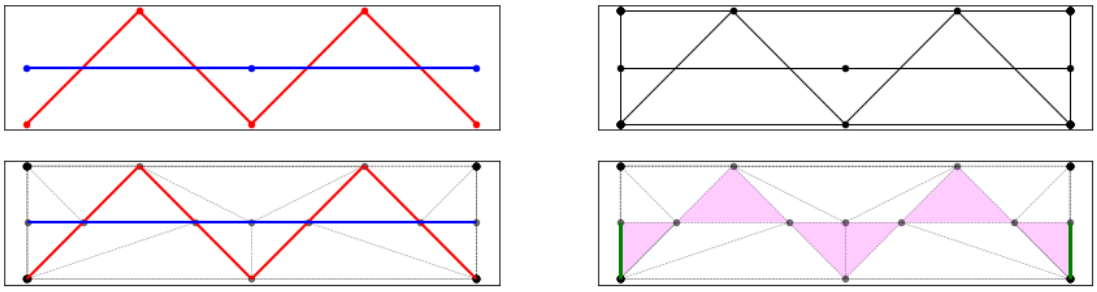

Figure 3.14: Steps in computing the flat norm for a pair of input geometries.

3.6.4 Proposed algorithm

Algorithm 5 describes how we compute the distance between a pair of geometries with the associated embedding on a metric space $\mathscr{M}$. We assume that the geometries (networks) $\mathcal{G}_1(\mathcal{V}_1, \mathcal{E}_1)$ and $\mathcal{G}_2(\mathcal{V}_2, \mathcal{E}_2)$ with respective node sets $\mathcal{V}_1, \mathcal{V}_2$ and edge sets $\mathcal{E}_1, \mathcal{E}_2$ have no one-to-one correspondence between the $\mathcal{V}_i$'s or $\mathcal{E}_i$'s. Note that each vertex $v \in \mathcal{V}_1, \mathcal{V}_2$ is a point and each edge $e \in \mathcal{E}_1, \mathcal{E}_2$ is a straight line segment in $\mathscr{M}$. We consider the collection of edges $\mathcal{E}_1, \mathcal{E}_2$ as input to our algorithm. First, we orient the edge geometries in a particular direction (left to right in our case) to define the currents T_1 and T_2, which have both magnitude and direction. Next, we consider the bounding rectangle $\mathcal{E}_{\text{bound}}$ for the edge geometries and define the set Σ to be triangulated as the set of all edges in either geometry and the bounding rectangle. We perform a constrained Delaunay triangulation (Si 2010) on the set Σ to construct the 2-dimensional simplicial complex K. The constrained triangulation ensures that the set of edges in Σ is included in the simplicial complex K. Then we define the currents T_1 and T_2 corresponding to the respective edge geometries $\mathcal{E}_1$ and $\mathcal{E}_2$ as

1-chains in K. Finally, the flat norm LP is solved to compute the simplicial flat norm.

Algorithm 5 Distance between a pair of geometries

Input: Geometries $\mathcal{E}_1, \mathcal{E}_2$
Parameter: Scale λ
Step 1: Orient each edge in the edge sets from left to right: $\tilde{\mathcal{E}}_1 := \mathsf{Orient}(\mathcal{E}_1)$; $\tilde{\mathcal{E}}_2 := \mathsf{Orient}(\mathcal{E}_2)$.
Step 2: Find bounding rectangle for the pair of geometries: $\mathcal{E}_{\text{bound}} = \mathsf{rect}\left(\tilde{\mathcal{E}}_1, \tilde{\mathcal{E}}_2\right)$.
Step 3: Define the set of line segments to be triangulated: $\Sigma = \tilde{\mathcal{E}}_1 \cup \tilde{\mathcal{E}}_2 \cup \mathcal{E}_{\text{bound}}$.
Step 4: Perform constrained triangulation on set Σ to construct 2-dimensional simplicial complex K.
Step 5: Define the currents T_1, T_2 as 1-chains of oriented edges
Step 6: $\tilde{\mathcal{E}}_1$ and $\tilde{\mathcal{E}}_2$ in K.
Step 7: Solve the flat norm LP to compute flat norm $\mathbb{F}_\lambda(T_1 - T_2)$.
Output: Flat norm distance $\mathbb{F}_\lambda(T_1 - T_2)$.

3.6.5 Normalized flat norm

Recall that in our context of synthetic power distribution networks, the primary goal of comparing a synthetic network to its actual counterpart is to infer the quality of the replica or the *digital duplicate* synthesized by the framework. The proposed approach using the flat norm for structural comparison of a pair of geometries provides us a method to perform global as well as local comparison. While we can produce a global comparison by computing the flat norm distance between the two networks, it may not provide us with complete information on the quality of the synthetic replicate. On the other hand, a local comparison can provide us details about the framework generating the synthetic networks. For example, a synthetic network generation framework might produce higher quality digital replicates of actual power distribution networks for urban regions as compared to rural areas. A local comparison highlights this attribute and identifies potential use case scenarios of a given synthetic network generation framework.

Furthermore, availability of actual power distribution network data is sparse due to its pro-

prietary nature. We may not be able to produce a global comparison between two networks due to unavailability of network data from one of the sources. Hence, we want to restrict our comparison to only the portions in the region where data from either network is available, which also necessitates a local comparison between the networks.

For a local comparison, we consider uniform sized regions and compute the flat norm distance between the pair of geometries within the region. However, the computed flat norm is dependent on the length of edges present within the region from either network. Hence we define the *normalized* multiscale flat norm, denoted by $\widetilde{\mathbb{F}}_\lambda$, for a given region as

$$\widetilde{\mathbb{F}}_\lambda\left(T_1 - T_2\right) = \frac{\mathbb{F}_\lambda\left(T_1 - T_2\right)}{|T_1| + |T_2|}. \tag{3.8}$$

For a given parameter ε, a local region is defined as a square of size $2\varepsilon \times 2\varepsilon$ steradians. Let $T_{1,\varepsilon}$ and $T_{2,\varepsilon}$ denote the currents representing the input geometries inside the local region characterized by ε. Note that the "amount" or the total length of network geometries within a square region varies depending on the location of the local region. In this case, the lengths of the network geometries are respectively $|T_{1,\varepsilon}|$ and $|T_{2,\varepsilon}|$. Therefore, we use the ratio of the total length of network geometries inside a square region to the parameter ε to characterize this "amount" and denote it by $|T|/\varepsilon$ where

$$|T|/\varepsilon = \frac{|T_{1,\varepsilon}| + |T_{2,\varepsilon}|}{\varepsilon}. \tag{3.9}$$

Note that while performing a comparison between a pair of network geometries in a local region using the multiscale flat norm, we need to ensure that comparison is performed for similar length of the networks inside similar regions. Therefore, the ratio $|T|/\varepsilon$, which indicates the length of networks inside a region scaled to the size of the region, becomes an important aspect of characterization while performing the flat norm based comparison.

3.6.6 Flat norm computation for power networks

Figure 3.15: Steps showing simplicial flat norm distance computation between two networks (top left plot). The convex rectangular boundary around the pair of networks is identified (top right plot). A constrained triangulation is performed such that the edges in the networks and convex boundary are edges of triangles (bottom left plot). Finally the optimization problem is solved to compute the simplicial flat norm, which includes sum of areas of the magenta triangles and lengths of green edges (bottom right plot).

Fig. 3.15 shows the steps involved in computing the simplicial flat norm distance between a pair of networks. These include the actual power distribution network (red) for a region in a county from USA and the synthetic network (blue) constructed for the same region using the framework proposed in 2. First, we orient each edge in either network from left to right. Thereafter, we find the convex rectangular boundary around the pair of networks. We perform a constrained triangulation, such that the edges in the networks and the convex boundary are selected as edges of the triangles. Finally the optimization problem in Eq.(3.6) is solved to compute the flat norm distance between the networks.

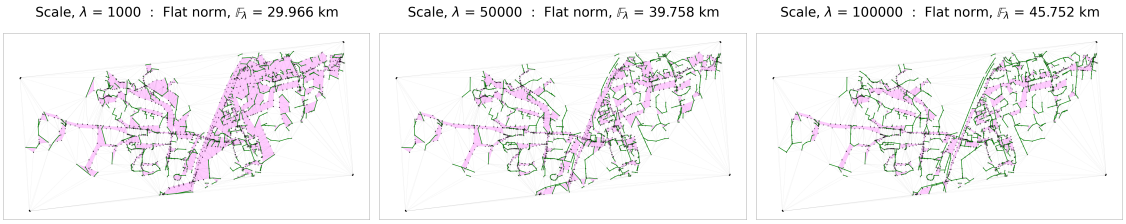

Figure 3.16: The flat norm computed between the pair of network geometries for three values of the scale parameter λ ranging between $\lambda = 1000$ to $\lambda = 10000$.

The multi-scale flat norm produces different norm values for three values of the scale parameter λ. Fig. 3.16 shows the simplicial flat norm between the actual and synthetic power

network for the same region for multiple values of the scale parameter λ. We observe that for different values of the scale parameter, the 2-D patches considered in the computed flat norm also changes.

The variation of the computed flat norm for different values of the scale parameter is summarized in Fig. 3.17. As the scale parameter is increased, less number of area patches are considered in the simplicial flat norm computation. This is denoted by the blue decaying line in the plot. The computed flat norm increases for larger values of the scale parameter λ as more lengths of individual currents is considered in its computation. We show the plot on two different scales: the left scale indicates the deviation in length (measured in km), and the right scale denotes the deviation expressed through the area patches (measured in sq.km).

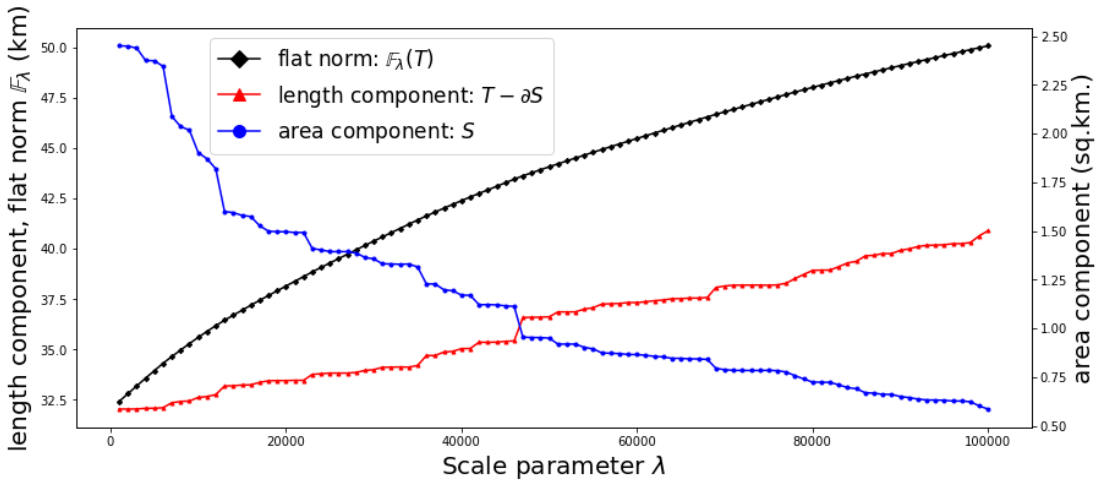

Figure 3.17: Plot showing the effect of varying scale parameter λ in the flat norm computation. The flat norm for a 1-dimensional current embedded on a 2-dimensional consists of two parts: length component and a scaled surface area component. The variations in the length component and the unscaled surface area component are also shown in the plot.

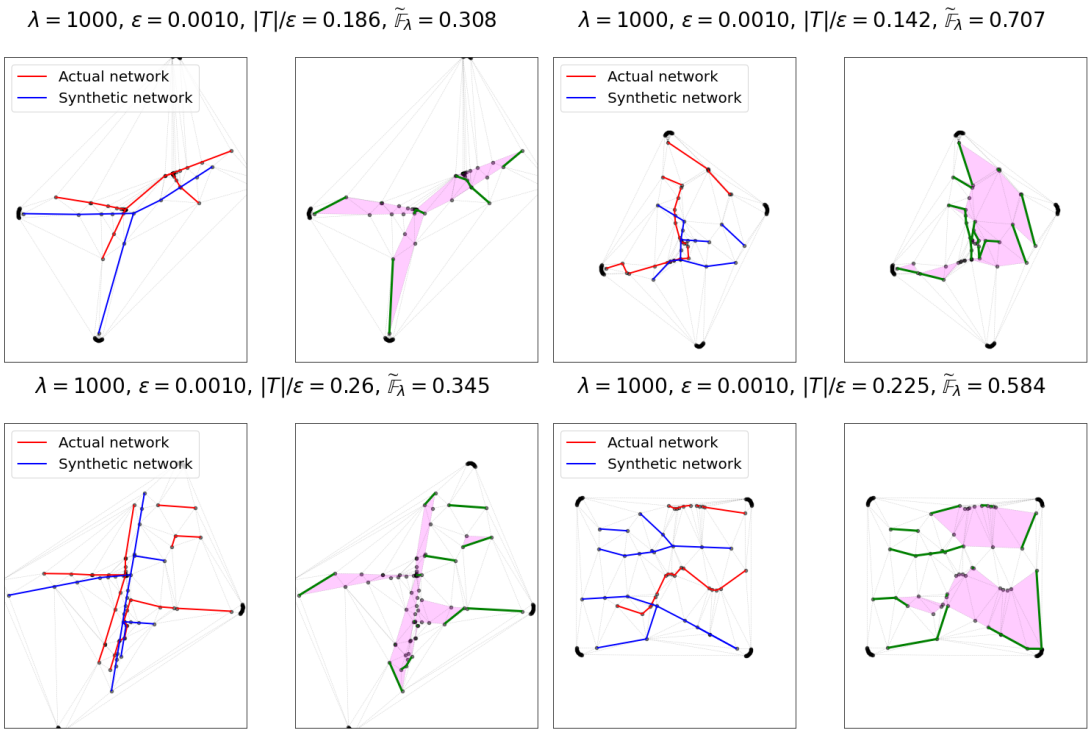

Figure 3.18: Plots showing normalized flat norm (with scale $\lambda = 1000$) computed for different regions in the network of same size ($\varepsilon = 0.001$) with similar $|T|/\varepsilon$ ratio along each row (top and bottom). From a mere visual comparison, we notice that the pair of geometries for the left plots are almost similar which is reflected in the low flat norm distance between them. The network geometries on the right plots do not resemble and hence the flat norm distance is high.

3.6.7 Comparing network geometries using the normalized flat norm

The primary goal of computing the flat norm is to compare the pair of input geometries. As mentioned earlier, the flat norm provides an accurate measure of difference between the geometries by considering the length deviation and area patches in between the geometries. Further, we normalize the computed flat norm to the total length of the geometries. In this section, we show few examples where we compute the normalized flat norm for the pair of network geometries (actual and synthetic) for a few regions.

The top two plots in Fig. 3.18 show two regions characterized by $\varepsilon = 0.001$ and almost similar $|T|/\varepsilon$ ratios. This indicates that the length of network scaled to the region size is almost equal for the two regions. From a mere visual perspective, we can conclude that the first pair of network geometries resemble each other where as the second pair are fairly different. This is further validated from the results of the flat norm distance between the network geometries computed with the scale $\lambda = 1000$, since the first case produces a smaller flat norm distance compared to the latter. The bottom two plots show another example of two regions with almost similar $|T|/\varepsilon$ ratios and enable us to infer similar conclusions. The results strengthens our case of using flat norm as an appropriate measure to perform a local comparison of network geometries.

3.6.8 Statistical considerations

Empirical distribution of normalized flat norm

In this section we study the empirical distribution of the normalized flat norm $\widetilde{\mathbb{F}}_\lambda$ for different local regions and argue that it indeed captures the similarity between input geometries. To this end, we use Algorithm 6 to sample random square shaped regions of size $2\varepsilon \times 2\varepsilon$

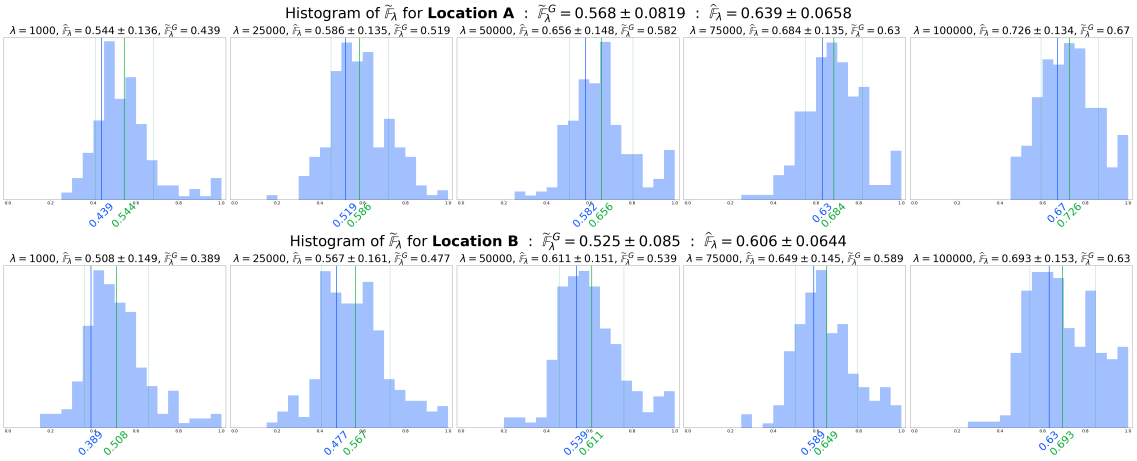

Figure 3.19: Distribution of normalized flat norm computed using five different values of λ for 200 uniformly sampled local regions in Location A (top) and Location B (bottom). The blue line in each histogram denotes the global normalized flat norm $\widetilde{\mathbb{F}}_\lambda^G$ computed for the location with the corresponding scale λ. The solid green line denotes the mean normalized flat norm $\widehat{\mathbb{F}}_\lambda$ for the uniformly sampled local regions computed with scale λ. The dashed green lines show the spread of the distribution.

steradians from a given geographic location.

We perform our empirical studies for two urban locations of a county in USA. These locations have been identified as 'Location A' and 'Location B' for the remainder of this section. We consider local regions of sizes characterized by $\varepsilon \in \{0.0005, 0.001, 0.0015, 0.002\}$. For each location, we randomly sample $N = 50$ local regions for each value of ε using Algorithm 6 and hence we consider $50 \times 4 = 200$ regions. For every sampled region, we use Algorithm (5) to compute the simplicial flat norm between the network geometries contained within the region with scale parameter $\lambda \in \{10^3, 25 \times 10^3, 50 \times 10^3, 75 \times 10^3, 10^5\}$. Thereafter, we normalize the computed flat norm using Eq.3.8. Additionally, we compute the global normalized flat norm for the entire location and indicate it by $\widetilde{\mathbb{F}}_\lambda^G$. The corresponding square box bounding the entire location is characterized by ε_G. We also denote the total length of networks in each location scaled by the size of the location using the ratio $|T_G|/\varepsilon_G$.

Algorithm 6 Sample square regions from location

Input: Geometries $\mathcal{E}_1, \mathcal{E}_2$, number of regions N
Parameter: Size of region ε

Step 1: Find bounding rectangle for the pair of geometries: $\mathcal{E}_{\text{bound}} = \mathsf{rect}(\mathcal{E}_1, \mathcal{E}_2)$.
Step 2: Initialize set of regions: $\mathcal{R} \leftarrow \{\}$.
Step 3: **while** $|\mathcal{R}| \leq N$ **do**
Step 4: Sample a point (x,y) uniformly from region bounded by $\mathcal{E}_{\text{bound}}$.
Step 5: Define the square region $r(x,y)$ formed by the corner points $\{(x-\varepsilon, y-\varepsilon), (x+\varepsilon, y_i+\varepsilon)\}$.
Step 6: **if** $r(x,y) \cap \mathcal{E}_1 \cap \mathcal{E}_2 \neq \emptyset$ **then**
Step 7: Add region $r(x,y)$ to the set of sampled regions: $\mathcal{R} \leftarrow \mathcal{R} \cup \{r(x,y)\}$.
Step 8: **end if**
Step 9: **end while**

Output: Set of sampled regions: $\mathcal{R}$.

First, we show the histogram of normalized flat norms computed for the two locations: Location A and Location B with the five different scale parameter λ. Each histogram shows the empirical distribution of normalized flat norm values $\widetilde{\mathbb{F}}_\lambda$ for uniformly sampled 200 local regions (50 regions for each ε). In each histogram, we additionally record the global normalized flat norm between the network geometries of the location $\widetilde{\mathbb{F}}^G_\lambda$ and denote it by the solid blue line. We also show the mean normalized flat norm $\widehat{\mathbb{F}}_\lambda$ using the solid green line and the dashed green lines indicate the standard deviation of the distribution. Note that for a high scale parameter λ, the distribution is skewed towards the right. This observation follows from our previous discussion of the dependence of scale parameter on the computed flat norm. For a large λ, the area patches are weighed higher in the objective function of the optimization problem Eq. 3.6. Therefore, the lengths of the input currents T_1 and T_2 are included in the flat norm, that is, $\mathbb{F}_\lambda(T_1 - T_2) \to |T_1| + |T_2|$. Hence, the normalized flat norm closer to 1. For the remainder of the section, we will continue our discussion with scale parameter $\lambda = 1000$, since the empirical distributions of normalized flat norm corresponding to $\lambda = 1000$ indicate almost Gaussian distribution.

Next, we consider the empirical distribution of normalized flat norm computed with scale

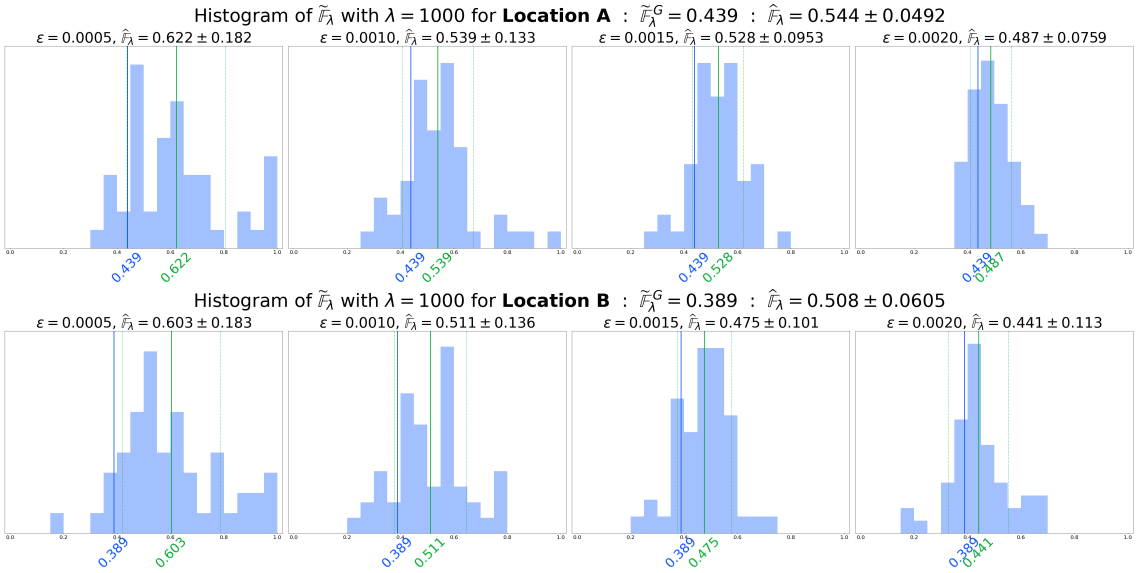

Figure 3.20: Distribution of normalized flat norm computed using $\lambda = 1000$ for 50 uniformly sampled local regions with four different sizes ε in Location A (top) and Location B (bottom). The blue line in each histogram denotes the global normalized flat norm $\widetilde{\mathbb{F}}_\lambda^G$. The solid green line denotes the mean normalized flat norm $\widehat{\mathbb{F}}_\lambda$ for the uniformly sampled local regions. The dashed green lines show the spread of the distribution.

parameter $\lambda = 1000$ for uniformly sampled local regions in the two locations: Location A and Location B. We show separate histogram for four different sized local regions (different values of ε). Note that for small sized local region (low ε), the distribution is skewed towards the right. This is because when we consider small regions, we often capture very isolated network geometries and the flat norm computation is close to the total network length, that is, $\mathbb{F}_\lambda\,(T_1 - T_2) \to |T_1| + |T_2|$ which again leads the normalized flat norm to be close to 1. For larger sized local regions, such occurrences are avoided and therefore, we do not observe any skewed distribution.

Analysis of normalized flat norm for local regions

The scatter plot in the top left of Fig. 3.21 shows the empirical distribution of $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values. The scatter plot highlights (blue star) the global value $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}_\lambda^G\right)$ of Location

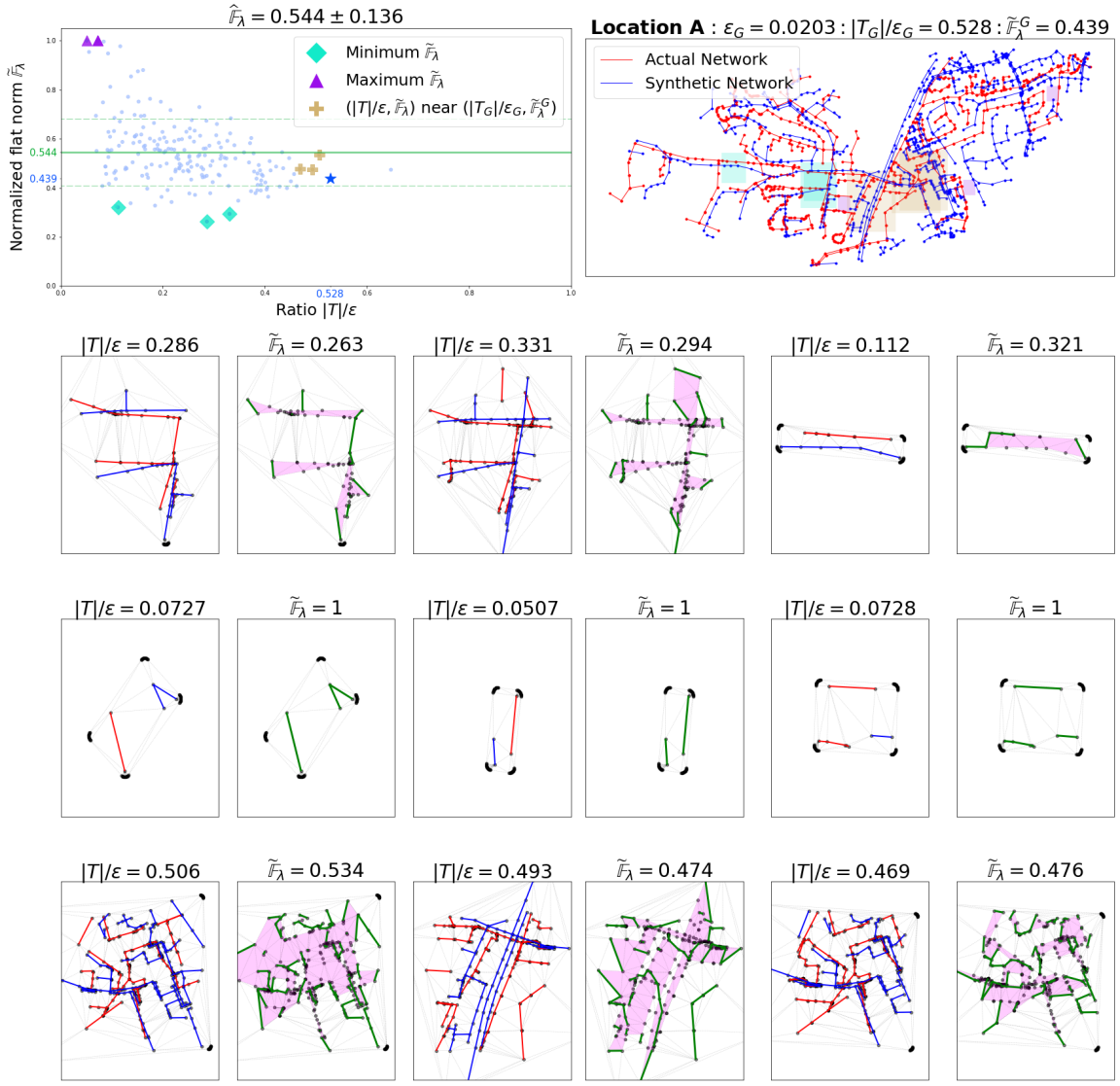

Figure 3.21: Plots showing normalized flat norm computed for entire Location A and few local regions within it. The scatter plot (top left plot) shows the empirical distribution of $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values with the global normalized flat norm $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}^G_\lambda\right)$ for the region (blue star). Nine local regions (three with small $\widetilde{\mathbb{F}}_\lambda$, three with large $\widetilde{\mathbb{F}}_\lambda$ and three with $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values close to the global value $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}^G_\lambda\right)$) are additionally highlighted. The local regions are highlighted along with the pair of network geometries (top right plot). The normalized flat norm computation (with scale $\lambda = 1000$) for the local regions are shown in bottom plots.

A, which indicates the normalized flat norm computed for the entire location. The global normalized flat norm (with a scale parameter $\lambda = 1000$) for Location A is $\widetilde{\mathbb{F}}^G_\lambda = 0.439$ and the ratio $|T_G|/\varepsilon_G = 0.528$. Further, nine additional points are highlighted in the scatter plot denoting nine local regions within Location A. The solid green line denotes the mean of the normalized flat norm values and the dashed green lines indicate the spread of the values around the mean.

The nine local regions are selected such that three of them have the minimum $\widetilde{\mathbb{F}}_\lambda$ in the location (highlighted by cyan colored diamond), three of them have the maximum $\widetilde{\mathbb{F}}_\lambda$ in the location (highlighted by purple triangles) and the remaining three local regions have the $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values close to the global value $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}^G_\lambda\right)$ for the location (highlighted by tan plus symbols). The network geometries within each region and the flat norm computation with scale $\lambda = 1000$ is shown in the bottom plots. The computed flat norm $\widetilde{\mathbb{F}}_\lambda$ and ratio $|T|/\varepsilon$ values are shown above each plot. The local regions are also highlighted (cyan, purple and tan colored boxes) in the top right plot where the actual and synthetic network geometries within the entire location is displayed.

Fig. 3.22 shows similar local regions from Location B. From a mere visual inspection from either of Figs. 3.21 or 3.22, we notice that the network geometries in each local region shown in the first row of the bottom plots resemble and almost overlap each other. The computed normalized flat norm $\widetilde{\mathbb{F}}_\lambda$ values for these local regions agree to this observation. Similarly, the large value of the normalized flat norm justifies the observation that network geometries for local regions depicted in the second row of the bottom plots do not resemble each other. These observations validate our choice of using normalized flat norm as a suitable measure to compare network geometries for local regions.

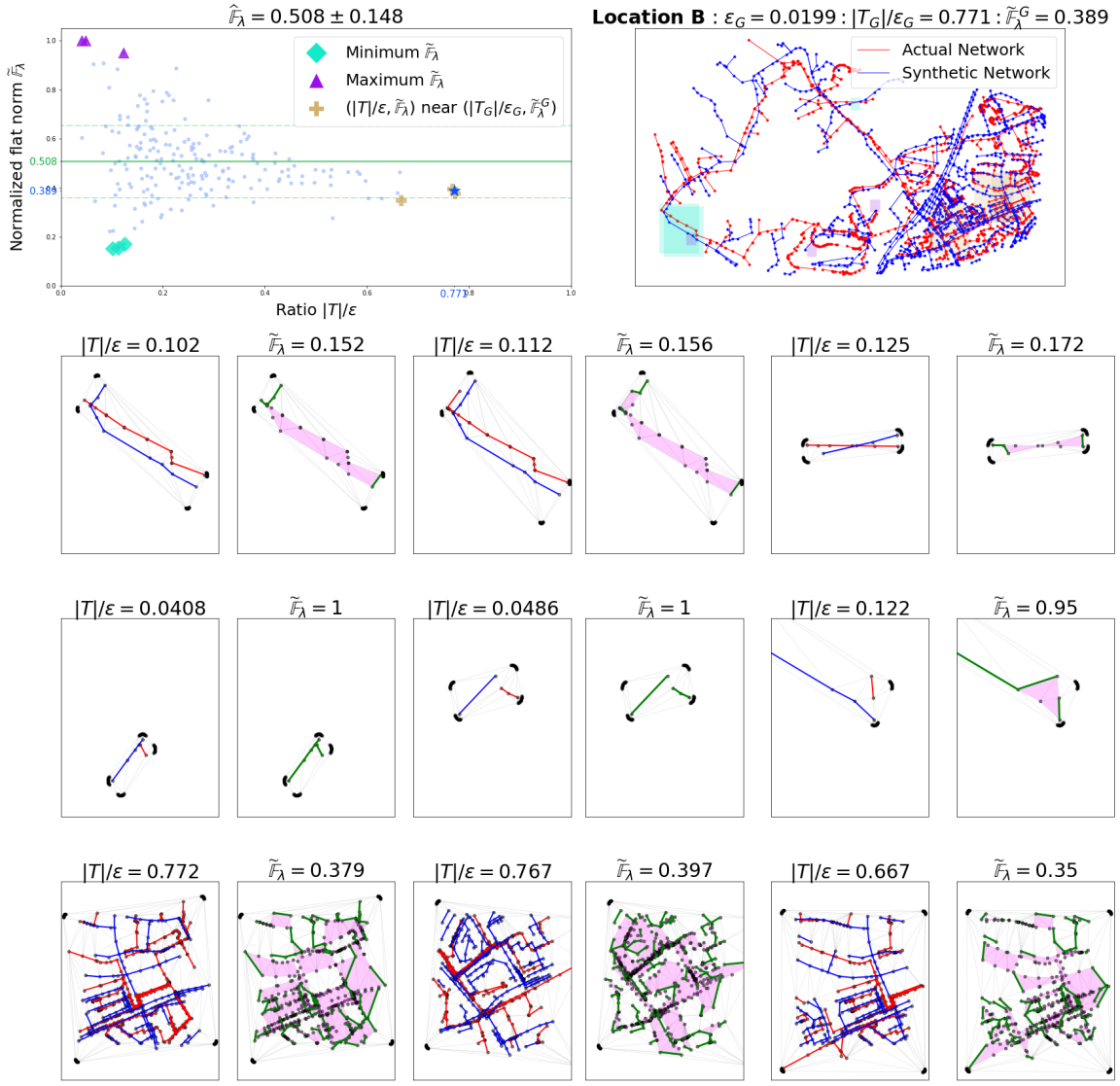

Figure 3.22: Plots showing normalized flat norm computed for entire Location B and few local regions within it. The scatter plot (top left plot) shows the empirical distribution of $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values with the global normalized flat norm $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}^G_\lambda\right)$ for the region (blue star). Nine local regions (three with small $\widetilde{\mathbb{F}}_\lambda$, three with large $\widetilde{\mathbb{F}}_\lambda$ and three with $\left(|T|/\varepsilon, \widetilde{\mathbb{F}}_\lambda\right)$ values close to the global value $\left(|T_G|/\varepsilon_G, \widetilde{\mathbb{F}}^G_\lambda\right)$) are additionally highlighted. The local regions are highlighted along with the pair of network geometries (top right plot). The normalized flat norm computation (with scale $\lambda = 1000$) for the local regions are shown in bottom plots.

Chapter 4

Using Synthetic Power Distribution Networks

Publications

1. Rounak Meyur, Anil Vullikanti, Samarth Swarup, Henning S. Mortveit, Virgilio Centeno, Arun Phadke, H. Vincent Poor, and Madhav Marathe, "Ensembles of realistic power distribution networks" in Proceedings of the National Academy of Sciences, Vol. 119 No. 42, Oct 2022.

2. Rounak Meyur, Swapna Thorve, Madhav Marathe, Anil Vullikanti, Samarth Swarup, and Henning S. Mortveit, "A reliability-aware distributed framework to schedule residential charging of electric vehicles.", in the Proceedings of the thirty-first international joint conference on artificial intelligence, IJCAI-22 (pp. 5115–5121). AI for Good track, Jul 2022, Vienna, Austria.

The main aim of the synthetic distribution network framework presented in Chapter 2 is to create an ensemble of power distribution networks for a given geographic region and perform planning studies for energy systems. In this chapter, we consider few of the use cases of synthetic distribution networks which have been created by our framework. First, we consider the case of photovoltaic (PV) penetration in distribution networks. Next, we

use the created networks to perform short circuit analysis. Finally, we consider the problem of reliability aware residential electric vehicle (EV) charging and discuss it in detail. In all the problems, we use the created synthetic power distribution networks for performing the experimental and simulation studies.

4.1 Impact of photovoltaic penetration

Installing photovoltaic (PV) generation on residence rooftops is an easy and viable option to improve the resilience of distribution systems. However, the continuous increase in PV penetration has led to serious operational issues such as overvoltage and unbalances in the distribution grid (Fatima, Püvi, and Lehtonen 2020). Therefore, it is necessary to identify the maximum allowable PV penetration without violating operational and performance constraints.

It has been observed that a penetration as low as 2.5% is capable of causing voltage violation when a large PV generation is installed at a single point in the MV network. On the contrary, the LV network can withstand multiple PV generators with penetration levels up to 110% (Aziz and Ketjoy 2017). In this work, we focus on the impact of PV penetration in the LV networks through the installation of PV generation on residence rooftops. We aim to address the following question: does PV penetration provide any advantage to grid operation, and how much PV penetration is acceptable without violating operational constraints?

While planning a practical network, the LV feeders are designed with a suitable rating to supply a certain load to certain groups of residences without considering the impact of distributed generation (DGs). Most grid connected DGs need to confer with the *CAN/CSA-C22.2 No.257-06. Interconnecting Inverter-Based Micro-Distributed Resources To Distri-*

bution Systems (2006) standard, which specifies the electrical requirements for the interconnection of inverter-based micro-distributed resource systems with grid-connected LV systems. For single phase connection, it is considered normal operating conditions (NR-Normal Range) when the voltage level is between 0.917 and 1.042 pu. On extreme operation conditions (ER-Emergency Range), the under and over steady state voltage limits are 0.88 and 1.058 pu, respectively. It is worth mentioning that although networks are allowed to operate under extreme conditions, the utility would need to take a corrective action.

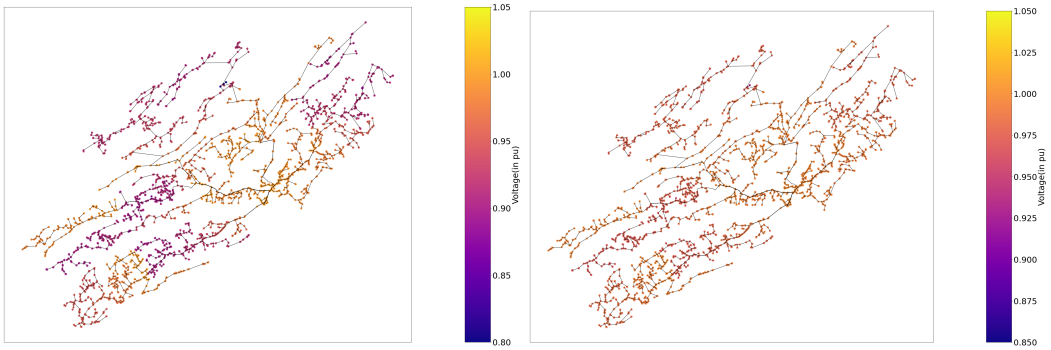

Figure 4.1: Comparison of voltage profile without (left) and with (right) PV penetration in the distribution network. With the introduction of rooftop PVs, the voltage profile is alleviated to close to 1 pu, which is recommended by engineering standards.

If a distribution network is operated without any PV penetration, the substation feeder is solely responsible for delivering power to all residences in the network. Therefore, the farthest connected residence suffers the maximum voltage drop and experience a voltage sag. This is denoted in the left figure of Fig. 4.1 where we observe that a significant percentage of nodes experiences undervoltage (less than 0.9 pu). The introduction of rooftop PVs installed on randomly chosen 50% of residences alleviates the voltage profile of the entire grid as shown in the right figure of Fig. 4.1. In this example, we have considered a 30% PV penetration , that is the total PV generation in the distribution network amounts to 30% of the total load. Therefore, the substation feeders are responsible for delivering the remaining 70% of the power.

We perform the comparison on two ensembles of synthetic networks - one belonging to an urban area and the other corresponds to a rural region. Fig. 4.2 shows the histogram of node voltages for comparing the impact of PV penetration in the urban feeder (top three figures) and the rural feeder (bottom three figures). The urban feeder network consists of shorter length lines. We observe that LV level PV generation is less likely to cause overvoltage issues as compared to a single node MV level PV integration. In the case of a rural feeder with longer lines, we observe a similar trend. However, the percentage of nodes experiencing severe overvoltage (around 1.05 pu which is the extreme limit of an acceptable overvoltage) is higher for rural networks as compared to urban feeder networks for all penetration levels when we consider a single node PV integration at the MV level. Therefore an optimal placement of PV generators is required for the rural feeders so that they do not suffer from overvoltage issues.

We now present a representative study where we analyze the impact of photovoltaic (PV) penetration on the system node voltages. We compare the PV penetration in multiple levels of the network (MV primary network or LV secondary network). We consider the following two cases: (i) PV penetration in LV network where PV generators are installed on residence rooftops and (ii) PV penetration in MV network where a single PV generator is installed at a location in the MV network. In the first case, we randomly identify a group of residences (for example, 50% of all residences) and assign PV generation to them. The penetration level indicates the rating of the PV generation installed on these residences. For the latter case, a single node PV penetration represents a 'solar farm' which is connected to the distribution grid and the penetration level indicates the PV generation rating as a fraction of the total load.

We perform the comparison on two different synthetic feeders: urban and rural. An urban distribution network is characterized with shorter lines as compared to rural networks

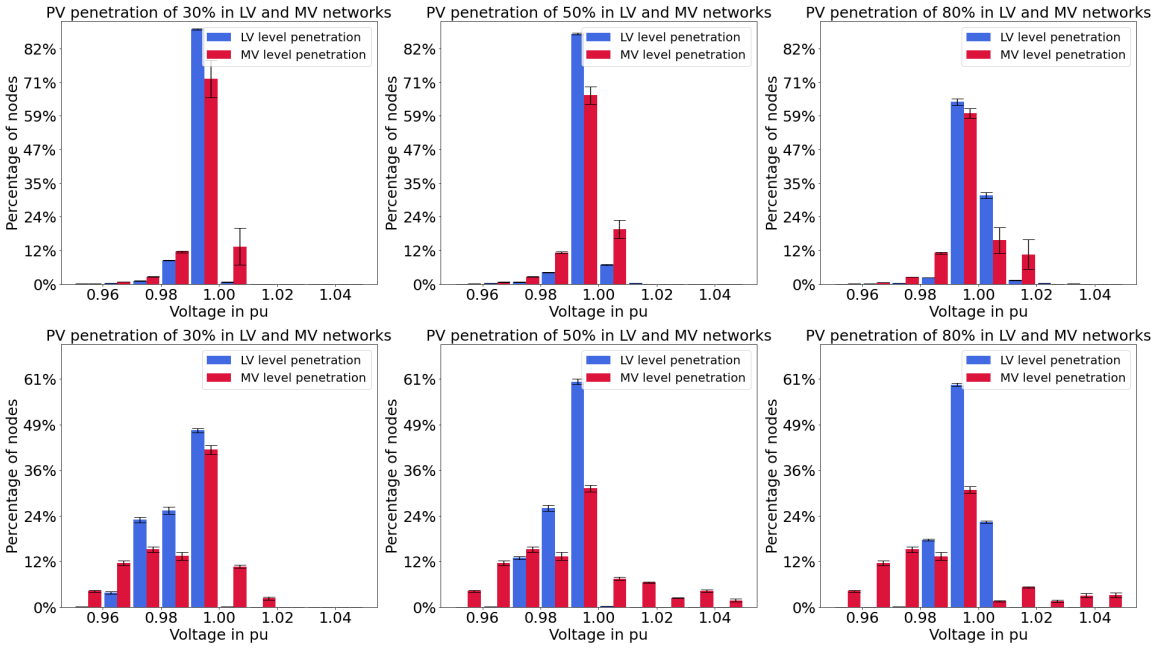

Figure 4.2: Comparing the impact of PV integration in an urban feeder (top figures) and rural feeder (bottom figures) for three different penetration levels. Colors depict PV addition at multiple locations in LV network (blue) and single node PV addition in MV network (red). The error bar shows the variation in impact when analyzed over an ensemble of 20 networks. LV level penetration is less likely to cause overvoltage as compared to similar penetration in the MV network. Rural feeders are more prone to overvoltage issues.

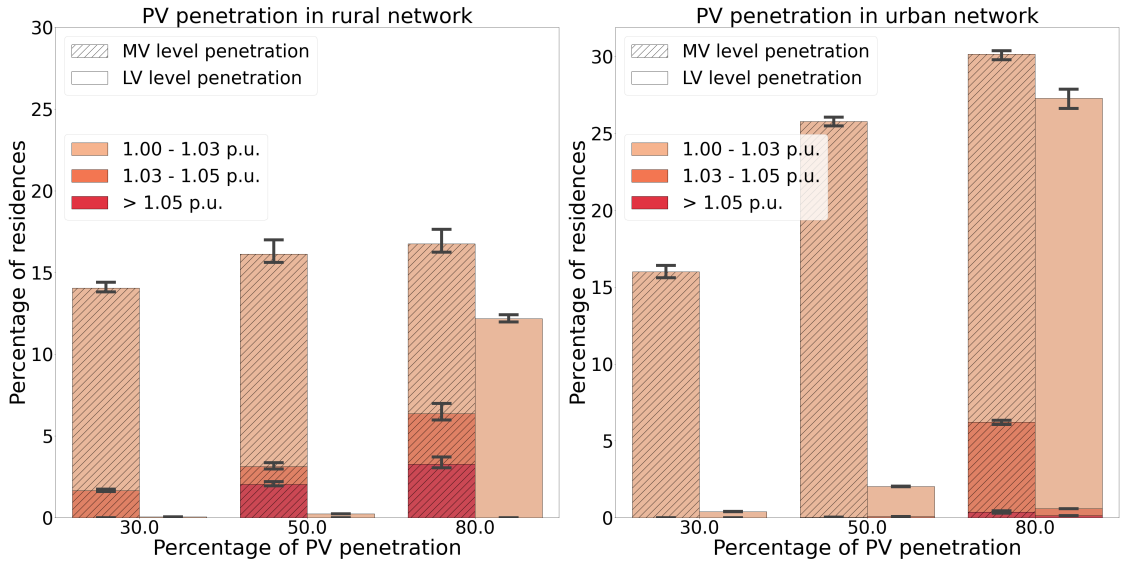

Figure 4.3: Plots showing impact of PV penetration in rural and urban networks. Colors depict the percentage of nodes with various levels of overvoltages. Shaded and non-shaded bars denote MV and LV-level penetration. LV-level penetration is less likely to cause severe overvoltages as compared to MV level penetration. PV penetration in rural networks is more likely to cause overvoltage issues (greater than 1.05 p.u.).

where remote nodes are connected by long lines. Fig. 4.3 compares impact of LV-level and MV-level PV penetration for two networks. Here, we focus on the percentage of nodes which face overvoltage issues due to different levels of PV penetration. We observe that for either case, the percentage of nodes with overvoltage increases with higher penetration level. Further, we see that LV-level PV generation is less likely to cause overvoltage issues as compared to a single node MV-level PV integration. Additionally, in the case of rural feeders, the percentage of nodes experiencing severe overvoltage (around 1.05 pu, which is the extreme limit of acceptable overvoltage) is higher as compared to urban feeder networks. Therefore an optimal placement of PV generators is required for the rural feeders so that they do not suffer from overvoltage issues.

4.2 Short circuit analysis

We can use the created networks to perform short circuit analysis. Since we have created a positive sequence network for the distribution system, we can carry out a short circuit analysis for three phase symmetrical faults. To this end, we construct the bus impedance matrix $\mathbf{Z}_{\text{bus}}$ using the traditional methodology (Grainger and Stevenson 1994).Thereafter, we compute the post-fault voltage at bus j for a three phase symmetrical fault at bus k using Eq. [4.1].

$$V_j^{\text{post-fault}} = V_j^{\text{pre-fault}} - \frac{Z_{jk}}{Z_{kk}} V_k^{\text{pre-fault}}, \tag{4.1}$$

where Z_{jk} and Z_{kk} are entries in the bus impedance matrix $\mathbf{Z}_{\text{bus}}$. Note that the pre-fault voltages are computed by solving the power flow problem using the LDF model (Bolognani and Dörfler 2015) with each residence consuming average hourly demand. Fig. 4.4 shows the percentage of residences experiencing undervoltage instantaneously after the occurrence of a symmetrical three phase fault at different locations in the network. We perform the short circuit analysis for two feeders in the same synthetic network. We simulate symmetrical three phase fault at various locations along each of the two feeders and observe the post-fault voltages at the residence nodes.

Though we have created positive sequence networks, we are able to perform short circuit analysis for unsymmetrical faults (line-ground and line-line). This can be performed under the assumption that all lines in the distribution network are overhead lines. Under such an assumption, the zero sequence impedance is three times the positive sequence impedance of the lines. Thereafter, the fault currents can be computed accordingly (Horowitz and Phadke 1995; Wadhwa 2009; Grainger and Stevenson 1994).

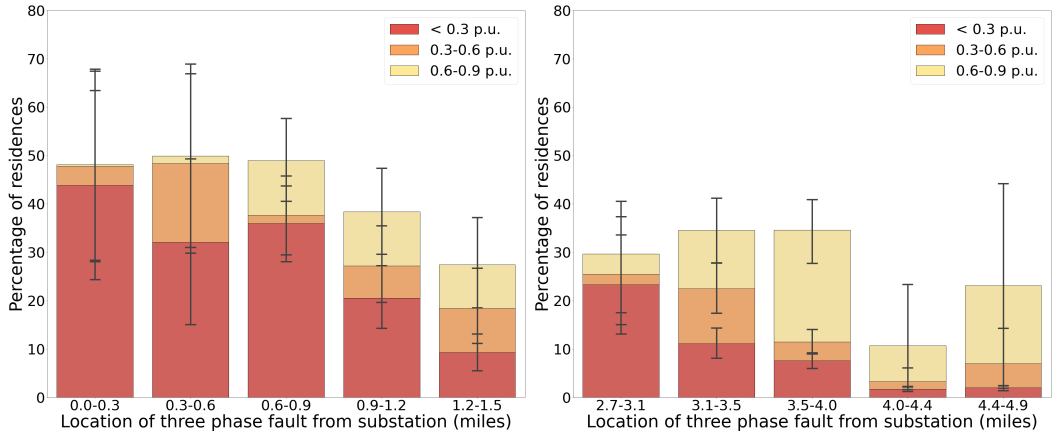

Figure 4.4: Plots showing short circuit analysis on two feeders in a synthetic network. Three phase symmetrical faults are simulated at multiple locations along each of the two feeders and post-fault voltages at residences are computed. The error bars in the plots show the variation in residence voltages for different fault locations along a feeder

4.3 Residential charging of electric vehicles

Studies have shown that home-charging units are pivotal infrastructure for promoting electric vehicle (EV) adoption (Wei et al. 2021). As EV adoption increases over the next few years, the power drawn from the grid will increase and may cause disturbances in the distribution power network. It is quite possible that the network operation may undergo modifications in order to accommodate these unconventional loads in the future without affecting the reliability of the grid.

From the standpoint of power engineering, a *reliable* power grid is one which has adequate generation to support the consumer load demand and can be operated without violating standard power engineering constraints (Billinton and W. Li 1994). Here, we consider an operational problem and therefore are not concerned about the aspect of adequate generation. Thus, we use the term *reliable distribution network* to mean a network that can satisfy the loads without violating the node voltages and line flow (edge flow) capacities. While flows above line rating causes overheating of conductors and subsequent physical dam-

age, node voltages represent the quality of power delivered at the node. Undervoltage and overvoltage at residence nodes lead to eventual failure of household appliances (Kersting 2012).

Traditionally, the distribution network is designed to sustain the peak load demand of consumers (Heidari, Fotuhi-Firuzabad, and Kazemi 2015). The predictable growths in consumer peak demands and energy consumption enables the network operator to plan and operate the network reliably. However, adoption of EVs in residential communities leads to a significant deviation in the predictability of consumer loads (Shao, Pipattanasomporn, and Rahman 2009). Residential EV charging constitutes a significant percentage of net household demand leading to large power consumed from the grid. The problem is more dominant when residential consumers opt to charge EVs according to their personal convenience (Putrus et al. 2009). The excess load consumed by the EV charging units adversely affects distribution grid by causing transformer overloading or high voltage drop at feeder ends (Farkas, Szabó, and Prikler 2011). As a result, it is desirable to develop a framework which aids residential consumers with their goal of scheduling EV charging based on their individual preferences, and simultaneously taking into account the grid reliability requirements of the network operator.

The main contributions of our work in this context are summarized here: (i) A novel 'reliability-aware distributed EV charging scheduling framework' is proposed. It uses information such as the hourly electricity rate, household energy demand profiles & their preferences as inputs for consumers, and power engineering constraints of distribution network as inputs for operator to aid residential EV adopters in scheduling their EV charging units in an optimal manner without affecting the reliability of the power grid. (ii) The distributed framework uses alternating direction method of multipliers (ADMM) based iterative methodology which guarantees an optimal solution for our problem. Each iteration

involves solving a mixed integer quadratic program (MIQP) for each residential consumer and a quadratic program (QP) for the operator. The optimal solutions are exchanged and used in succeeding iterations until a consensus is reached. This minimum exchange of information between the consumers/residences/households and the network operator can be executed using present smart grid infrastructure and avoids sharing of private and proprietary data. (iii) We use digital duplicates of residential consumer load demand profile and power distribution networks resembling the actual physical counterparts for our case studies. This facilitates conducting real-world test scenarios to explore the impact of using the proposed framework while considering multiple levels of EV adoption. Our experiment results demonstrate that the proposed distributed framework helps maintain network reliability compared to the case where EV adopters charge their vehicles based on their personal (individualized) preferences.

4.3.1 Related works

Several works have been presented in the literature for scheduling EV charging. In general, optimization techniques are a popular choice for solving the problem of scheduling EV charging at household level (Cao, S. Tang, et al. 2012; Zhao, Y. Chen, and Keel 2018; S. Lee and Choi 2020; Blonsky, Munankarmi, and Balamurugan 2021; Wi, J. Lee, and Joo 2013; W. Tang, Bi, and Y. J. Zhang 2016; Gonçalves, Gomes, and Antunes 2018; Khonji, Chau, and Elbassioni 2018). Recently, machine learning (ML) techniques such as neural networks (Shuvo and Yilmaz 2021) and reinforcement learning frameworks (Cao, H. Wang, et al. 2022) have been proposed to study the problem of scheduling EV charging loads in smart homes.

Khonji, Chau, and Elbassioni (2018) develop an approximation algorithm and a fast heuris-

tic to solve the scheduling optimization problem of EV charging. A genetic algorithm optimization framework is used to schedule different types of loads in a household such as HVAC, appliances, energy storage system, and EV (Gonçalves, Gomes, and Antunes 2018). Stochastic optimization techniques such as quadratic programming and dynamic programming are proposed for coordinated EV charging with the goal of minimizing the power losses and to maximize the main grid load factor (Clement-Nyns, Haesen, and Driesen 2010). Gan, Topcu, and Low (2012) formulate EV charging scheduling as a discrete optimization problem and is solved in an iterative fashion via communication between EV and transformer.

Most of these works implement a centralized approach to schedule EV charging. A centralized optimization algorithm/framework evaluates the optimal power consumption patterns which are beneficial to only one of the entities – consumers or network operators (M. Liu et al. 2015). This approach may not be realistic since details of the individual consumer load is usually not accessible to the network operator. At the same time, the network topology and parameters are unknown to the consumers. Under such circumstances, a de-centralized/distributed framework is useful since it can help network operators and consumers communicate essential information that can respect both – network reliability and consumer preferences. The current smart grid infrastructure supports the development of such a framework due to the availability of two-way communication.

Dall'Anese et al. (2014) use an alternating direction method of multipliers (ADMM) based approach to evaluate inverter set points at different locations in a network while maintaining network reliability. The results show that this method provides superior convergence guarantees in comparison with other methods while dealing with mixed integer linear programs (MILP). In this work, we propose a distributed optimization framework based on the ADMM method for scheduling EV charging in a power distribution network. Our frame-

work satisfies two goals – maintaining grid reliability while respecting consumer preferences.

4.3.2 Problem Formulation

Here, we are interested in power consumption trajectory over a finite horizon time window of T intervals from time instant $k = 0$ to $k = T$. We define an interval t as the duration between time instants $k = t - 1$ and $k = t$. Table 4.1 summarizes the index variables and sets used in the section.

Table 4.1: Summary of index variables and sets used.

Symbol	Description
T	Number of intervals in time window
N	Number of non-substation nodes in network
i	Index of node in network
t	Index of time interval
k	Index of time instant
α, β	Limits of squared voltage magnitude
$\mathcal{V}$	Set of all nodes in network
$\mathcal{H}$	Set of all residence nodes in network
$\mathcal{N}$	Set of all non-substation nodes in network

Distribution network model

The power distribution network is a *tree* comprising of $N+1$ nodes collected in the set $\mathcal{V} := \mathcal{N} \cup \{0\}, \mathcal{N} := \{1, 2, \cdots, N\}$. The tree is rooted at substation node $\{0\}$ and consists of primary and secondary distribution lines collected in the edge set. The set $\mathcal{N}$ includes residences, local transformers and auxiliary nodes required to connect the transformers and residences (Meyur, Marathe, et al. 2020; Kersting 2012; Bolognani and Dörfler 2015). Here, we are interested in the variables associated with the set of residence nodes $\mathcal{H} \subset \mathcal{N}$.

The power consumption and squared voltage magnitude at node i and time t are denoted respectively by p_i^t and v_i^t. We respectively stack these variables for $i \in \mathcal{N}$ to corresponding vectors $\mathbf{p}^t$ and $\mathbf{v}^t$ for every time interval t. The non-linear relation between power injections and squared voltages in the network can be simplified to linear expression using the Linearized Distribution Flow (LDF) model (Bolognani and Dörfler 2015).

$$\mathbf{v}^t = -2\mathbf{R}\mathbf{p}^t + \mathbf{1} \tag{4.2a}$$

$$\alpha\mathbf{1} \leq \mathbf{v}^t \leq \beta\mathbf{1} \tag{4.2b}$$

Here, $\mathbf{1}$ is a vector of all 1's. The matrix $\mathbf{R}$ is a function of network topology and edge parameters which is only accessible by the network operator. The primary objective of the operator is to maintain the network reliability where the node voltages are within acceptable ANSI C.84 Range A limits (ANSI 2020). In most practical distribution networks, these limits are 0.95 pu and 1.05 pu. The operator ensures that the power consumption at different nodes in the network satisfy (4.2) where α, β denote the squared voltage limits.

Residence Load Models

This section describes load demand of a residence node $i \in \mathcal{H}$. The aggregate power consumption for the time interval t is given by $p_i^t \in \mathbb{R}$. This load comprises of an uncontrollable base load demand denoted by $p_{i,0}^t$ and the controllable counterpart. Here, we use the synthetically generated residential load demand data described in Thorve et al. (2018) for the uncontrollable base load demand. We consider residence owned EV charging stations as the only controllable load which is denoted by $p_{i,\mathrm{EV}}^t$. The power consumption at node i and

over time interval t can be expressed as

$$p_i^t = p_{i,0}^t + p_{i,\mathrm{EV}}^t \quad \forall t = 1, 2, \cdots, T \tag{4.3}$$

EV Charging Model

We assume that the EV charging unit is responsible to charge only a single EV owned by the customer. Let the charge capacity of the EV be $Q_{i,\mathrm{EV}}$ and the power rating of the EV charging unit be $P_{i,\mathrm{EV}}$. The state of charge (SOC) evolves over the time interval t from $s_{i,\mathrm{EV}}^{t-1}$ to $s_{i,\mathrm{EV}}^t$ following (4.4b). Further, the constraint (4.4c) limits the SOC to suitable lower and upper bounds. The scheduling problem aims to find out the optimal time intervals when the EV can be charged while maintaining a secure power grid. Let $z_{i,\mathrm{EV}}^t \in \{0,1\}$ be a binary variable which takes the value 1 if the EV is charged at time interval t and 0 otherwise.

Further, the EV is available for charging only at particular time intervals denoted by the closed interval $\mathcal{T}_{i,\mathrm{EV}} := \left[t_{i,\mathrm{start}}, t_{i,\mathrm{end}}\right]$. Let the initial SOC be $s_{i,\mathrm{init}}$ and it is expected that by the end of the interval, the SOC needs to be at least $s_{i,\mathrm{final}}$. Note that we consider a very simple case where the EV is available to be charged within a single continuous time interval $\mathcal{T}_{i,\mathrm{EV}}$ in the time window of T intervals. We also assume that the EV is not used in this interval.

$$p_{i,\mathrm{EV}}^t = z_{i,\mathrm{EV}}^t P_{i,\mathrm{EV}} \qquad \forall t = 1, 2, \cdots, T \tag{4.4a}$$

$$s_{i,\mathrm{EV}}^t = s_{i,\mathrm{EV}}^{t-1} + \frac{p_{i,\mathrm{EV}}^t}{Q_{i,\mathrm{EV}}} \qquad \forall t = 1, 2, \cdots, T \tag{4.4b}$$

$$0 \leq s_{i,\mathrm{EV}}^k \leq 1 \qquad \forall k = 0, 1, \cdots, T \tag{4.4c}$$

$$z_{i,\mathrm{EV}}^t = 0 \qquad \forall t \notin \mathcal{T}_{i,\mathrm{EV}} \tag{4.4d}$$

$$s_{i,\mathrm{EV}}^{t_{i,\mathrm{start}}} = s_{i,\mathrm{init}}, \quad s_{i,\mathrm{EV}}^{t_{i,\mathrm{end}}} \geq s_{i,\mathrm{final}} \tag{4.4e}$$

In realistic scenarios, these intervals of charging are discontinuous and usage of EV would result in different SOC at different time intervals.

4.3.3 Optimization problem

Each residence aims to compute the optimal power usage trajectory of its EV charging unit over a finite horizon time window of length T denoted by $\{p^t_{i,\mathrm{EV}}\}^T_{t=1}$. Given the hourly rate of electricity c^t for each time interval in the time window, the optimization problem for each residence involves minimizing the total cost of consumption given the EV constraints (4.4). This results in the MILP (4.5).

$$\begin{aligned}
\min \quad & \sum_{t=1}^{T} c^t p^t_i && && && \text{(4.5a)}\\
\text{over} \quad & p^t_i && \forall t && && \text{(4.5b)}\\
\text{s.t.} \quad & (4.3), (4.4) && && \forall t && \text{(4.5c)}
\end{aligned}$$

At the same time, the network operator needs to ensure node voltages are within acceptable limits to maintain a reliable system. Additionally the operator might aim to optimize other aspects such as minimize losses or reduce voltage deviation (Dall'Anese et al. 2014). This can be expressed by $C(\mathbf{p}^t)$ which is a function of power usage of all residences at interval t. Here, we do not consider any particular objective of the network operator and treat $C(\mathbf{p}^t) = 0$.

$$\min \quad \sum_{t=1}^{T} C\left(\mathbf{p}^t\right) + \sum_{i\in\mathcal{H}} \sum_{t=1}^{T} c^t p_i^t \tag{4.6a}$$

$$\text{over} \quad p_{i,\text{EV}}^t \qquad \forall t \ \forall i \in \mathcal{H} \tag{4.6b}$$

$$\text{s.t.} \quad (4.2) \qquad \forall t \tag{4.6c}$$

$$(4.3), (4.4) \qquad \forall t \ \ \forall i \in \mathcal{H} \tag{4.6d}$$

$$p_i^t = 0 \qquad \forall t \ \ \forall i \notin \mathcal{H} \tag{4.6e}$$

The above equation (Eq 4.6) defines the *Reliability-aware EV Charge Scheduling* (REVS) problem which satisfies consumer preferences as well as ensures network reliability.

4.3.4 Proposed methodology

The REVS problem in (4.6) is an MILP with binary variables arising from the on/off status of the EV charging unit. This problem can be solved from a central location (such as the operator) if the load information of residences are known. However, this is not always the case due to privacy concern associated with sharing personal data of consumers. Similarly, the network topology and parameters are considered proprietary information and cannot be shared with the consumers. However, limited information exchange such as total power consumption can be done without violating privacy concerns using the current smart grid infrastructure. In this section, we propose an iterative method based on the ADMM technique to reach the optimal solution for the REVS problem.

To this end, we separate the problem for the network operator and individual residences. Each residence i aims to compute the optimal power usage trajectory $\{p_i^t\}_{t=1}^T$ over the time window given the EV charging constraints. The network operator computes consumption

trajectories $\{\tilde{p}_i^t\}$ for all nodes (in the vector form $\tilde{\mathbf{p}}^t$) such that the network reliability constraints are satisfied. Additionally, we add constraint (4.7f) to force these trajectories to match each other. Therefore, we get (4.7) as the alternate version of the REVS problem.

$$\min \quad \sum_{i\in\mathcal{H}}\sum_{t=1}^{T} c^t p_i^t \tag{4.7a}$$

$$\text{over} \quad p_i^t, \tilde{p}_i^t \qquad \forall t \;\; \forall i \tag{4.7b}$$

$$\text{s.t.} \quad (4.2) \qquad \forall t \tag{4.7c}$$

$$(4.3), (4.4) \qquad \forall t \;\; \forall i \in \mathcal{H} \tag{4.7d}$$

$$p_i^t = 0 = \tilde{p}_i^t \qquad \forall t \;\; \forall i \notin \mathcal{H} \tag{4.7e}$$

$$\tilde{p}_i^t = p_i^t \qquad \forall t \;\; \forall i \in \mathcal{H} \tag{4.7f}$$

Table 4.2: Summary of variables in optimization problem

Symbol	Description
v_i^t	Voltage at node i for interval t
p_i^t	Power consumed at node i for interval t
$\tilde{p}_i^t$	Power consumption computed by operator
$p_{i,0}^t$	Power consumed by fixed load
$p_{i,\text{EV}}^t$	Power consumed by EV charging unit
$z_{i,\text{EV}}^t$	On/Off status of EV charging unit
$s_{i,\text{EV}}^k$	SOC of EV at time instant k
γ_i^t	Dual variable corresponding to (4.7f)
$\mathbf{v}^t$	Vector of v_i^t for all nodes $i \in \mathcal{N}$
$\mathbf{p}^t$	Vector of p_i^t for all nodes $i \in \mathcal{N}$
$\tilde{\mathbf{p}}^t$	Vector of $\tilde{p}_i^t$ for all nodes $i \in \mathcal{N}$

Now, we use the conventional ADMM steps to iteratively update the optimization variables for the operator and residences. Let $\tilde{\mathscr{P}}[l] := \{\tilde{\mathbf{p}}^t[l]\}_{t=1}^T$ denote the optimal trajectories for all nodes computed by the operator for iteration l. Similarly, let $\mathscr{P}_i[l] := \{p_i^t[l]\}_{t=1}^T$ denote the optimal power usage trajectory computed by residence i. We abuse the notation

$\{\mathscr{P}_i[l]\}$ to denote the optimal trajectories $\{p_i^t[l]\}_{t=1}^T$ computed by all residences $i \in \mathcal{H}$ individually. The two steps of iteration are listed below. Note that the first step is carried out simultaneously for all residences and network operator. Fig. 4.5 illustrates the proposed message passing based distributed framework.

S1a. At the operator side, we update the operator estimated power consumption $\tilde{p}_i^t$ for all residences using (4.8).

$$\tilde{\mathscr{P}}[l+1] := \arg\min F(\tilde{\mathscr{P}}[l], \{\mathscr{P}_i[l]\}) \tag{4.8a}$$

$$\text{s.to.} \quad \alpha \leq 1 - 2\sum_{j=1}^{N} R_{ij}\tilde{p}_j^t \leq \beta \quad \forall t \ \ \forall i \tag{4.8b}$$

where the function $F(\tilde{\mathscr{P}}[l], \{\mathscr{P}_i[l]\})$ is defined as

$$\begin{aligned} F(\tilde{\mathscr{P}}[l], \{\mathscr{P}_i[l]\}) := & \sum_{i\in\mathcal{H}}\sum_{t=1}^{T} \frac{\kappa}{2}\left(\tilde{p}_i^t\right)^2 \\ & + \sum_{i\in\mathcal{H}}\sum_{t=1}^{T} \tilde{p}_i^t\left(\gamma_i^t[l] - \frac{\kappa}{2}\tilde{p}_i^t[l] - \frac{\kappa}{2}p_i^t[l]\right) \end{aligned} \tag{4.9}$$

S1b. For residence i, we update using (4.10).

$$\mathscr{P}_i[l+1] := \arg\min \sum_{t=1}^{T} c^t p_i^t + F_i(\tilde{p}_i^t[l], p_i^t[l]) \tag{4.10a}$$

$$\text{s.to.} \quad (4.3) - (4.4) \tag{4.10b}$$

where the function $F_i\left(\tilde{p}_i^t[l], p_i^t[l]\right)$ is defined as

$$\begin{aligned} F_i\left(\tilde{p}_i^t[l], p_i^t[l]\right) = & \sum_{t=1}^{T} \frac{\kappa}{2}\left(p_i^t\right)^2 \\ & - \sum_{t=1}^{T} p_i^t\left(\gamma_i^t[l] + \frac{\kappa}{2}\tilde{p}_i^t[l] + \frac{\kappa}{2}p_i^t[l]\right) \end{aligned} \tag{4.11}$$

S2. At the operator and residence sides, the dual variable is updated.

$$\gamma_i^t[l+1] = \gamma_i^t[l] + \frac{\kappa}{2}\left(\tilde{p}_i^t[l+1] - p_i^t[l+1]\right) \tag{4.12}$$

The resulting decentralized procedure involves a two-way message exchange of the iterates $\{\tilde{\mathbf{p}}^t[l]\}_{t=1}^T$ and $\{\mathbf{p}^t[l]\}_{t=1}^T$ between the network operator and residential consumers. At an iteration $l > 0$, the network operator updates the power trajectories based on (4.8) whose objective includes a regularization term $F(\tilde{\mathscr{P}}[l], \{\mathscr{P}_i[l]\})$. This term enforces consensus with the power usage trajectories computed at the residences. The constraints ensure the reliability aspects of the network. Note that (4.8) is a QP because of the quadratic regularization term. The operator relays to each residential consumer i a copy of the iterate value $\{\tilde{p}_i^t[l+1]\}_{t=1}^T$. At the same time, the consumer optimal trajectories are updated using (4.10) and copy of the iterate value $\{p_i^t[l+1]\}_{t=1}^T$ is sent to the operator. We note that (4.10) is a MIQP because of the quadratic regularization term ensuring consensus with the operator objective and binary constraints for the EV charging unit. Thus, the REVS problem which was originally an MILP is converted to a QP for the operator and MIQPs for individual residences using the proposed ADMM based framework. Once the updated local iterates are exchanged, the operator and residences update the local dual variables using (4.12).

The centralized approach to solve the MILP guarantees convergence to the global optimum solution. However, the concern of sharing private consumer information with the network operator hinders the approach. The proposed ADMM based distributed framework avoids sharing of private and proprietary information and only uses exchange of power consumption data. The approach converts the problem into a QP for the operator and MIQPs for each residence. However, the size of each problem is significantly smaller than the original MILP. The convergence of the algorithm to the optimal solution of (4.6) is formally stated

next.

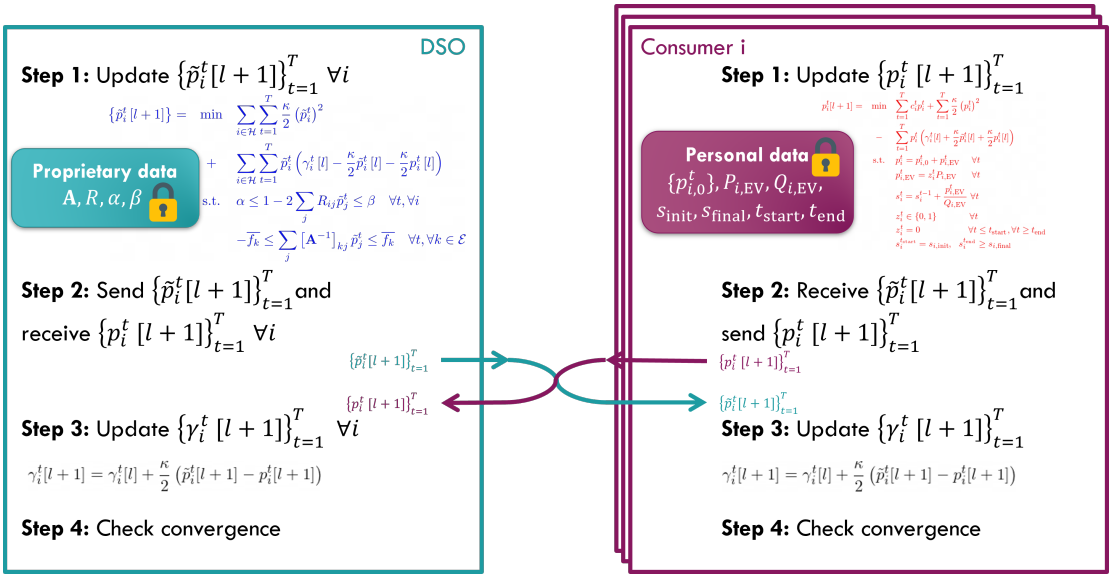

Figure 4.5: Message exchange aided proposed distributed framework to schedule residential EV charging.

Proposition 4.1. *The iterates* $\{\tilde{\mathbf{p}}^t[l]\}_{t=1}^T$ *and* $\{\mathbf{p}^t[l]\}_{t=1}^T$ *produced by* [**S1**] − [**S2**] *are convergent, for any* $\kappa > 0$. *Further,*

$$\lim_{l\to\infty}\{\tilde{\mathbf{p}}^t[l]\}_{t=1}^T = \lim_{l\to\infty}\{\mathbf{p}^t[l]\}_{t=1}^T = \{\mathbf{p}_{\text{opt}}^t\}_{t=1}^T$$

where $\mathbf{p}_{\text{opt}}^t$ *denotes the optimal power usage trajectory.*

ADMM has been proved to converge to the optimal solution for convex problems and for specific non-convex problems involving binary constraints (Boyd et al. 2011). Therefore, we can guarantee that the proposed framework converges to the optimal solution in the exact sense.

4.3.5 Experiments

The experiments are conducted in order to study the effects of EV adoptions at different levels (30%, 60%, 90%). We compare effects of two optimization scenarios (individual vs. distributed) on EV scheduling behavior in different communities. Under the *individual optimization* scenario, customers charge their EVs based on individual preferences without considering the impact on network reliability. The optimal schedule is obtained by solving (4.5) for each EV adopter. With *distributed optimization*, the customers coordinate with the network operator to achieve an EV charging schedule where their personal preferences are accommodated as well as the network reliability is maintained. The optimal schedule is obtained by iteratively solving Equations (4.8), (4.10), and (4.12) until convergence.

Particularly, we aim to compare the reliability of the network when these two methods are used to schedule residential EV charging for varied levels of EV adoption. Note that network reliability is the ability to operate with edge power flows within the line capacities and node voltage within the bandwidth ($0.95 - 1.05$ p.u) ANSI 2020. Hence, these two measures – node voltage and edge power flow are used to quantify the impact of network reliability at different levels (30%, 60%, 90%) of EV adoptions in multiple communities in the distribution network.

A small area of Montgomery county in Virginia is considered as the region of interest for our study. Household level synthetic hourly electricity consumption profiles are used. These timeseries are created using several population surveys, statistical models, and physics based models of household devices and validated using real data Thorve et al. 2018. The data also has household level demographics and spatial attributes. Synthetically generated distribution networks created using electrical engineering concepts and resembling actual networks are used for the purpose of our analysis Meyur, Marathe, et al. 2020. Hourly

electricity rate (in $/kWhr) is known to the residential customers (Table 4.3). Time of use (TOU) hourly electricity rate provided by the off-peak plan of a utility company serving the particular geographical region Dominion Energy 2021 are used.

Assumptions. All EVs have a uniform charge capacity of 20kWhr and are available to be charged between 4:00p.m. and 5:00a.m. The initial state of charge is assumed to be 20% and the EVs are required to be charged to at least 90%. Households are randomly selected as EV adopters in the network. All adopters have necessary provisions to charge their EVs at their residential premises. We consider a uniform power rating of 4.8kW for all residential EV chargers.

4.3.6 Results

Figure 4.6 describes edge power flow as a percentage of line capacities and node voltages in 'Com-A' community in the network when EV adoption has reached 90%. Figure 4.6 (top) shows that the edge flow (line rating capacity) levels in the network are well within limits even when 90% of residences have adopted EV. However, the same cannot be observed for node voltages in the network. Figure 4.6 (bottom) shows that the node voltages at several residences are outside the acceptable limits of 0.95-1.05 p.u. We notice that maximum number of node voltage violations occur in the period where the hourly electricity rate (in $/kWhr) is minimum.

Table 4.3: Hourly electricity rate for the experiments.

Time interval (HH:MM)	00:00--05:00	05:00--15:00	15:00--18:00	18:00--00:00
Cost ($/kWhr)	0.07866	0.09511	0.21436	0.09511

We further explore effects of node voltage violation, at different adoption levels in two

Figure 4.6: Comparison of line loading level (edge flows) and node voltages for residential EV adoption of 90% in 'Com-A' of network. The high EV adoption does not significantly affect the line loading level. However, node voltages at multiple residences in the network are outside the acceptable voltage limits of $0.95 - 1.05$ p.u.

communities ('Com-A', 'Com-B') of residences in the network. Figure 4.7 shows the selected communities in the network and node voltage violation at different adoption levels. The results are obtained after performing optimizations under two scenarios : individual and distributed. We focus our attention only on the time intervals where the hourly electricity rate (in \$/kWhr) is minimum (i.e. time windows of maximum node violations). The node voltage violation is divided into 3 ranges : less than 0.92 p.u., between $0.92 - 0.95$ p.u. and between $0.95 - 0.98$ p.u. Though the last voltage range is not considered as voltage violation by the practised ANSI standard ANSI 2020, it can be considered as a sign of reduced network reliability. The clustered bar chart shows significantly higher number of residences with voltage violation under individual optimization scenario as compared to the distributed optimization scenario. The number of residences with voltage less than 0.95 p.u. is close to zero at all considered time intervals. The observations are similar for both communities in the network. This shows that, if the proposed distributed framework is used to schedule EV charging units, the network operator is able to dispatch the power without compromising on system reliability.

Figure 4.7 shows that under the individual optimization approach the number of residences with undervoltage during the cheap electricity hours increases with an increase in the level of adoption. However, this trend is not consistent in the distributed optimization approach. This is because the later approach also ensures system reliability along with consuming electricity during cheap hourly rates. The distributed framework does this by allocating small amount of EV charging during time intervals where the hourly electricity rates are relatively higher. We also notice that the number of residences experiencing undervoltage issues for the same level of adoption differs significantly when we consider different communities for EV adoption. These differences can be attributed to location and energy usage of adopter households in the network and the resulting voltages at different nodes. The

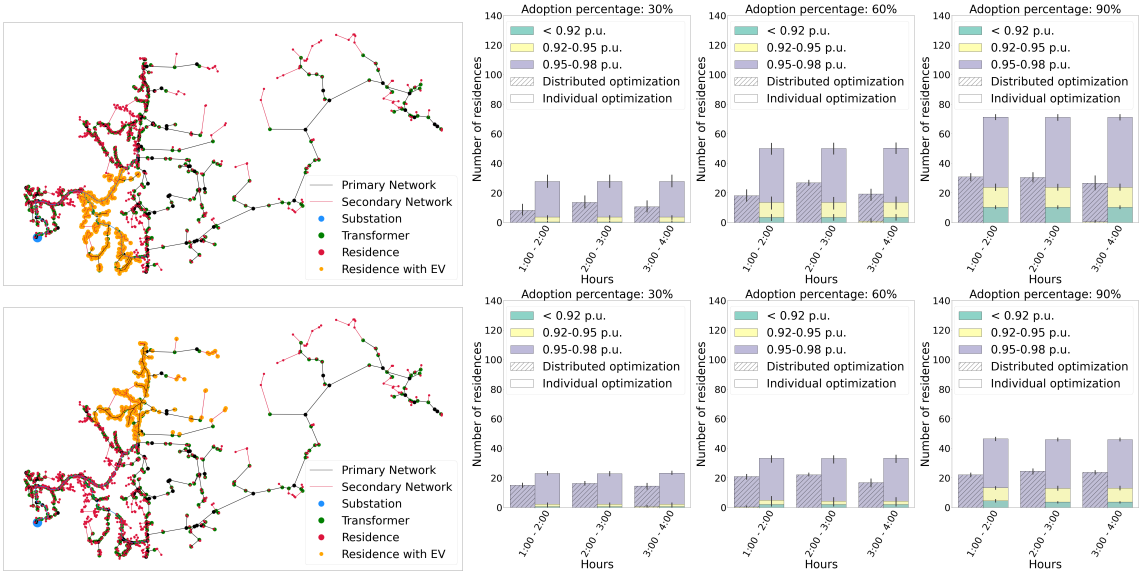

Figure 4.7: Impact of residential EV charging is analyzed for two different residential communities within the same network: 'Com-A'(top) and 'Com-B'(bottom). The orange nodes shown in the network denote the residences in the two communities. The individual optimization leads to undervoltage (less than 0.95 p.u.) at a significant number of residences. This can be avoided by using the proposed distributed optimization method even for higher levels of EV adoption.

error bars on the bar chart (Figure 4.7) show variation in number of residences violating node voltage in each category for multiple random group of EV adopters.

Chapter 5

Framework for Realistic Analysis of Cascading Failures

Publications

1. Rounak Meyur, Anil Vullikanti, Madhav Marathe, Anamitra Pal, Mina Youssef, Virgilio Centeno, and Arun Phadke, "Vulnerability of the power grids to targeted physical attacks" in Proceedings of the National Academy of Sciences, 2022 (in preparation for submission).

2. Rounak Meyur, Anil Vullikanti, Madhav Marathe, Anamitra Pal, Mina Youssef, Virgilio Centeno, and Arun Phadke, "Cascading Effects of Targeted Attacks on the Power Grid", in Complex Networks and Their Applications VII (pp. 155–167). Dec 2018, Cambridge, UK.

3. Rounak Meyur, "A Bayesian Attack Tree Based Approach to Assess Cyber-Physical Security of Power System", in IEEE Texas Power and Energy Conference (TPEC) (pp. 1–6). Feb 2020, College Station, Texas, USA.

5.1 Introduction

Critical infrastructures are defined as *those physical and cyber-based systems that are essential to the minimum operations of the economy and the government* (The White House 1998; Republican Policy Committee 2021). Since they provide crucial support for the delivery of basic services to almost all segments of society, they form the backbone of any nation's economy. As one of the most complex, large-scale networked systems, electric power system has become increasingly automated in the past few decades. However, the increased automation has introduced new vulnerabilities to equipment failures, human errors (NERC 2021; NERC 2016; NERC 2015; NERC 2014; NERC 2013), weather and other natural disasters (Dumas, Kc, and Cunliff 2019; Weiss and Weiss 2019), and physical and cyber-attacks (Republican Policy Committee 2021; B. M. Amin et al. 2020). The ever-increasing system scale and the strong reliance on automatic devices increase the likelihood of turning a local disturbance into a large-scale cascading failure (Y. Liu et al. 2021; Guo et al. 2021; Schäfer et al. 2018; Meyur, Vullikanti, et al. 2018; C. Barrett et al. 2012; Panzieri and Setola 2008). This kind of wide-area failure may have a catastrophic impact on the whole society. Reports of recent major power system blackouts (King, Rhodes, and Zarnikau 2021; Busby et al. 2021; Beers 2021; Feuerstein 2022; Romero 2012; Pourbeik, Kundur, and Taylor 2006; Force 2004; Kosterev, Taylor, and Mittelstadt 1999) have shown how several events ranging from minor equipment failure and operator errors to severe weather events (such as forest fires, hurricanes and winter storms) have triggered widespread system wide power disruption affecting millions of customers. This necessitates the development of a framework which would assess the vulnerability of the power grid subjected to any of these events, and thereby allowing energy policy makers to identify critical components in the grid and subsequently allocate budgets to harden them.

The inclusion of information and communication technology (ICT) in the supervisory control and data acquisition (SCADA) system has increased the resiliency and facilitated self healing capabilities of the power grid (C. Liu et al. 2012). This has been made possible through the large connected communication network with several remote access points enabling coordinated monitoring and control functions on the power grid (S. M. Amin and Wollenberg 2005). However, this has exposed the system to numerous possibilities of cyber threat increasing the risk of a combined catastrophic failure of the power grid along with the communication network (CIGRE 2007). Therefore, the smart power grid consisting of the traditional power system with the intertwined ICT elements is identified as a critical interdependent infrastructure where a failure in either network can result in severe impact on the combined system (W. Wang et al. 2018).

The usage of standardized communication protocol in the SCADA system has opened up several vulnerabilities in the commonly used protocols like distributed network protocol (DNP) and IEC 61850 (Mackiewicz 2006). These may be known and zero-day type which can be exploited by an adversary to gain unauthorized access to control assets in the SCADA system (L. Wang et al. 2014). For example in the cyber physical system depicted in Fig. 5.1 the lower part represents the physical power system. The upper part denotes the hierarchical cyber system or the SCADA network associated with the power grid. The substations communicate with the regional control centers through Ethernet communication protocol, the control centers interchange information using the inter control center protocol (ICCP) and also communicate with the transmission operator (TO) through Ethernet routers. In such a setup, an intruder can exploit the vulnerabilities of the control center or substation LAN to gain administrator privilege in one of the human machine interfaces (HMIs) (Y. Zhang, L. Wang, et al. 2015). By obtaining access, the control commands might be manipulated to operate the circuit breakers in the physical power system resulting

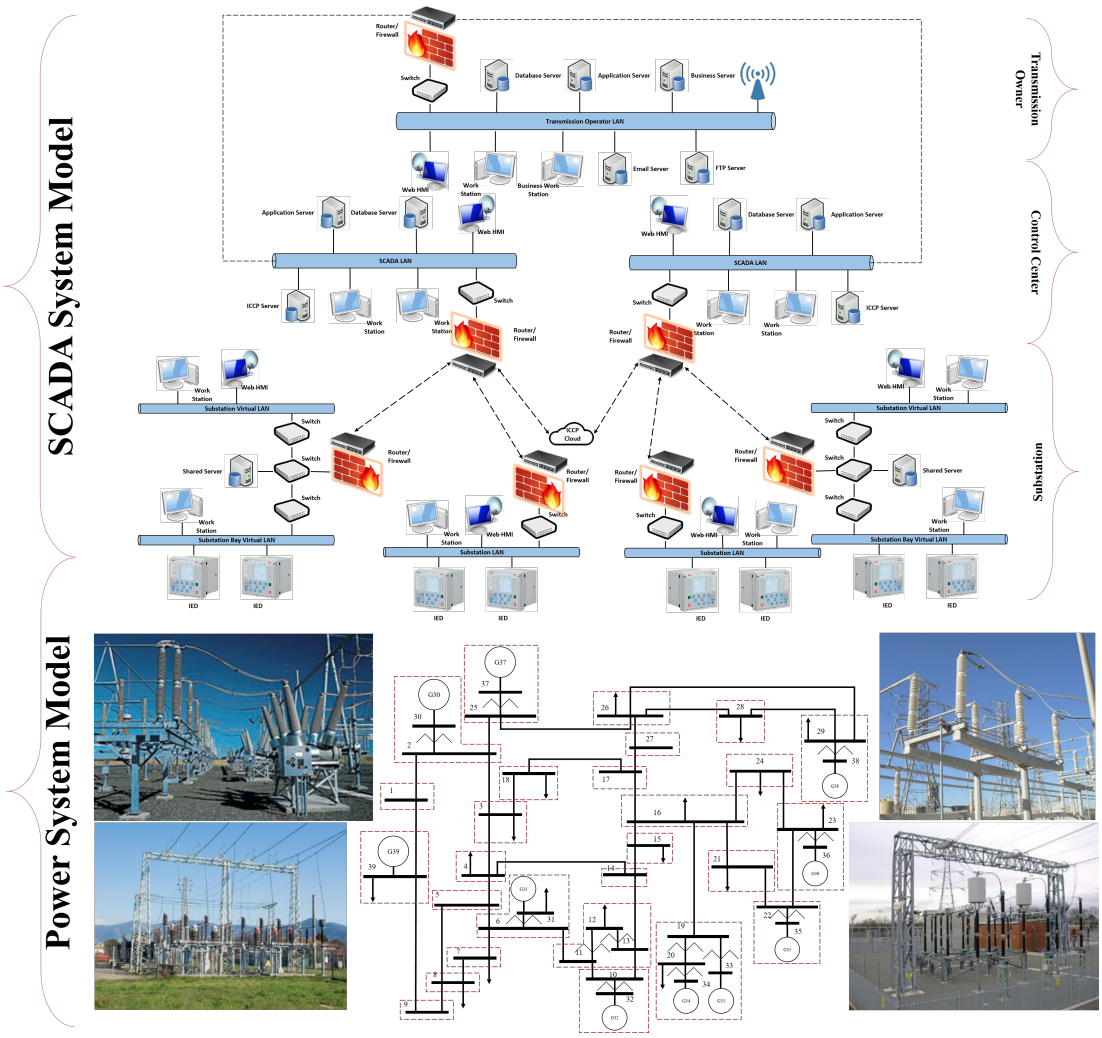

Figure 5.1: Cyber-physical model of a typical power system.

in instability in the grid from load-generation imbalance (Stamp, McIntyre, and Ricardson 2009). Often such attack can cause cascading events in the system leading to widespread blackout as in the case of Stuxnet malware attack in Ukraine in 2015 Pultarova 2016; Gary and Prananto 2017. This is due to the fact that the power system is operated with security analysis performed for at most 2 contingencies. A planned cyber attack can lead to multiple contingencies at the same time exacerbating the disturbance in the grid or lead to unauthorized access to SCADA system giving way to man-in-the-middle attacks (R. Liu et al. 2015).

5.1.1 Related works

Statistical analysis of more than 400 blackouts in USA from 1984 to 1999 indicates that a large blackout, though rare, is more likely to occur than expected (heavy tails of a power law distribution) (Carreras, Newman, et al. 2000). Therefore, large blackouts require more attention not only due to their higher probability of occurrence, but also due to the enormous societal damage caused by such events. Following this observation, several works (Guo et al. 2021; Dobson, Carreras, et al. 2001; Carreras, Lynch, et al. 2002b; Carreras, Lynch, et al. 2002a; Dobson, J. Chen, et al. 2002; J. Chen, Thorp, and Dobson 2005; Y. Zhang and Yaan 2016; Pahwa, Scoglio, and Scala 2014; Soltan, Mazauric, and Zussman 2014; Bernstein et al. 2012) have proposed multiple failure models to represent the system dynamics leading to a cascading outage. They have studied cascading failures in power grids using quasi-steady state analysis with DC power flow. With any reactive power component being ignored and the assumption of a flat voltage profile, the DC power flow analysis may produce good approximations under some circumstances, e.g., when performing steady-state planning level studies. However, the increased penetration of converter-based generator technologies, loads and transmission devices have contributed to newly evolved dynamic stability behaviors of the power grid (N. Hatziargyriou et al. 2021). Major cascading outages are caused when transient rotor angle stability and voltage stability of the power grid are affected (Pourbeik, Kundur, and Taylor 2006; P. Hines, Cotilla-Sanchez, and Blumsack 2010; P. Hines, Blumsack, et al. 2010; Song et al. 2016). Therefore, a simple cascading failure model based on DC power flow analysis is not a suitable tool to simulate such events. In this chapter, we consider the AC power flow model to accurately simulate the actual operating point in the power system.

Several physics-based models have been used to study cascading failures in power grid networks and interdependent power and communication networks (Buldyrev et al. 2010;

Parshani, Buldyrev, and Havlin 2010; Huang et al. 2011; Brummitt, D'Souza, and Leicht 2012; W. Wang et al. 2018; Z. Wang, G. Chen, et al. 2020; Son et al. 2012; Valdez et al. 2020). The authors have considered the effect of connectivity between layered networks on the cascade probability in each network, and used the sandpile dynamics (Bak, C. Tang, and Wiesenfeld 1988) to represent the cascade tripping of loads in the power grids. These papers are useful in that one can often either obtain analytical results, or carry out large number of simulations to get a detailed understanding of cascade dynamics. The physics based models are simplified models capable of showcasing mechanistic possible behavior of complex network systems, rather than providing precise predictions which requires engineering models with a large number of parameters (Valdez et al. 2020). The models fail to replicate the actual system conditions in a power grid where a node (or bus) trips due to under-voltage or under-frequency and not due to overload. Further, stability of a power system subjected to cascading events is evaluated either from the network structure point of view (evaluating the degree distribution of nodes) (Brummitt, D'Souza, and Leicht 2012; Buldyrev et al. 2010; Huang et al. 2011; W. Wang et al. 2018) or from the convergence of steady-state power flow solution (Dobson, Carreras, et al. 2001; Carreras, Lynch, et al. 2002b; Carreras, Lynch, et al. 2002a; Dobson, J. Chen, et al. 2002; J. Chen, Thorp, and Dobson 2005; Y. Zhang and Yaan 2016; Pahwa, Scoglio, and Scala 2014). Such measures do not necessarily cover all possibilities of grid instability (N. Hatziargyriou et al. 2021), as non-linear mechanisms such as rotor angle stability or voltage collapse are not accurately captured in these methods (P. Hines, Cotilla-Sanchez, and Blumsack 2010). In this work, dynamic transient analysis has been used to assess stability of the power system.

The reports of certain major blackouts (Force 2004; Pourbeik, Kundur, and Taylor 2006) suggest that cascades need not propagate locally due to the complex non-linear nature of the power grid. Furthermore, Kosterev, Taylor, and Mittelstadt (1999) discusses the various

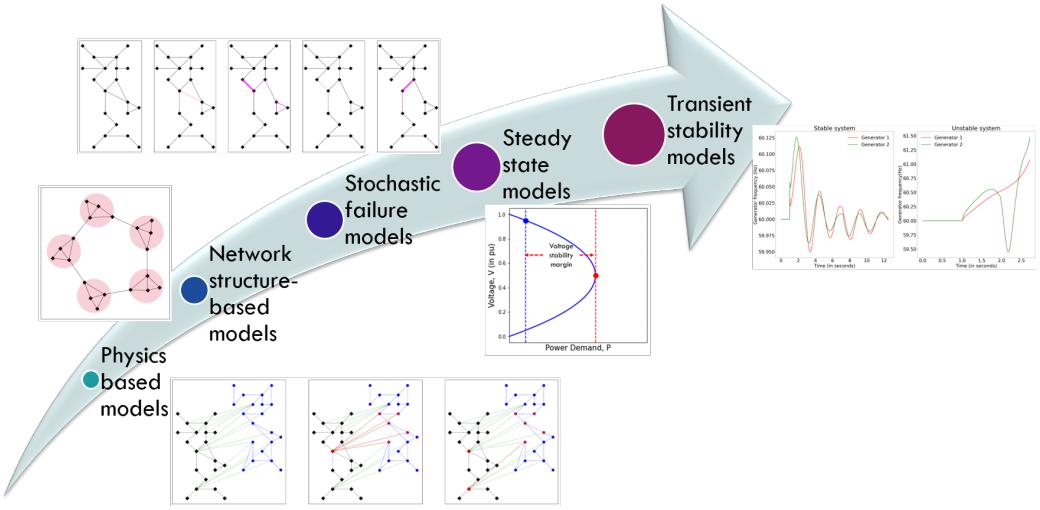

Figure 5.2: Summary of different failure models in literature – physics based models represent cascades in inter-dependent networks where dependent nodes in either network fail simultaneously, network structure-based failure models proposes importance of network motifs, stochastic failure models assign probabilities to edges adjacent to failed nodes/edges and cascade propagates. The steady-state and time domain simulation models depict a more realistic version of cascading outages. The operation of protection systems is not considered in any of these works.

reasons leading to the historic 1996 WSCC outage, the most important being the operation of relays. Based on the NERC data, in more than 70% of the major disturbances, failures in protective relays are found to be a contributing factor (J. Chen, Thorp, and Dobson 2005). Among these failures, a failed protection system that remains dormant in normal operating conditions and becomes exposed when an abnormal condition in the system forms, is the most troublesome to tackle (Tamronglak et al. 1996). Such failures are termed as hidden failures and these are capable of causing widespread cascading failures in the power system network leading to a major blackout (Thorp et al. 1998). This is equivalent to the human immune system where an immune response following immunization might be more damaging than the pathogen it is supposed to protect against (Schrom et al. 2021). It is evident from the above discussion that protection systems play a key role in cascading events. Though most of the papers consider line outages due to overload, the protection system in

the power network respond to measured impedance, voltage and current. The vital contribution of the proposed work is the inclusion of a stochastic model to simulate hidden failures in the power system whose effects surface in the aftermath of a human initiated attack on the network.

An important step in modeling cascading failures is to evaluate the probability of tripping of each component in the power system network. The NERC statistics over the past decade (NERC 2013; NERC 2014; NERC 2015; NERC 2016; NERC 2021) show that relay misoperations due to unnecessary trips are more probable than failure to trip. Such relay failures or faulty settings are often the principal determinant of the occurrence of a hidden failure in the power system. In this work the AC power flow model is used to obtain the system conditions at each instant of simulation and transient stability analysis is performed to assess the stability of the grid. The operation of protection systems for generators, transmission lines and transformers is modeled along with the stochastic occurrence of hidden failures in them. The trip signals of these relays are considered as the sole contributors of node and edge outages in the network.

Another important aspect of studying cascading events in the power system network is the impact of different initiating events. For example, (Brummitt, D'Souza, and Leicht 2012; Buldyrev et al. 2010; Y. Zhang and Yaan 2016) have initiated cascaded failures by targeting random node(s) in the network. On the contrary, cascaded outages triggered by weather events are initiated by targeting geographically correlated nodes (Weiss and Weiss 2019). Given the complex non-linear nature of the power system, the *optimal critical set* problem is worth mentioning (Meyur, Vullikanti, et al. 2018). The goal is to find the optimal set of nodes that results in maximum damage to the network. The results show that a greedy choice of high voltage (500kV) nodes leads to significant impact on the power network, often resulting in an unstable system causing widespread power outages.

In the context of cyber attacks, several models have been proposed for identifying vulnerabilities in the CPS by evaluating the impact resulted from such attacks (W. Wang et al. 2018; R. Liu et al. 2015; Ten, C. Liu, and Manimaran 2008; T. M. Chen, Sanchez-Aarnoutse, and Buford 2011; M. Govindarasu and C. Liu 2012; Xie, Stefanov, and C. Liu 2016; Stefanov, C. Liu, et al. 2013; Stefanov and C. Liu 2012). A very simple statistical model based on graph theory results has been proposed in W. Wang et al. (2018) where neither the SCADA nor the power network is accurately modeled. A Petri-Net based cyber model consisting of firewalls and passwords has been proposed in Ten, C. Liu, and Manimaran (2008). Though the model was capable of replicating the operation of firewalls precisely, the probability of intrusion through a firewall was randomly selected irrespective of the hierarchy at which it is present. An improved usage of hierarchical Petri Nets is seen in T. M. Chen, Sanchez-Aarnoutse, and Buford (2011) where the vulnerabilities of smart meters are modeled. However, the model was not developed to be applicable in the hierarchy of the SCADA system for the power grid. A comprehensive CPS model has been used in Xie, Stefanov, and C. Liu (2016), Stefanov, C. Liu, et al. (2013), and Stefanov and C. Liu (2012) where every element of the communication system is modeled using queues. The attack efficiency of a possible cyber threat is evaluated as the time required to send a packet of data from the source vulnerability to the target. However, the time of arrival of the data packet is selected randomly based on the processing rate. Therefore, the relative difficulty in exploiting a particular vulnerability is not considered; nor the skill level of the intruder is used to determine the time to compromise a target vulnerability. To this end, a statistical model is proposed in McQueen et al. (2006) where the skill level of the intruder and the difficulty of exploiting a vulnerability is considered to determine the time to compromise a given vulnerability.

5.1.2 Contributions

In this work, we build a framework to analyze cascading failures on a given power network that has been subjected to physical attacks on multiple nodes. We compare the extent of cascading outages in case of a detonation of a large tactical device in Washington DC, and a strategic targeted physical attack performed simultaneously on different power system substations located far apart from each other. The following are the main conclusions of our analysis: (a) A time-domain, AC analysis is essential in capturing the full effects of a cascading event on the electric power grid. (b) A strategic targeted attack on few critical substations (as few as 2) is capable of leading the power system to collapse within a few seconds. (c) In context of a strategic attack on a critical target set, the addition of target nodes to an existing target set does not necessarily increase the impact on the power grid. (d) The load-generation balance plays a key role in the extent of cascading outages.

We also propose a Bayesian attack tree based CPS model to model cyber attacks on the power grid SCADA system. Such a model allows us to identify vulnerabilities in the SCADA architectures. This model can be augmented to the cascading failure framework to evaluate the physical impact of a cyber-attack on the power grid.

5.2 Protection Systems in Power Grid

The vast transmission network is exposed to mercy of the nature making it highly susceptible to faults. These might have varied degree of impact on the power system ranging from sagging of lines to damage of important assets in the grid Horowitz and Phadke 1995. In order to protect the equipment from such severe phenomena, the power system is equipped with protection elements. The purpose of protection system is two fold: (*i*) detect a fault

and (*ii*) isolate the faulted section from the healthy part of the power grid. The first task is performed by a relay where a fault is detected based on an algorithm. The second action is carried out by a circuit breaker (CB), which receives a trip signal from the relay as soon as a fault is detected. In this section the operation of different relays employed for protection of transmission lines, transformers and generators has been discussed. The relays widely used for transmission line protection are directional overcurrent relays, mho distance protection relays, and carrier communication based directional comparison blocking relays. The relays employed for transformer protection are percentage differential relays. For generator protection, a large number of relays are involved; in this study, we are interested in the voltage based generator protection and out-of-step protection.

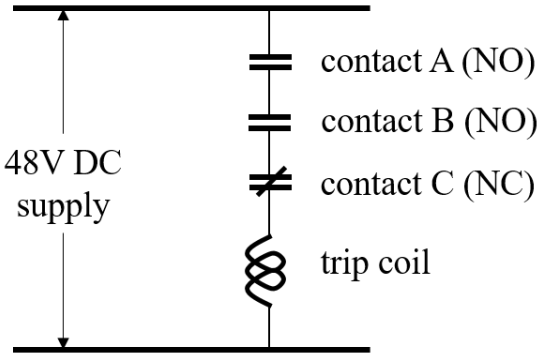

Figure 5.3: DC schematic of a typical protection relay. The trip coil is connected to 48V DC source through two NO contacts (contact A, contact B) and an NC contact (contact C). The trip signal is issued to the circuit breaker if all contacts are closed and trip coil is energized by DC supply.

In this section, we discuss mode of operation of these protective relays. Current transformers (CTs) and potential transformers (PTs) are deployed in the power network to measure current and voltage respectively. The relays use these measurements to detect faults in the power grid and issue a trip signal to the associated CB. In our discussion, we focus our attention to the DC schematic of each relay like the one showed in Fig. 5.3. This consists of a DC circuit with one or more contacts and connects a $48V$ DC source to the coil which

issues the trip signal to the CB. The contacts are either normally closed (NC) or normally open (NO) based on the type of relay and its mode of operation. Each NO contact is closed (or, NC contact is opened) when certain conditions are satisfied. Some examples of such conditions include but are not limited to CT measurement exceeding a threshold value, CT measurement in a definite direction, PT measurement below a threshold value. The trip signal is issued when all contacts are closed and the trip coil is energized by the DC source. A mechanical failure in one of these contacts might result in an NC contact being open or an NO contact being closed. In case of relays with more than one contact, an above-mentioned mechanical failure might remain *hidden* as long as all contacts are closed and a trip signal is issued.

For example, we consider the relay showed in Fig. 5.3. For the relay to issue a trip signal to the associated CB, NO contacts A and B need to be closed so that the trip coil is energized by the DC supply. This requires the conditions for contacts A and B to be closed to be satisfied simultaneously. A mechanical failure in contact B results in the NO contact to be closed. This does not immediately result in a trip signal (since contact A is open) and therefore such a failure remains hidden. However, if the condition to close contact A is satisfied during an event in the power grid, a trip signal is issued instantaneously irrespective of the condition to close contact B being satisfied or not. These are detrimental to the power grid operation since they cause unwanted cascaded trips and eventual system-wide blackout.

5.2.1 Directional overcurrent protection scheme

The most widely used non-pilot protection system for transmission lines is the directional overcurrent relays (DOCRs) which detects faults in a particular direction. Fig 5.4 shows the

schematic of a directional overcurrent protection scheme. Each end of the transmission line is provided with an overcurrent element (L_{AB},L_{BA}) and a directional element (D_{AB},D_{BA}). The overcurrent element operates if magnitude of the current measured at that end exceeds the pickup value. The directional element determines the direction in which the fault is present from the direction of measured current.

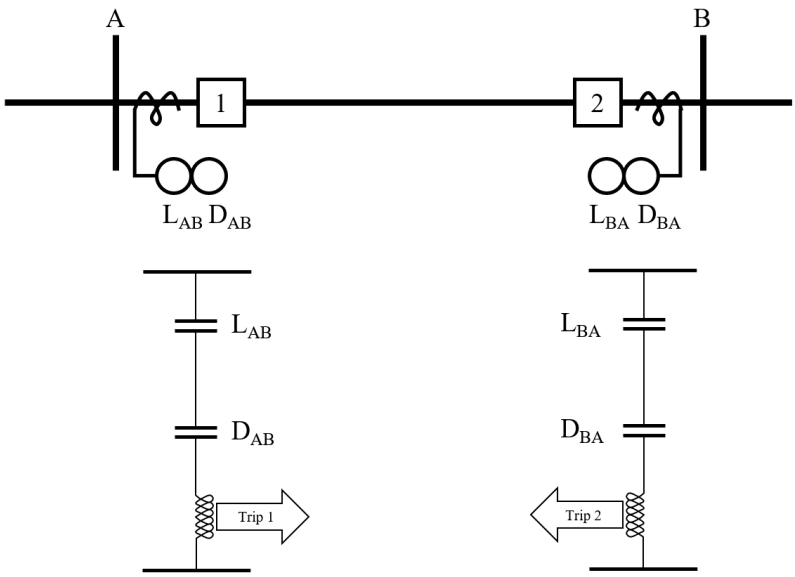

Figure 5.4: Schematic diagram of directional overcurrent protection scheme.

5.2.2 Hidden failures in directional overcurrent relays

Consider a mechanical failure in the directional contact where the contacts are permanently closed. This failure remains hidden in normal operating condition since the relay does not issue a trip signal unless the overcurrent element operates. However, in case of a disturbance, if the load pickup is exceeded, the overcurrent element operates and the relay issues a trip signal irrespective of the current direction. Fig. 5.4 shows the operating region of a directional overcurrent relay in the absence an presence of a hidden failure in the

directional contact. It is to be noted that the directional element responds to the phase angle between the measured voltage and current.

Let the current measured at the ends A and B in Fig 5.4 be I_A and I_B respectively. The voltage and current phase angles at end A are $\phi_{v,A}$ and $\phi_{i,A}$ respectively and $\phi_{v,B}$ and $\phi_{i,B}$ at end B. Let the load pickup magnitudes for the overcurrent elements at A and B be respectively given by I_A^{pick} and I_B^{pick}. The condition for directional relay to generate a trip signal is given by

$$\begin{aligned} \text{Breaker at A trips:} \quad & |I_A| > I_A^{pick}, \quad 0 \leq |\phi_{v,A} - \phi_{i,A}| \leq \frac{\pi}{2} \\ \text{Breaker at B trips:} \quad & |I_B| > I_B^{pick} \quad 0 \leq |\phi_{v,B} - \phi_{i,B}| \leq \frac{\pi}{2} \end{aligned} \tag{5.1}$$

The DOCR issues a trip signal if the current magnitude is above the load pickup value and phase angle between measured current and voltage is less than 90 degrees. If the directional contact is permanently damaged due to a hidden failure, the relay will trip for all phase difference between voltage and current, as soon as the current magnitude exceeds the load pickup value.

$$\begin{aligned} \text{Breaker at A trips:} \quad & |I_A| > I_A^{pick} \text{ and failure in } D_{AB} \\ \text{Breaker at B trips:} \quad & |I_B| > I_B^{pick} \text{ and failure in } D_{BA} \end{aligned} \tag{5.2}$$

5.2.3 Transmission line distance protection

The second type of transmission line relays is the three zone mho distance protection scheme. Protection zones play an important role in this over-reaching distance relay operation Horowitz and Phadke 1995. Zone-1 is the primary protection scheme while Zone-2 and Zone-3 are back-up protection schemes. Zone-1 provides high speed protection to ap-

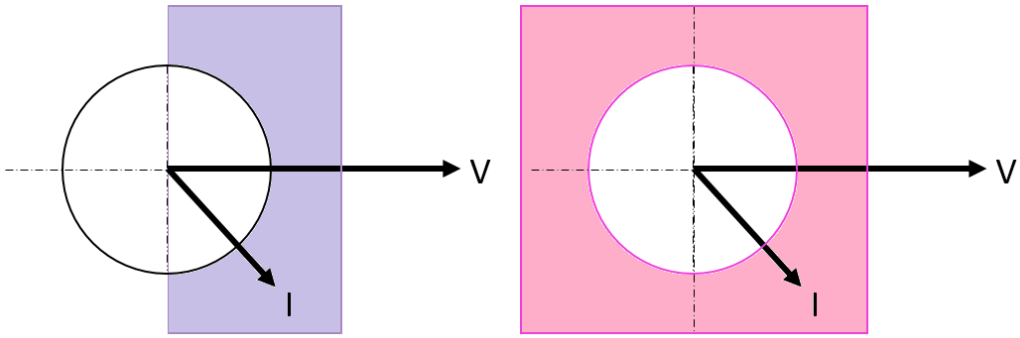

Figure 5.5: Plots showing operating region of DOCR in the absence (left) and presence (right) of hidden failure in the directional contact. The arrowheads show the phasor plots of voltage (V) and current (I) measured at the relay terminals. The purple region indicates the intended region of operation while pink region indicates the region of operation in presence of hidden failure.

proximately 80% of the line which it is designed to protect. It never reaches the bus at the other end of that line. Zone-2 completely covers the protected line and overreaches to a portion of the next line. The primary purpose of Zone-2 is to detect faults in the protected line beyond Zone-1. Zone-2 also provides backup for a failed Zone-1 element, both in the protected line as well as in the next line. Zone-2 is typically set to reach less than the Zone-1 reach of the next line. Zone-3 provides remote back-up protection typically by detecting a fault in the event a remote breaker (which was expected to trip) does not trip. Zone-3 is set to cover 100% or more of the next line beyond the line that is to be protected. Sometimes Zone-3 setting becomes high enough to operate on high load or on power swings. Adequate measures are taken to prevent Zone-3 operation for such situations by using shaped characteristics, load encroachment detection and power swing blocking PSRC 2009.

5.2.4 Mho distance protection scheme

Coordination in time is essential for successful operation of the three zones especially when the power system is under stress. Zone-1 is designed to be instantaneous whereas Zone-

2 and Zone-3 have inherent time delays in their operation. This time delay is called the *fault clearing time*. In case one of the previous zone/s fail to clear the fault, the next zone which detects the fault operates after the fault clearing time is crossed. Thus, in case of an unsuccessful zone operation, the fault persists in the system for a time that is equal to the fault clearing time of the zone which eventually clears the fault. More details about different protection zones can be found in PSRC 2009.

For the simulations done here, the fault clearing time for a Zone-2 operation was set at 0.5 seconds, while that for a Zone-3 operation was set at 1 second (industry standard for the EI). The reach of Zone-1 is considered to be 80% of the line it is designed to protect. Zone-2 has a reach of 150% which covers the entire length of the protected line and 50% of the adjacent line. Zone-3 has a reach of 250% which covers the entire length of the protected line and the adjacent line and covers some portion of the next adjacent line. Fig 5.6 represents the schematic for a three zone mho distance relay which is designed to protect the line A-B. The inherent time delays for the operation of Zone-2 and Zone-3 is implemented through the timer contacts (T_{AB2},T_{BA2} for Zone-2 and T_{AB3},T_{BA3} for Zone-3) and their corresponding timer coils.

The right plot in Fig 5.6 shows the operating zones of the mho distance relay at A designed to protect the line A-B. Let $M_{AB1}, M_{AB2}, M_{AB3}$ indicate operating zones Zone-1, Zone-2 and Zone-3 for the mho relay at A. If the apparent impedance (the impedance measured by the relay) encroaches a particular zone, the zone element contact ($Z_{AB1}, Z_{AB2}, Z_{AB3}$) closes instantaneously. However for Zone-2 or Zone-3, the trip coil is not energized since the timer contact (T_{AB2}, T_{AB3}) remains open. The timer contact closes only after the inherent time delay.

It is evident that the diameter of an operating zone lies on the impedance line A-B-C. Let the complex impedance at the point on an operating zone be $Z\angle\theta$ where θ is the impedance

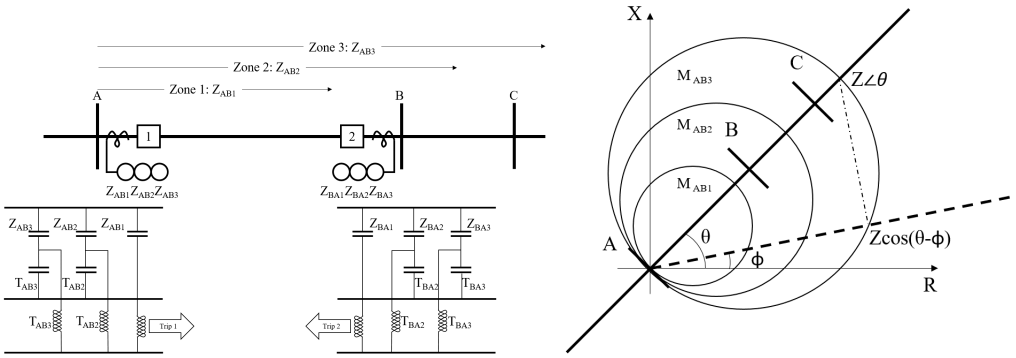

Figure 5.6: Plots showing three zone mho distance protection scheme (left) and the three zones of protection (right).

angle of the line. It is to be noted that this point lies along the impedance line A-B-C. Therefore, the zone is designed to issue trip signal if the apparent impedance (Z^{app}) is less than $Z\angle\theta$. However, under normal operating condition, a power system operates at high power factor. Hence the apparent impedance of a line lies at a considerably smaller impedance angle, such as along the dotted line with an angle of ϕ. Therefore, the apparent impedance encroaches the operating zone of the mho distance element at an impedance given by $Z\cos(\theta-\phi)$. If the operating power factor of the line ($\cos\phi$) is varied, the zone encroachment impedance also varies. Fig. 5.7 depicts the variation of zone encroachment impedance for a zone setting of $1\angle 80^o$ with line power factor varying from 0.1 to 1.0. This implies that trip signal is issued based on the operating power factor of line.

5.2.5 Hidden failures in mho distance protection relays

Consider the case of a hidden failure (contacts get permanently closed) in the timer contacts of Zone-2 or Zone-3. The trip coil is not energized until the apparent impedance encroaches any of these zones (since contacts Z_{AB2}, Z_{AB3} are open). Therefore, the failure in timer contacts remain hidden until the instant when apparent impedance encroaches the zone

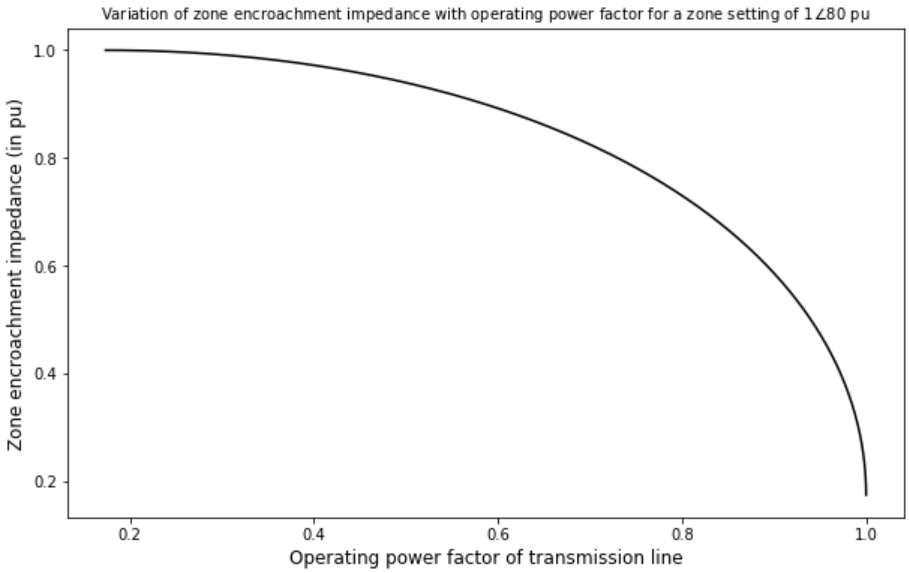

Figure 5.7: Variation in zone encroachment impedance with operating line power factor

which has a failed timer contact. This leads to an instantaneous trip in these backup zones which is unnecessary (since they are supposed to issue a trip signal with a particular delay). It is to be noted that the Zone 1 does not have any timer contact and hence there is no possibility of the occurrence of a hidden failure. Fig. 5.8 compares the operating region of mho distance protection element in the presence and absence of hidden failures in its timer contacts.

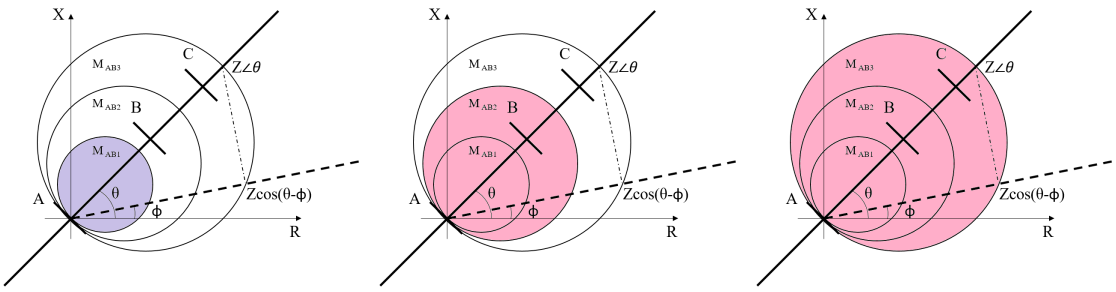

Figure 5.8: Plots showing operating regions of mho distance protection relay in the absence (left plot) and presence (middle and right plots) of hidden failures in the timer contacts. The violet region is the intended zone of instantaneous relay operation while pink zones indicate zones of operation if Zone-2 (middle plot) or Zone-3 (right plot) have hidden failure.

Let Z_A and Z_B be the apparent impedances measured at ends A and B respectively. Let $M_{AB1}, M_{AB2}, M_{AB3}$ indicate operating zones Zone-1, Zone-2 and Zone-3 for mho relay at A and $M_{BA1}, M_{BA2}, M_{BA3}$ be the same for mho relay at B. The condition of mho distance relay to issue an instantaneous trip signal for line A-B is given by

$$\begin{aligned} &\text{Breaker at A trips:} \quad Z_A \in M_{AB1} \\ &\text{Breaker at B trips:} \quad Z_B \in M_{BA1} \end{aligned} \tag{5.3}$$

For hidden failure in timer contacts, the condition for which instantaneous trip signal is issued for the line A-B is given by

$$\begin{aligned} &\text{Breaker at A trips:} \quad Z_A \in M_{AB2} \text{ and failure in } T_{AB2} \\ &\text{Breaker at A trips:} \quad Z_A \in M_{AB3} \text{ and failure in } T_{AB3} \\ &\text{Breaker at B trips:} \quad Z_B \in M_{BA2} \text{ and failure in } T_{BA2} \\ &\text{Breaker at B trips:} \quad Z_B \in M_{BA3} \text{ and failure in } T_{BA3} \end{aligned} \tag{5.4}$$

5.2.6 Overview of directional comparison blocking scheme

The third widely used protection scheme for transmission lines is the power line carrier (PLC) communication based directional comparison blocking scheme Horowitz and Phadke 1995. In this type of protection system, the direction of a fault on the transmission line is identified (within the protected line or external to it) and the information is sent to the remote end to allow/block the relay operation.

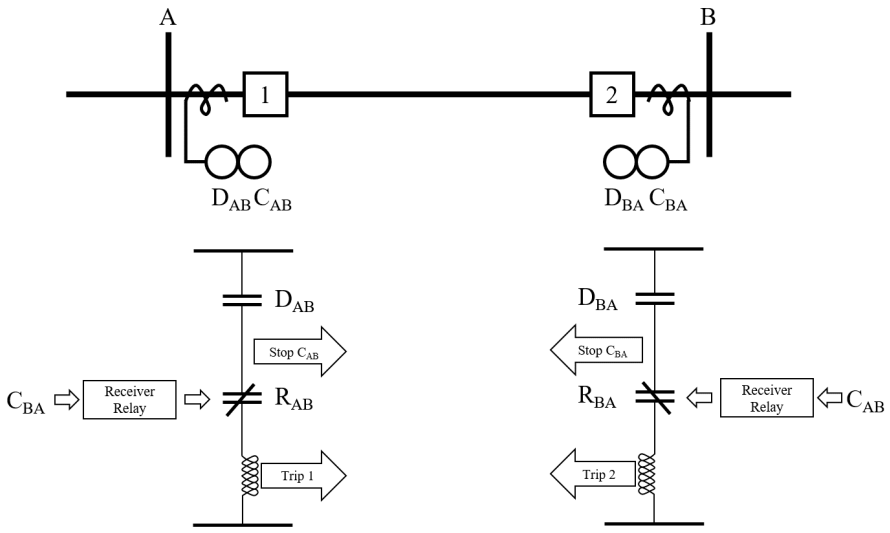

Figure 5.9: Schematic diagram of PLC based directional comparison blocking scheme

5.2.7 PLC based directional comparison block scheme

Fig. 5.9 shows a schematic of the directional comparison blocking scheme. Each end of the transmission line has a directional mho distance element (D_{AB},D_{BA}) and a reversed mho carrier start relay (C_{AB},C_{BA}) as shown in Fig. 5.10. The directional mho element is set to detect faults in the direction of the remote end. The reverse mho element detects a fault in the opposite direction and sends a carrier *block* signal to the receiver relay at the remote end. Transmission of this signal is stopped if the directional element detects a fault in its zone of operation. The receiver relay at the remote end opens the normally closed receiver contact if it receives a carrier *block* signal. Therefore, a trip signal is issued to the circuit breaker if the directional element has operated at an end and no blocking signal is received from the remote end.

If a fault occurs on the transmission line A-B in Fig 5.9, the directional elements (D_{AB},D_{BA}) detects the fault and stops the carrier start relays (C_{AB},C_{BA}) from transmitting the carrier block signal. Therefore, the directional contact operates and the receiver contacts remain

closed, thereby tripping the circuit breakers at A and B. If the fault is external to the line and beyond B, the directional element at A (D_{AB}) detects it and operates. It also stops transmitting the carrier block signal (C_{AB}) to the other end (B). The directional element at B (D_{BA}) does not operate since the fault is not in its operating region and hence breaker (2) at B does not trip. However, the fault is detected by the reversed mho carrier start relay at B (C_{BA}) which transmits carrier blocking signal to end A. This signal is received by the receiver element at A and opens the receiver contact (R_{AB}). Therefore, breaker (1) at A is inhibited from operation.

5.2.8 Hidden failures in PLC based directional comparison blocking relays

Similar to the previous two types of relays, the carrier based directional overcurrent protection relays are equally susceptible to hidden failures. Consider a mechanical failure in the receiver contact at any end where the contacts are permanently closed. In such a case, the relay issues a trip signal if the fault is detected in the zone of operation of the directional element at that end. A carrier blocking signal from the other end has no effect on the trip logic. Furthermore, such a failure remains hidden as the relay does not issue trip signal until the directional element detects a fault. We consider the mho relay characteristics as shown in Fig. 5.10 which compare the operating regions in the absence and presence of the hidden failures in the receiver contacts.

Let $M_{\text{Trip A}}$ and $M_{\text{Trip B}}$ be the operating characteristics of the directional mho elements at ends A and B respectively. Similarly, let $M_{\text{Block A}}$ and $M_{\text{Block B}}$ be the operating characteristics of the reversed mho carrier start relays responsible for transmitting blocking signals to ends A and B respectively. The condition for the PLC based directional comparison block-

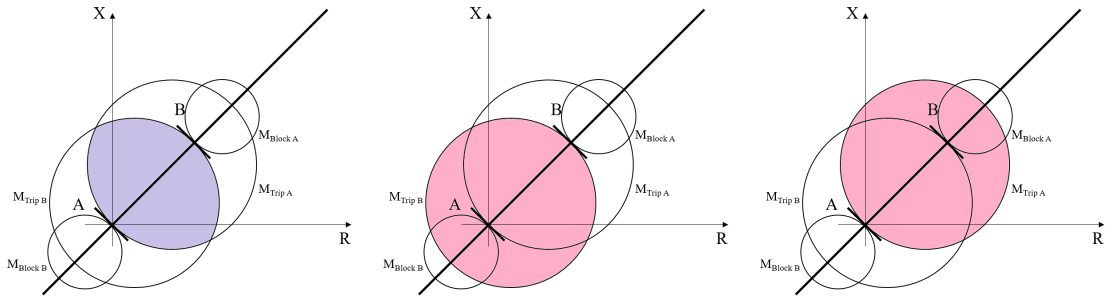

Figure 5.10: Plots showing operating region of PLC based directional comparison blocking relay in the absence (left plot) and presence (middle and right plots) of hidden failure in the receiver contact. The violet color region is the intended region of operation while pink region shows the region of operation when one of the receiver contact has a hidden failure.

ing relay to generate a trip signal based on the complex impedances Z_A and Z_B measured at A and B respectively is given by

$$\text{Breakers at A,B trip:} \quad Z_A \in M_{Trip\ A} \text{ and } Z_B \in M_{Trip\ B} \tag{5.5}$$

If the receiver contact of the relay at an end is permanently damaged due to a hidden failure, the relay issues a trip signal based on the directional element at the same end.

$$\begin{aligned} &\text{Breaker at A trips:} \quad Z_A \in M_{Trip\ A} \text{ and failure in } R_{AB} \\ &\text{Breaker at B trips:} \quad Z_B \in M_{Trip\ B} \text{ and failure in } R_{BA} \end{aligned} \tag{5.6}$$

5.2.9 Overview of percentage differential relay

The protection system used to detect internal faults in a transformer is the percentage differential protection scheme. This relay is a variation of the differential protection, where the currents at the two ends of the transformer are compared to identify internal faults. However, the traditional differential protection is sensitive to inrush currents which are treated as internal faults. Therefore, the percentage differential protection scheme is employed to

restrain the relay from tripping during such events.

5.2.10 Percentage differential protection scheme

Fig. 5.11 shows the schematic of a percentage differential relay for a transformer. The relay has two coils namely the operating coil and restraining coil. The operating coil responds to the difference in measured currents ($I_{As} - I_{Bs}$) at the two ends of the transformer A-B. The restraining coil responds to the average of the measured currents $\left(\frac{I_{As} + I_{Bs}}{2}\right)$. The restraining coil is present to avoid false trips due to heavy inrush currents.

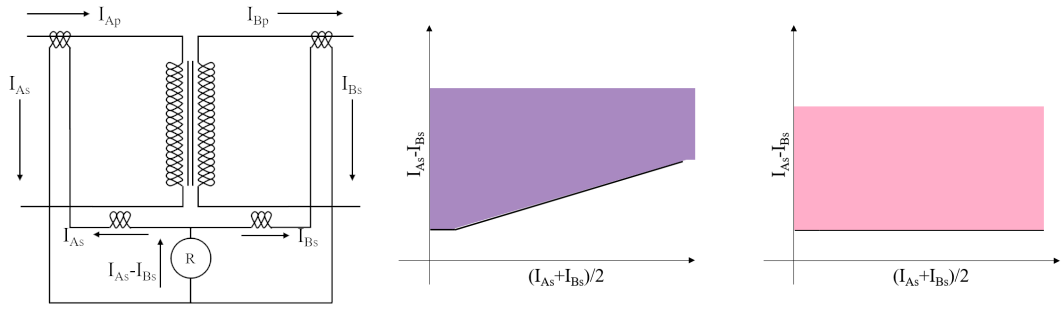

Figure 5.11: Plots showing schematic diagram (left plot) of percentage differential protection of transformer and operating region in the absence (middle plot) and presence (right plot) of hidden failure in the restraining coil.

Under normal operating condition, the differential current is caused due to the magnetizing component of the transformer equivalent circuit which accounts for $1-4\%$ of the rating Thompson 2011. The trip signal is sent to the transformer circuit breakers if

$$\left(I_{As} - I_{Bs}\right) - K\left(\frac{I_{As} + I_{Bs}}{2}\right) \geq I_{min}^{pick} \tag{5.7}$$

where K denotes the sensitivity of the percentage differential relay and is expressed as a percentage. I_{min}^{pick} is the minimum pickup value of the differential current.

5.2.11 Hidden failure in percentage differential relays

The middle and right plots in Fig. 5.11 show the operating region (shaded region) of a percentage differential relay. Similar to transmission line relays, hidden failures in the percentage differential relays can affect the tripping logic. If the restraining coil is shorted internally, the trip signal is generated if difference in the measured currents exceeds the minimum pickup value. Such false trip signals are generated for high currents flowing through the transformer. This failure remains hidden since the relay does not issue a trip signal until the difference in measured currents at the two ends of the transformer is considerable. Therefore, tripping condition of a percentage differential relay with hidden failure in the restraining coil is given by

$$I_{As} - I_{Bs} \geq I_{min}^{pick} \tag{5.8}$$

5.2.12 Generator Protection

Generators are the most vital equipment in a power grid and is responsible for maintaining the stability of the system. Therefore, they are provided with a wide range of protection systems. In this work, we are considering the overvoltage and undervoltage generator relays. If a generator bus violates the allowable voltage limits for a definite duration of time, the over/undervoltage relays issue corresponding trip signal.

In practical cases, when multiple HV transmission lines are tripped during a disturbance, the reactive support drops. This causes the generators to provide reactive power through excitation of the field (rotor) circuit. If it exceeds the generator capability, the overexcitation protection issues a trip signal. However, if it does not trip and the generator is made to operate above reactive power limits, the generator bus voltage drops and the undervoltage

relay issues a trip signal. In this study, we consider the trip signal of the undervoltage relays to be responsible for generator outages.

Since the generators are important from the stability aspect of the power system, the overvoltage and undervoltage relays are provided with fault clearance time (or delay). A separate timer is activated for each voltage limit violation. The generator trips if any of the timers reaches their respective fault clearance time. Generally, two overvoltage and two undervoltage limits are provided for every generator: one for moderate limit violation with longer clearance time and the other for severe limit violation with shorter clearance time.

In previous works, probability of occurrence of hidden failure for generator protection relays have been proposed in Bae and Thorp 1999 as a function of the generator voltages. However, keeping in consideration, the strict maintenance guidelines provided by the Federal Regulatory Commission NRC 1988, the occurrence of failures in the generator protection elements is considered less probable. Therefore, hidden failures in generator relays is not studied in the present work.

In the present study, the transient rotor angle stability is used to identify a system collapse. If a set of generator rotor angles differ from the rotor angles of another set by more than 180 degrees, the two sets of generators are said to be operating out of step Horowitz and Phadke 1995. In such a scenario, the generators trip due to the operation of out-of-step relays causing load-generation imbalance in the power grid. This causes frequency to drop below the allowable range and thereby triggers automatic under-frequency load shedding. Since this load-shedding is automatic, it can result in further load-generation imbalance. For example, over-frequency at the less loaded generator buses can cause the generators to trip and this process, if allowed to continue, can result in a widespread blackout.

5.3 Proposed Framework

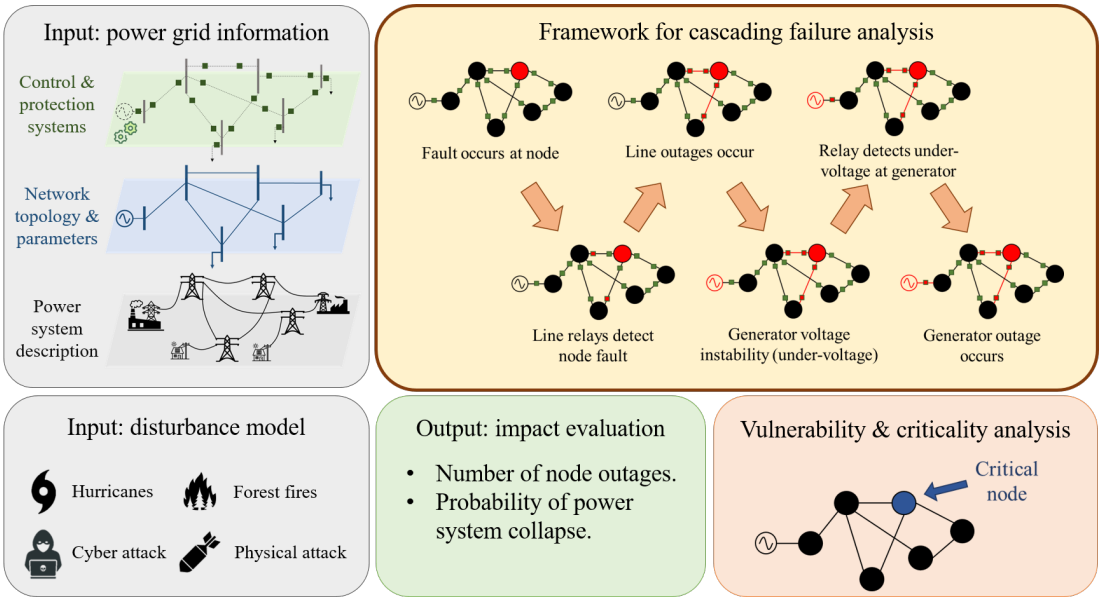

Figure 5.12: Proposed cascading failure framework with relevant input, output and analysis modules. The contribution of this work is the framework with its 'plug and play' capability rendering it useful to analyze any power system under any arbitrary disturbance model.

Protocol 1 lists the steps required for using our proposed framework. The same has been outlined in Fig. 5.12. The first step is to obtain the information regarding the power grid network (Step 1). This is stored as a graph network with a list of edges and nodes and their associated parameters. Thereafter, the generator control system is modeled from the parameters of the governor and excitation system (Step 2). The next step (Step 3) is to define the deterministic $1/0$ logic for the protection relays in the power grid. The list of hidden failures are also identified in this step. Then, the initiating event is to be defined (Step 4). Note that our proposed framework can be used for analyzing impact of cascading failures for any given contingency. To this end, we need to provide the framework with the list of nodes which are directly impacted by the event. We model the event as a three phase fault on these nodes to initialize the cascading failure analysis. Finally, we use Algorithm 7 to compute the impact of the initiating event on the power grid (Step 5).

Protocol 1 Outline of the proposed framework

Step 1: Construct the power grid network.
- a: Get the list of labeled edges from the table of transformers and transmission lines.
- b: Get the list of labeled nodes from table of substations, generators and load buses.

Step 2: Model the control and protection systems.
- a: Get parameters of generator governors.
- b: Get parameters of generator excitation systems.
- c: Define protection logic for generators, transmission lines and transformers.
- d: Identify list of relays with one or more hidden failures.

Step 3: Model the initiating disturbance event.
- a: Evaluate the list of nodes affected by the event.
- b: Model the initiating event as fault in power grid.

Step 4: Evaluate impact of the event.
- a: Perform transient stability analysis.
- b: Evaluate operation of protection relays.
- c: Compute the total number of node outages.
- d: Determine the stability of power grid.

5.3.1 Power System Model

The power grid network is represented as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n nodes and m edges contained respectively in the sets $\mathcal{V}$ and $\mathcal{E}$. The nodes of the graph are known as 'buses' in the power grid and the edges constitute transmission lines and transformers in the power system which connect pair of nodes. Let $\mathcal{E}_L$ denote the set of transmission lines and $\mathcal{E}_T$ be the set of transformers. We have $\mathcal{E} = \mathcal{E}_L \cup \mathcal{E}_T$.

We denote voltage of a node $i \in \mathcal{V}$ as the complex phasor $V_i = |V_i| e^{\mathrm{j}\theta_i}$ where $\mathrm{j} = \sqrt{-1}$. Note that the voltage magnitude is $|V_i|$ and phase angle is θ_i. Each node can consist of a generating unit injecting power into the network or a load which consumes power from the network or a combination of both generation and load. The complex power generation $S_i^g = P_i^g + \mathrm{j}Q_i^g$ includes real and reactive power generation $\left(P_i^g, Q_i^g\right)$. Similarly, the complex power consumption denoted by $S_i^c = P_i^c + \mathrm{j}Q_i^c$ includes real and reactive power consumption (P_i^c, Q_i^c). Also, the current injected into the node $i \in \mathcal{V}$ is denoted by the complex phasor

$I_i = |I_i|\, e^{\mathrm{j}\phi_i}$, where $|I_i|$ is its magnitude and ϕ_i is the phase angle. The power injection at a node $i \in \mathcal{V}$ is given by:

$$S_i = V_i I_i^{\star} = P_i + \mathrm{j}Q_i = \left(P_i^g - P_i^c\right) + \mathrm{j}\left(Q_i^g - Q_i^c\right) = S_i^g - S_i^c$$

where $I_i^{\star}$ is the complex conjugate of I_i. The current flowing along edge $e := (i,k) \in \mathcal{E}$ from node $i \in \mathcal{V}$ and directed towards node $k \in \mathcal{V}$ is a complex phasor denoted by $I_{ik} = |I_e|\, e^{\mathrm{j}\phi_{ik}}$. Similar to the voltage phasor, $|I_e|$ denotes the current magnitude and ϕ_{ik} is the current phase angle. It is trivial to note that the current flowing from node k towards node i is $I_{ki} = |I_e|\, e^{\mathrm{j}\phi_{ki}}$ where the phase angles are related as $\phi_{ki} = \pi + \phi_{ik}$ and the magnitude is the same. The complex power is denoted by $S_{i,k} = P_{i,k} + \mathrm{j}Q_{i,j}$ flowing along edge $e := (i,k) \in \mathcal{E}$ from node i to k, where $P_{i,k}$ and $Q_{i,k}$ respectively denote real and reactive power flowing along the edge. The resistance, reactance and shunt susceptance of edge $e \in \mathcal{E}$ are r_e, x_e and b_e^c respectively. The complex impedance of the edge is denoted by $z_e = r_e + \mathrm{j}x_e$ and the admittance by

$$y_e = \frac{1}{z_e} = \frac{r_e}{\sqrt{r_e^2 + x_e^2}} - j\frac{x_e}{\sqrt{r_e^2 + x_e^2}} = g_e + jb_e.$$

5.3.2 Power Flow Problem

The bus admittance matrix for the power system network is given by $\mathbf{Y} = \mathbf{G} + j\mathbf{B}$. The element along i^{th} row and k^{th} column of the matrix is as follows:

$$Y_{ik} = G_{ik} + jB_{ik} = \begin{cases} \sum_{l \neq i} \left(y_{il} + j\frac{b_{il}^c}{2}\right) & , \; i = k \\ -y_{il} & , \exists\, e := (i,k) \in \mathcal{E} \\ 0 & , \text{ otherwise} \end{cases} .$$

Table 5.1: Variables and parameters in power system model for cascading failure analysis

Notation	Description
$V_i = \lvert V_i \rvert e^{\mathrm{j}\theta_i}$	complex bus voltage phasor at bus $i \in \mathcal{V}$; magnitude: $\lvert V_i \rvert$, angle: θ_i
$I_i = \lvert I_i \rvert e^{\mathrm{j}\phi_i}$	complex bus current injection phasor at bus $i \in \mathcal{V}$; magnitude: $\lvert I_i \rvert$, angle: ϕ_i
$I_{ik} = \lvert I_e \rvert e^{\mathrm{j}\phi_{ik}}$	complex current flowing through edge $e : (i,k) \in \mathcal{E}$ from node i to node k; magnitude: $\lvert I_e \rvert$, angle: ϕ_{ik}
$S_{ik} = P_{ik} + \mathrm{j}Q_{ik}$	complex power flowing through edge $e := (i,k) \in \mathcal{E}$ from node i to node k; real power flow: P_{ik}, reactive power flow: Q_{ik}
$S_i = P_i + \mathrm{j}Q_i$	complex power injection at bus $i \in \mathcal{V}$; real power injection: P_i, reactive power injection: Q_i
$S_i^g = P_i^g + \mathrm{j}Q_i^g$	complex power generation at bus $i \in \mathcal{V}$; real power generation: P_i^g, reactive power generation: Q_i^g
$S_i^c = P_i^c + \mathrm{j}Q_i^c$	complex power consumption at bus $i \in \mathcal{V}$; real power consumption: P_i^c, reactive power consumption: Q_i^c
$z_e := r_e + \mathrm{j}x_e$	complex impedance of edge $e \in \mathcal{E}_R$; resistance: r_e, reactance: x_e
$y_e := y_{ik} = g_e + \mathrm{j}b_e$	complex admittance of edge $e := (i,k) \in \mathcal{E}_R$; conductance: g_e, susceptance: b_e
$b_e^c := b_{ik}^c$	shunt charging susceptance of edge $e := (i,k) \in \mathcal{E}_R$.

The power flow problem pertaining to any power grid is stated as: *given power injections at all nodes in the network, compute the node voltages and edge flows in the network* Bergen 1986; Powell 2004. The governing equations are as follows:

$$P_i = V_i \sum_{k=1}^{n} V_k \left[G_{ik} \cos(\theta_i - \theta_k) + B_{ik} \sin(\theta_i - \theta_k) \right] \tag{5.9a}$$

$$Q_i = V_i \sum_{k=1}^{n} V_k \left[G_{ik} \sin(\theta_i - \theta_k) - B_{ik} \cos(\theta_i - \theta_k) \right] \tag{5.9b}$$

Transient Stability Analysis

The ability of a power system to maintain synchronism when subjected to a large disturbance is termed as transient stability Kundur 1994; Sauer and Pai 1998; N. Hatziargyriou et al. 2021. In response to a rapid loss of load (or generation), the power system frequency

will increase (or decrease). However, the generator controls respond to this change by changing the power output to meet the electric load demand based on a set of differential equations. In the present study we have followed a numerical integration method to solve these equations. This means that for each time instant the AC power flow problem is solved to obtain the states of the power system. This solution is used as initial values for the differential equations required to solve the transient stability problem.

The stability study which occurs in the time range of 0.1 second after the initiating event to 10-30 seconds is called transient stability analysis Sauer and Pai 1998; N. Hatziargyriou et al. 2021. In response to a rapid loss of load (or generation), in the initial seconds of this time-frame, the system frequency will increase (or decrease). However, within a few seconds after that, the governors will respond (matching mechanical power input with electrical power output) and try to match the power outputs of the controllable generators with the load that exists in that time-frame. This dynamic balancing is described by the generator swing equation as shown below.

$$P_m - P_e(\delta) = M\ddot{\delta} + D\dot{\delta} \tag{5.10}$$

In Eq. 5.10, P_m is the mechanical power input, $P_e(\delta)$ is the electrical power output as a function of the electrical angle δ, M is a function of the machine's inertia, and D corresponds to the damping coefficient. Eq. 5.10 is the governing equation for a transient stability analysis.

Typically two approaches have been suggested for solving transient stability problems Bergen 1986:

1. Direct or energy methods: Mostly provides an intuitive insight into the transient stability problem. For a two bus system this method is known as the equal area

criteria.

2. Numerical integration methods: Most common technique, especially for large systems. In this approach, during the fault and after the fault the power system differential equations are solved using numerical methods.

In our study we have followed the numerical integration method. We considered three states of the system: (a) Pre-fault state $\mathcal{G}$: This is the state in which the system lies before the fault occurs. The system is assumed to be at a stable equilibrium point and this state provides the initial conditions; (b) Fault state $\mathcal{G}'$: The fault changes the system equations, moving the system away from its equilibrium point; (c) Post-fault state $\mathcal{G}''$ and $\mathcal{G}'''$: This is the state in which the system lies after the fault has been cleared. In this state, the system may or may not return to a stable equilibrium point or may even collapse altogether (all three outcomes possible).

The standard power system model is a set of non-linear differential algebraic equations. For all the three states of the system, we solved the algebraic network power balance equations and the differential equations using a partitioned approach where the solution of the algebraic equations alternates with the solution of the differential equations. The combined equations are given below.

$$\begin{aligned} \dot{x} &= f(x,y,p) \\ 0 &= g(x,y,p) \end{aligned} \tag{5.11}$$

In Eq. 5.11, x are the state variables $x \in R^n$; y are the algebraic variables $y \in R^m$; p are the independent variables $p \in R^l$; f are the differential equations $f : R^n \times R^m \times R^l \mapsto R^n$; and g are the algebraic equations $g : R^m \times R^m \times R^l \mapsto R^m$ Milano 2005. The algebraic equations g are obtained from the power flow equations in Eq. 5.9 and the differential equations f are from the swing equation Eq. 5.10. For this study, the Newton-Raphson (NR) method was

used for solving the power flow equations and determining the voltage magnitude and angle at every bus in the network. The numerical integration method is used to solve the swing equations. Eq. 5.11 is solved at every time step of the simulation which then accounts for any change that might have happened in the network conditions/topology due to the dynamics of the system.

5.3.3 Steady State/DC Analysis vs. Time Varying-AC Analysis

Given the simplicity of the DC power flow model, a variety of researchers have used this model to analyze the power system Pepyne 2007; Chertkov, Pan, and Stepanov 2011. Similarly, since steady state analysis on large networks is easier to perform, it was used for making assessments of the system state in C. L. Barrett et al. 2013; Pinar et al. 2010. However, the modern power system is a dynamic entity and to analyze it most accurately performing a detailed time-varying AC analysis is essential. Some of the advantages of the AC analysis over the DC analysis are as follows Powell 2004:

- The mismatch between generation and load is fully considered through active power and voltage angle components, while voltage stability and voltage regulation are fully addressed through reactive power and shunt reactance devices. Since the DC power flow assumes that voltage magnitudes remain fixed when the system undergoes dynamic changes, the control network that requires both real as well as reactive power components cannot be integrated with the power grid using such a model.

- As the DC load flow typically ignores the reactive power component altogether, its results are not very accurate. The results may be good approximations under some circumstances, but not all.

- The internal resistance of the transmission lines and transformers are neglected in the DC load flow. This means that using this analysis one cannot correctly evaluate the heating of the lines when the power flow is close to its thermal limits.
- The DC load flow model assumes the system to be loss-less. As this is not true for the actual system, the solution obtained using the DC load flow model is not an accurate depiction of the actual state of the power grid.

Similarly the steady state analysis does not consider the fact that a substation might trip and become out-of-service because of voltage instability caused by a power surge, and rotor angle instability during the transient swings. The generators slow down due to a sudden increase in load, or they speed up due to a sudden decrease in loads. This aspect is not considered in a steady state stability analysis but is an important component of transient stability analysis.

5.3.4 Power system collapse

A power system collapse can occur due to different reasons that can be attributed to the transient rotor angle stability and voltage stability of the grid. In the present study, the transient rotor angle stability is used to identify a system collapse. If a set of generator rotor angles differ from the rotor angles of another set by more than 180 degrees, the two sets of generators are said to be operating out of step Horowitz and Phadke 1995. In such a scenario, the generators trip due to the operation of out-of-step relays causing load-generation imbalance in the power grid. This causes frequency to drop below the allowable range and thereby triggers automatic under-frequency load shedding. Since this load-shedding is automatic, it can result in further load-generation imbalance. For example, over-frequency at the less loaded generator buses can cause the generators to trip and this

process, if allowed to continue, can result in a widespread blackout.

5.3.5 Operation of protection systems

The protection system in a power system detects faults and issues a trip signal to separate the faulted section from the healthy section. These protective relays play an important role in cascading outages since these are solely responsible for tripping edges and nodes in the network. We consider a generic protection system P_i that protects an element i (node or edge) and operates based on measured quantity W_{P_i}. Examples of measured quantities are current, voltage and impedance. Each protection system has a zone of operation, denoted by $\mathcal{R}_{\mathsf{P}_i}$. The trip decision $u_{\mathsf{P}_i} = \{0,1\}$ issued by the relay is a binary quantity where $u_{\mathsf{P}_i} = 1$ represents that the protected element i (node or edge) is disconnected and vice-versa. The element i is disconnected or *tripped* if the measured quantity W_{P_i} encroaches the zone of operation. That is, $u_{\mathsf{P}_i} = 1$ if $W_{\mathsf{P}_i} \in \mathcal{R}_{\mathsf{P}_i}$, and $u_{\mathsf{P}_i} = 0$ otherwise.

The zone of relay operation is contingent upon a set of conditions occurring simultaneously. Such conditions are physically realized through electro-mechanical/digital contacts, such that the relay issues a trip signal when all of them are satisfied. When a relay has a hidden failure in one or more of these contacts, only certain conditions are required to be satisfied to send the trip signal. Under such circumstances, its zone of operation alters to $\mathcal{H}_{\mathsf{P}_i}$ where $\mathcal{R}_{\mathsf{P}_i} \subset \mathcal{H}_{\mathsf{P}_i}$. For example, a directional overcurrent relay issues a trip signal to the circuit breaker if an overcurrent is detected in a particular direction. This is designed through two contacts - overcurrent and directional contacts. If a hidden failure occurs in the directional contact, the relay issues a trip signal as soon as it detects an overcurrent irrespective of the direction. Further, such a failure remains hidden because the relay does not issue any trip signal until an overcurrent is detected. Other details regarding hidden failures and their

way of occurrence in different relays are provided in the SI.

Drawing an analogy to the human immune response, the action of protective relays in a power grid is similar to the role played by active immunity developed in a human through direct clinical infection or by specific immunization. The antigens of infected cells are detected either by B-cells (humoral immunity) to create antibodies or by T-cells (cellular immunity) to initiate a chain of response required to neutralize the microbe or its toxin. In a similar fashion, a protective relay detects a fault in a power system and sends a trip signal to the circuit breaker in order to isolate the faulted section. An unwanted trip caused by a hidden failure in a protection relay is equivalent to an immune system which attacks an innocuous substance or its own uninfected cells causing *anaphylaxis* Schrom et al. 2021; Park 2015.

The hidden failures in relays have been modeled using a stochastic approach, wherein a set of relays with hidden failure are randomly sampled from a probability distribution. Let $\mathcal{K} = \left\{\mathsf{P}_{h_1}, \mathsf{P}_{h_2}, \cdots, \mathsf{P}_{h_n}\right\}$ denote the set of n relays with hidden failures that are randomly sampled from the entire set of relays.

5.3.6 Cascading failure model

The cascading failure model uses an AC power flow based time domain analysis to evaluate the states of the power system at each time instant. The simulation is initiated at time $t = 0$ with initial number of nodes $N_0 = |\mathcal{V}|$ (where $|\cdot|$ denotes cardinality of set) and condition of power system $c = 0$ (denoting stable system). At $t = t_f$, the target set $\mathcal{S}$ is attacked, which is done by simulating a three phase fault at the targeted nodes. The connected edges experience an outage due to the targeted attack. Thereafter, the time domain simulation is carried on until $t = t_{\text{end}}$ seconds with a time step of Δt. At each time step, the operation

of protective relays is monitored to identify tripped nodes and edges. Additionally, any isolated nodes in the network are considered as node outage. The simulation is stopped if a power system collapse occurs in between or if time reaches $t = t_{\text{end}}$ and the condition of power system is altered to $c = 1$. Algorithm 7 depicts the pseudocode for the proposed AC transient analysis based cascading failure model.

Algorithm 7 Cascading failure model

Input network topology ($\mathcal{G}(\mathcal{V}, \mathcal{E})$), network parameters and settings, relays with hidden failures ($\mathcal{K}$), target set ($\mathcal{S}$), time of attack (t_f), time increment (Δt), simulation end time (t_{end}).

Step 1: Initialize time instant $t = 0$.
Step 2: Initialize condition of power system $c \leftarrow 0$.
Step 3: Initialize number of nodes $N_0 \leftarrow |\mathcal{V}|$.
Step 4: Create three phase fault at $t = t_f$ on all nodes in $\mathcal{S}$.
Step 5: **while** time instant $t \leq t_{\text{end}}$ **do**
Step 6: Perform transient stability analysis.
Step 7: Identify tripped edges $\mathsf{F}(t) = \{e \in \mathscr{E} \big| u_{\mathsf{P}_e} = 1, \forall \mathsf{P}_e\}$.
Step 8: Remove tripped edges from graph $\mathscr{E} \leftarrow \mathscr{E} \setminus \mathsf{F}(t)$.
Step 9: Identify tripped nodes $\mathsf{G}(t) = \{g \in \mathscr{V} \big| u_{\mathsf{P}_g} = 1, \forall \mathsf{P}_g\}$.
Step 10: Remove tripped nodes from graph $\mathscr{V} \leftarrow \mathscr{V} \setminus \mathsf{G}(t)$.
Step 11: Identify isolated nodes $\mathsf{H}(t) = \{v \in \mathscr{V} \big| \mathsf{degree}(v) = 0\}$.
Step 12: Remove isolated nodes from graph $\mathscr{V} \leftarrow \mathscr{V} \setminus \mathsf{H}(t)$.
Step 13: **if** power system collapse occurs **then**
Step 14: Change condition to unstable $c \leftarrow 1$.
Step 15: Compute node outages $\Delta N \leftarrow N_0 - |\mathcal{V}|$.
Step 16: Stop the process.
Step 17: **end if**
Step 18: Increment time $t \leftarrow t + \Delta t$
Step 19: **end while**
Step 20: Compute node outages $\Delta N \leftarrow N_0 - |\mathcal{V}|$

Output number of node outages (ΔN), condition of power system (c).

5.4 Cyber attack model

In this section, a Bayesian attack tree based model is proposed to evaluate vulnerabilities in the SCADA system associated with the power grid. The proposed model identifies vulnerable SCADA architectures in the power grid which identifies potential target nodes for cyber attack on the power grid. This model can be augmented to the proposed cascading failure framework (discussed in the earlier section) to obtain the overall impact of a cyber attack on the power grid.

A planned cyber attack on the SCADA system takes place through multiple steps in which the software protection elements are compromised. This entire process can be effectively modeled using attack trees. A cyber intrusion consists of vulnerabilities in the cyber system and the dependency among them to be exploited. Therefore, a cyber attack can be represented as a directed graph with vulnerabilities denoted by the nodes and edges symbolizing the dependencies. In this section, the attack tree representation of a cyber attack on the SCADA system and the method to evaluate probability of successful intrusion are detailed.

5.4.1 Attack tree representation of vulnerabilities

An attack graph consists of two types of nodes: exploit to vulnerabilities and conditions required for exploiting. The preconditions needed to exploit a vulnerability are assumed to be either initial conditions of the attack or resulting output of some previously occurred exploit. In this case, three preconditions are considered to exploit a vulnerability: (*i*)service, (*ii*)connection and (*iii*)privilege required to access the vulnerability from previous exploit.

For example, a cyber intrusion scenario is considered for a control center SCADA system, where an adversary aims to gain unauthorized access to control assets in the power system.

The cyber intruder has to access the application server for this purpose which is dedicated to send control commands to open/close circuit breakers in the power system. In order to do so, the adversary needs to gain access of the historian server through a firewall and thereafter reach the application server through a different firewall as shown in Fig. 5.13.

Let there be two possible exploits to the vulnerabilities in the first firewall denoted by $\langle \mathrm{Ser}1, 0, 1 \rangle$ and $\langle \mathrm{Ser}2, 0, 1 \rangle$. The first one is assumed to be a zero day exploit and the second one is considered as a known exploitation. A zero day exploit to a vulnerability is one which may not be publicly known but identified by an intruder. In order to exploit either of the vulnerabilities, the intruder needs the privilege *user*(*0*) (which denotes him being present) and is required to be connected to the historian server through $\langle 0,1 \rangle$. Additionally, the vulnerabilities require the services *Ser1(1)* and *Ser2(1)* respectively to be available for them to be exploited. Once the vulnerability is successfully exploited, the intruder gains the user privilege *user*(*1*) of the historian server. This output of previously occurred exploit can be used as a precondition of the successive exploits. In the second firewall, we consider a single zero-day exploit to a vulnerability denoted by $\langle \mathrm{Ser}3, 1, 2 \rangle$. This can be successfully exploited by an intruder having privilege *user*(*1*) to obtain access to the application server.

Attack trees can be effectively modeled using Bayesian networks (BN) which are widely used to develop the probabilistic model for the same. BN is denoted by a directed graph $\mathscr{G}(\mathscr{V}, \mathscr{E})$ where the nodes are vulnerabilities and target conditions and the edges represent the path between them. In a BN, the probability of reaching each node is dependent on the conditional probability of its parent nodes. For this purpose, the individual probability of a successful exploitation of a vulnerability is required to be calculated. Every vulnerability can be scored based on its severity of being exploited by a standard Common Vulnerability Scoring System (CVSS) (Mell, Scarfone, and Romanosky 2006). Since the scores are provided on a scale from 0 to 10, they can be normalized through division by 10. If the

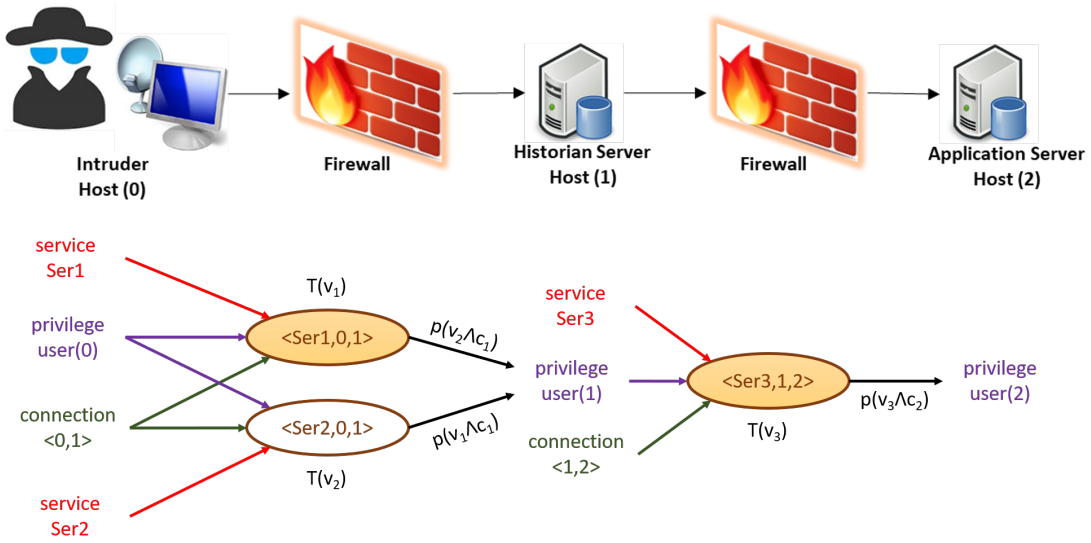

Figure 5.13: Cyber intrusion scenario in control center server.

preconditions are satisfied, each vulnerability node (v_i) can be exploited successfully with a probability equivalent to the normalized CVSS score.

$$\mathbb{P}(v_i|s_i,l_i) = \frac{\mathrm{CVSS}(v_i)}{10} \tag{5.12}$$

where s_i and l_i respectively denote that the service and connection required to exploit the vulnerability v_i are available. For known vulnerabilities, the CVSS scores can be evaluated (NIST 2007a; NIST 2007b) depending on the level of access complexity, authentication requirements and other factors. The CVSS score for the zero-day exploits are assumed to be 0.8. Thereafter, using the Bayes' theorem, the probability that a vulnerability (v_i) is successfully exploited is

$$\mathbb{P}(v_i) = \mathbb{P}(s_i)\cdot\mathbb{P}(l_i)\cdot\frac{\mathrm{CVSS}(v_i)}{10} \tag{5.13}$$

The probabilities of availability of service and connection given by $\mathbb{P}(s_i)$ and $\mathbb{P}(l_i)$ respectively are randomly selected from 0.85 to 1.0. The initial probability of availability of user privilege $\mathbb{P}(c_i)$ is considered to be 1.0 since it is assumed that the intruder is present to per-

form the cyber attack. For evaluating the availability of privileges in the successive target vulnerabilities, the probability of a successful intrusion through the preceding vulnerability is calculated. Therefore, the attack tree follows the structure of a *Markov Chain* where the probability of occurrence of a state is only dependent on the probability of occurrence of preceding state(s).

Certain access privileges can be achieved by exploiting more than one vulnerability from multiple prior access privileges. The probability of successfully reaching the condition c_i from n privileges ($c_j, j = 1,2,\cdots n$) through the m_j vulnerabilities $v_k, k = 1,2,\cdots m_j$ is given by

$$\mathbb{P}(c_i) = \sum_{j=1}^{n} \mathbb{P}\left(\bigcup_{k=1}^{m_j} v_k \right) \mathbb{P}(c_j) \tag{5.14}$$

It is assumed that the adversary does not waste any time in attacking multiple vulnerabilities of the same system while targeting a given goal condition. To this end, the probability of successfully exploiting a target vulnerability through a minimal attack sequence is considered. In order to calculate this, we need to evaluate individual probabilities of reaching target condition (c_i) from each of the possible vulnerabilities v_j. The probability of successful intrusion through each v_j to reach target c_i is given by

$$\mathbb{P}(v_j \wedge c_i) = \begin{cases} \mathbb{P}(v_j) \cdot \mathbb{P}(c_i|v_j), & j = 1 \\ \mathbb{P}(v_j) \cdot \prod_{k \neq j} \mathbb{P}(v_k = False), & j > 1 \\ \quad \cdot \mathbb{P}(c_i|v_j = True, v_{k \neq j} = False) & \end{cases} \tag{5.15}$$

5.4.2 Cyber system model of SCADA architectures

The cyber system model consists of the evaluation of probability of successfully exploiting the vulnerabilities and the calculation of time to compromise a vulnerability. The time to

compromise known and zero day vulnerabilities are evaluated in McQueen et al. (2006). These are denoted by $T(v_k)$ and expressed as an exponential function of k which represents the capability of the intruder in identifying a possible exploit to a vulnerability. Here, four possible skill levels are considered for the intruder with $k = 10, 2, 1, 0.01$. These intruders are identified as expert, professional, intermediate and amateur level adversaries respectively. The time to compromise decreases exponentially with increase in skill level.

Fig. 5.13 shows the simplified Bayesian attack tree for the attack path in the example. The time to compromise each vulnerability is shown beside each node and the probability of successfully reaching target condition c_i from vulnerability v_j denoted by $\mathbb{P}(v_j \wedge c_i)$ is shown beside each edge. The mean time to compromise (MTTC) and reach a target access level c_i from n possible prior access levels $c_j = 1, 2, \cdots, n$ is calculated. Let there be m_j vulnerabilities to reach from access level c_j to c_i given by $v_k, j = 1, 2, \cdots, m_j$.

$$\mathsf{MTTC}(c_i) = \frac{1}{\mathbb{P}(c_i)} \left[\sum_{j=1}^{n} \mathsf{MTTC}(c_j) + \sum_{k=1}^{m_j} \mathbb{P}(v_k \wedge c_i) T(v_k) \right] \tag{5.16}$$

The mean time to compromise a vulnerability with exploit code available is 1 day as evaluated in McQueen et al. (2006). Therefore, the attack efficiency (ζ) for the target c can be calculated as

$$\zeta(c) = \frac{1}{\mathsf{MTTC}(c)} \tag{5.17}$$

Here, three substation LAN models and a SCADA model for control center are considered and the attack tree for the same are generated. In each substation model, the intruder aims to attack the human machine interface (HMI) to gain administrator access and thereafter send control signals to trip breakers in the physical power system. In the control center model, the goal of intruder is to access the application server.

Substation LAN model A

In this model, the HMI, workstations and the IEDs at a substation are connected to a common LAN network as shown in Fig. 5.14. A single firewall with an ethernet switch controls the passage of information to and from the network. The attack graph is shown in the bottom figure. In this case, the intruder can exploit a vulnerability in the firewall to directly access the HMI which is connected to the LAN network. Two most popularly used protocols for remote access are the file transfer protocol (FTP) and the secure shell (SSH). It is assumed that the FTP vulnerability is a known type and SSH vulnerability is a zero day type. Once the intruder accesses the HMI, a buffer overflow vulnerability (bof) can be exploited to gain administrator privilege on the HMI system.

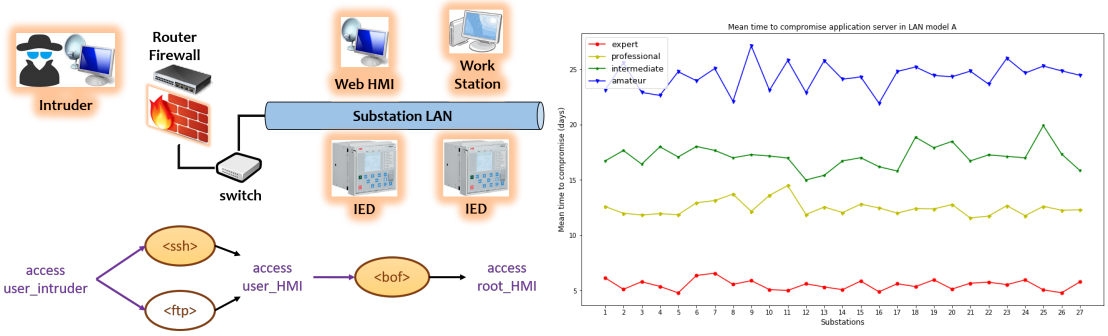

Figure 5.14: Intrusion scenario in LAN Model A

Substation LAN model B

In this model, the substation LAN is divided into two virtual LANs (VLANs). The substation VLAN connects the workstations, HMI and other control units and the bay VLAN connects the IEDs in the switchyard. These two VLANs communicate with a shared server which alternates between the networks through a pair of ethernet switches as shown in Fig. 5.15. Such an architecture increases the level of security of the SCADA system. The

bottom figure shows the attack graph based on the above architecture. An intruder can exploit a vulnerability in the firewall to access the shared server. In this case a known cross scripting vulnerability (XSS) is considered which allows remote attackers to arbitrarily inject web script to access the shared server (MITRE 1999a). Thereafter, the HMI can be accessed by exploiting a vulnerability from the shared server. The remote access of the HMI can be done from the the intruder system directly as in case of model A through the two popularly used protocols FTP and SSH.

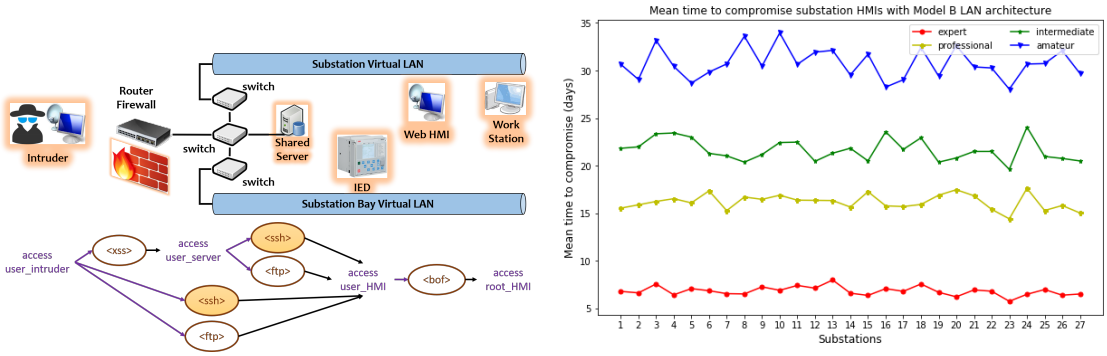

Figure 5.15: Intrusion scenario in LAN Model B

Substation LAN model C

Fig. 5.16 shows the architecture of this LAN model and corresponding attack graph. In this case, a local SCADA system connects all components in the substation LAN. The HMI cannot be accessed remotely and all communication has to pass through the local SCADA system. The intruder can exploit a HTTP vulnerability in the SCADA firewall to gain access of the local SCADA system. This vulnerability can cause denial of service (DoS) in the servers (MITRE 1999b). Thereafter, the HMI can be accessed by exploiting an FTP or SSH vulnerability from the local SCADA. Finally, vulnerabilities within the HMI can be exploited to gain administrator privilege on the system.

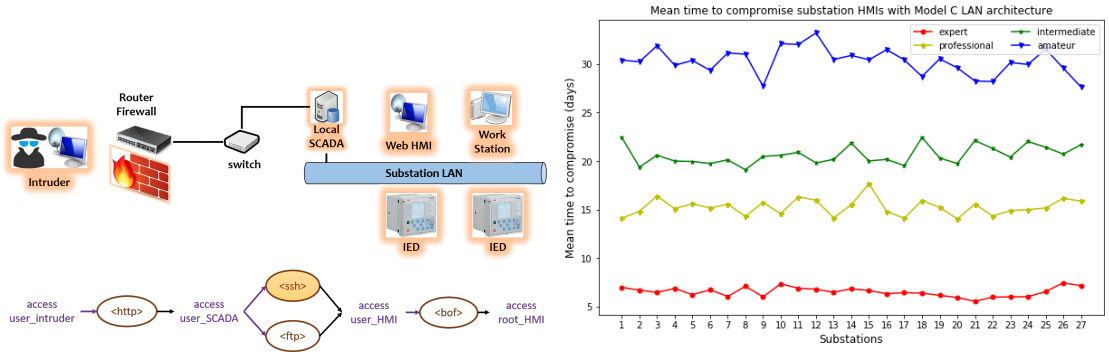

Figure 5.16: Intrusion scenario in LAN Model C

5.5 Case study: physical attack on the Washington DC power grid

In this chapter, we provide a case study where a targeted adversarial attack is considered. The scenario is caused by a detonation of a bomb at one or more substations located in and around Washington DC, USA. The attacks depict a scenario that is aimed at harming the power grid and thereby *indirectly* affecting the human population of the city. We present two different types of targeted attack on the power grid of Washington DC and its neighboring areas.

- *Type 1*: a large scale attack caused by detonation of a tactical device which results immense infrastructural damage in the attacked region, and
- *Type 2*: a simultaneous strategic targeted attack on critical substations with an aim of creating a cascading failure throughout the power network.

The first scenario can be considered as a large bomb blast in downtown Washington DC. This results in an immense loss of property. However, it needs to be examined whether the disturbance created in the power grid following the event results in a cascading outage in

the neighboring regions. The second scenario is a more planned attack on selected substations with the aim of maximizing impact by creating a large scale power outage throughout the grid. The property damage may be limited within the boundaries of the targeted substations; however, the cascaded outage, if one occurs, will severely affect the societal infrastructures in and around Washington DC. There can be multiple possible choices of targeted attack and therefore, we had studied the *optimal critical node* problem in Meyur, Vullikanti, et al. 2018. We have summarized the result in the Appendix. We had observed that a greedy choice of high voltage substations (500kV) leads to a significant impact on the power grid. Fig. 5.17 shows the two scenarios and the resulting impact obtained after analyzing each of them using the cascade failure model. We notice that for an attack of *Type 1*, the cascaded outages are contained within the boundaries of Washington DC. However, a *Type 2* targeted attack on a single critical node outside Washington DC has resulted in widespread cascading outages.

We analyze each of these attack scenarios through the proposed AC transient analysis based cascading failure model (Algorithm 7). At the beginning of each scenario, we randomly identify a set of relays with hidden failures denoted by $\mathcal{K}$. This is the only stochastic aspect of the simulation, following which the cascading failure model proceeds in a deterministic fashion. In order to cater for the stochastic presence of hidden failure, we run each scenario multiple times (20 times in our case). We model each attack as a three phase fault on the set of targeted nodes. The protection relays operate based on their settings and causes the following outages.

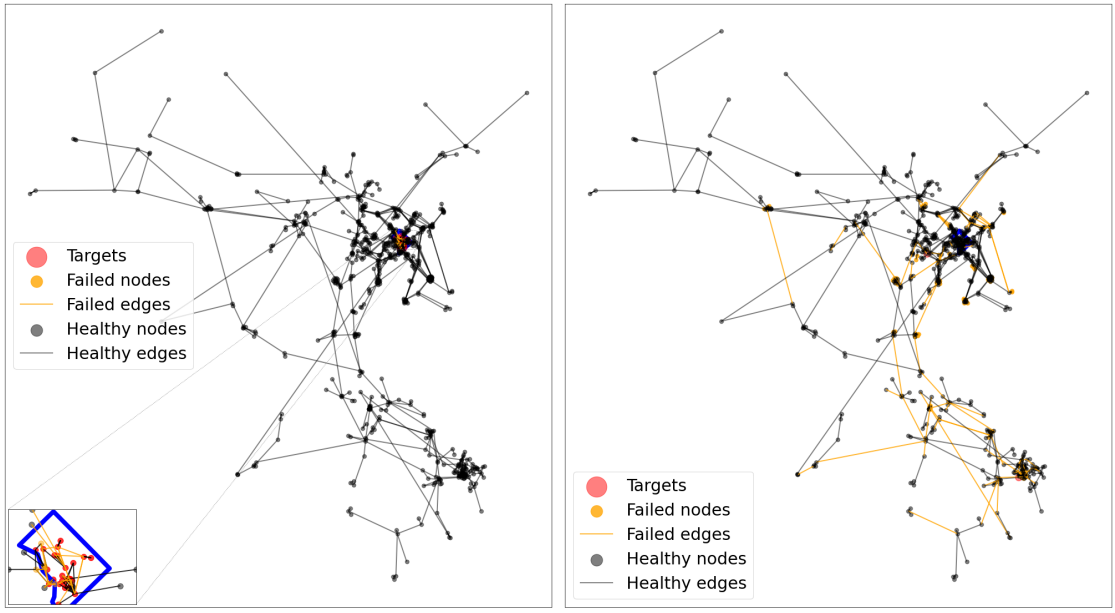

Figure 5.17: Figure showing a large scale physical attack on Washington DC (left) and targeted attack on strategically selected substation nodes (right). Blue boundary shows Washington DC, red nodes are physically damaged due to the attack and cascading failure propagates to the adjacent orange nodes and edges which trip due to operation of protective relays.

5.6 Results

Fig. 5.18 shows the number of node outages which follows a large scale *Type 1* attack and three different *Type 2* attacks on different critical sets of 500kV substations. We list down the key observations here.

1. A *Type 1* attack on Washington DC does not result in a cascading power outage outside the boundary of DC.

2. Increasing the number of nodes in the target set for a *Type 2* attack does not increase the resulting impact of attack.

3. An increase in hidden failure occurrence probability reduces the extent of cascading outages for *Type 2* attack on the Washington DC grid.

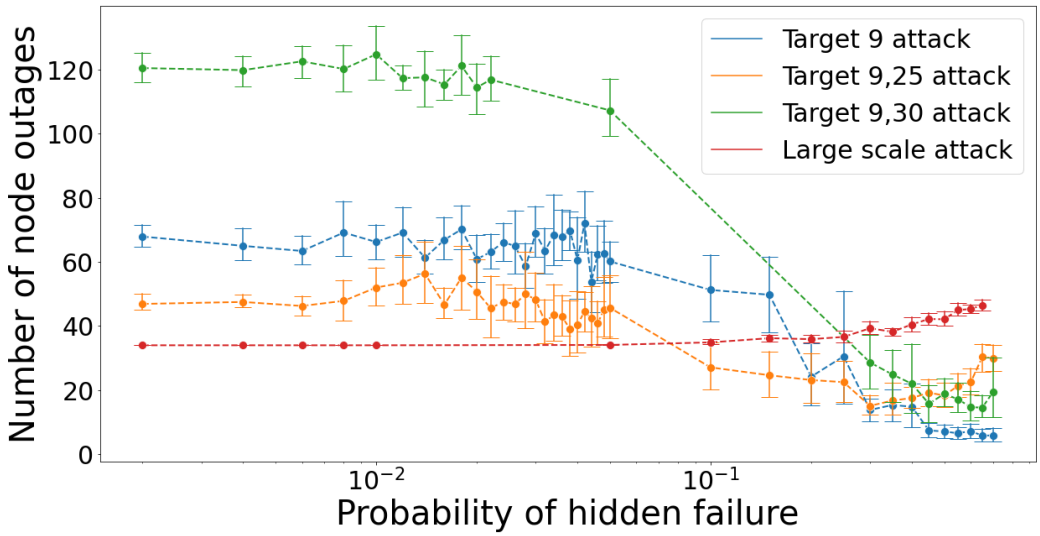

Figure 5.18: Plot showing variation sensitivity of number of cascading node failures to the probability of hidden failures in protection relays. The node outages caused due to a *Type 1* attack on Washington DC is almost immune to the occurrence of hidden failures. The cascading node outages caused by a targeted attack reduces with increase in hidden failure probability.

Observation 5.1. *A* Type 1 *attack on Washington DC is less likely to cause cascading outages as compared to a strategic targeted attack of* Type 2 *on selected 500kV substations.*

We notice that a *Type 1* attack has the minimum impact in terms of the number of node outages in the power grid. However, for a targeted attack of *Type 2*, the number of node outages is significantly higher and often results in system collapse. Washington DC is an area with high load consumption due to large number of residential and commercial establishments; but it does not have enough generating capacity within its geographic boundary. Therefore, there is a significant amount of power which is imported from the neighboring regions (see SI for detailed power flow values).

Fig. 5.19 compares the loss of generation and load in the power grid for the two types of adversarial attack. In case of a *Type 1* attack in Washington DC, there is a large loss of load within the boundary of the city. Therefore, little to no power is required to be imported from the neighboring regions. On the contrary, a targeted attack on selected substations along one of the power import paths causes outage of *important* transmission lines connecting Washington DC to the neighboring regions, thereby interrupting the power flow along them. The outage of significant 500kV lines result in a reduced power support in the power system. The deficit needs to be supplied by the generating units along the other import paths. This results in transient rotor angle and voltage instabilities in the power grid and further loss of generation.

Consequences of active power deficit. The active power generation is limited by the governor control in most generators. The lack of load-generation balance in the overloaded generators causes rotor angle instability. This is because overloaded generators decelerate and they operate *out-of-step* with other generators (rotor angles difference exceeds 180 degrees). The out-of-step protection relays causes generator outages during such an occurrence. This, in turn, creates further load-generation imbalance due to reduced generation

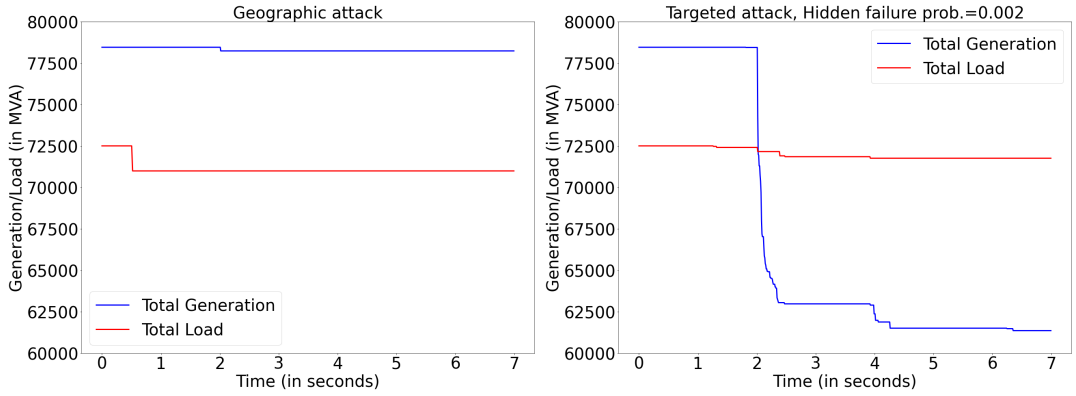

Figure 5.19: Plots showing change in total system load and generation in the aftermath of a *Type 1* attack on Washington DC (left) and a targeted attack of *Type 2* on a strategically selected substation node (right). For the *Type 1* attack, there is a large drop of load (nearly 2000 MW). For the *Type 2* attack, there is no immediate load drop; rather we observe massive loss of generation due to instability.

and thus eventually leads to system collapse. We show this comparison in Fig. 5.20. The *Type 1* attack on Washington DC does not result in any sustained oscillations of the rotor angles of major generators in the area. On the contrary, the generator rotor angles go *out-of-step* in the case of the *Type 2* targeted attack.

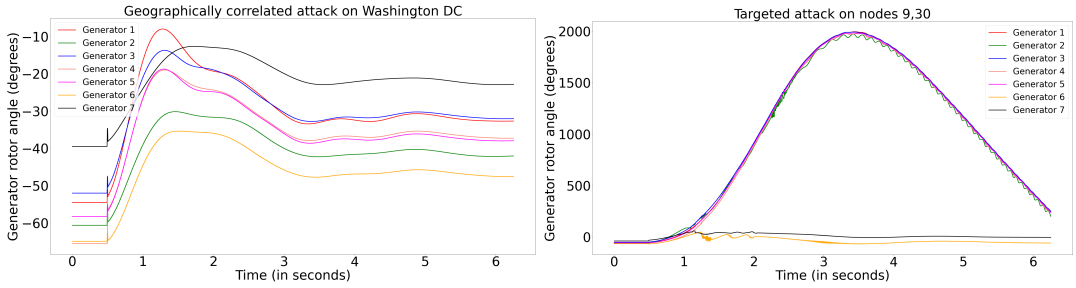

Figure 5.20: Plots showing the rotor angle oscillation of important generators around the Washington DC region. For the *Type 1* attack (left), the rotor angle oscillation stabilizes quickly. For a *Type 2* attack (right), the oscillations are severe and some generators are *out-of-step* with the other generators.

Consequences of reactive power deficit. The reactive power generation is limited by the field excitation of generators. The overcurrent protection in field circuit results in outages of the overexcited generators. The lack of reactive power support causes further voltage

collapse, more line outages and eventually leading to voltage instability and eventual system collapse as depicted for the case of targeted attack in Fig. 5.21. In case of the *Type 1* attack, the voltage oscillations stabilizes within a short duration.

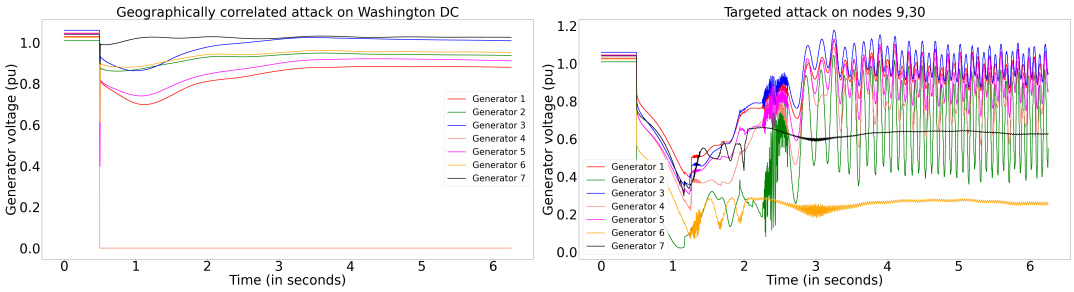

Figure 5.21: Plots showing the voltage oscillation of important generators near the Washington DC region. For the *Type 1* attack (left), the voltage oscillation stabilizes. For a *Type 2* attack (right), we observe voltage instability issues.

Observation 5.2. *Increasing the number of nodes in the target set does not increase the impact of attack. Additionally, the power flow magnitude along transmission line is not the primary determinant of the occurrence of a line outage.*

We observe the outage of a significant high voltage (500kV) transmission line connecting Washington DC and carrying almost 250MVA. We consider three different scenarios of *Type 2* attack for the purpose of comparison. The choice of the nodes in the target set is based on criticality analysis performed in Meyur, Vullikanti, et al. 2018. The necessary details are provided in the SI.

- *Scenario 1*: The target set consists of a single node (Target ID 9). It is known to produce maximum impact resulting in the occurrence of a system collapse when targeted singularly.
- *Scenario 2*: The target set consists of two nodes Target IDs 9,25. It has been observed to result in the minimum impact when Target ID 9 is combined with other 500kV nodes.

- *Scenario 3*: The target set consists of two nodes Target IDs 9,30. This produces the maximum impact when Target ID 9 is combined with other 500kV nodes.

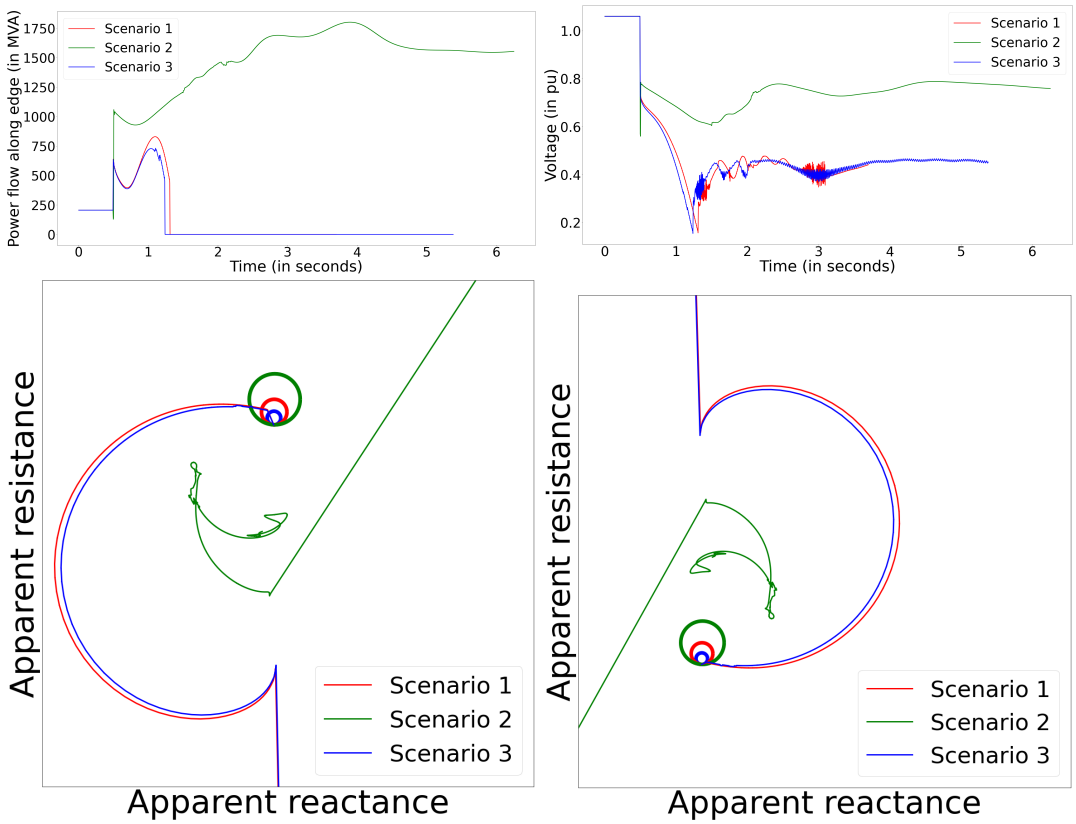

Figure 5.22: Plot showing variation of edge flow in a single transmission line (top left), node voltage at one of its end (top right), and apparent impedance trajectory measured by mho relays on two ends (bottom) for three targeted attack scenarios. Each color depicts a different targeted attack scenario. The node voltage drops considerably for Scenarios 1 and 3 as compared to Scenario 2. Apparent impedance trajectory enters the zone characteristics of the mho relays for Scenarios 1 and 3 which results in line trips. Scenario 3 results in the maximum power flow along the line, yet the mho relays do not trip.

The apparent power (in MVA) flowing through the transmission line in the three scenarios is shown in the left plot of Fig. 5.22. Note that after the targeted attack (at $t = 1$s), the power flow along the line increases in all the three scenarios. However, the line outage occurs in Scenarios 1 and 3 (at $t = 2.5$s), but there is no outage in Scenario 2. Further, we note that among the three scenarios considered in this example, the power flow along the

transmission line is maximum for Scenario 2; yet it does not result in a line outage.

The apparent impedance trajectories as measured by a *mho relay* installed to protect the transmission line is shown for the three scenarios in the middle plot of Fig. 5.22. While the apparent impedance encroaches the zones of protection (shown by the circular regions) for Scenarios 1 and 3, it is far away from the zones for Scenario 2. This is because of the voltage instability that occurs for Scenarios 1 and 3. We show the voltage profile at one of the connecting buses in the top-right plot of Fig. 5.22. We note that the voltage drops to a significantly low value which resulted in the apparent impedance to encroach the zones of protection.

In typical power grids, operation of transmission line relays is not dependent on power flow measured at one of its ends. It is rather protected using mho relays, directional overcurrent relays and carrier based directional comparison block relays. The operation of these relays are principal determinant of transmission line outages and are instrumental in causing cascading outages. Majority of prior works Carreras, Newman, et al. 2000; Dobson, Carreras, et al. 2001; Carreras, Lynch, et al. 2002b; Carreras, Lynch, et al. 2002a; J. Chen, Thorp, and Dobson 2005 model cascading outages using probabilities assigned based on power flowing through lines. Such approximate models are often reasonable due to dependence of current and power flowing through a line. However, they fail to represent the non-linear relationships between the power engineering quantities (current, voltage, power and impedance). A realistic representation of relay operation is necessary in the cascading failure models in order to capture the complex dynamics of cascading outages. Being devoid of such models, prior works often fail to accurately identify possible cascading outages. For instance, in the above example, a power flow based line outage model would result in a line trip for Scenario 2, since the post-event power flow through the line is maximum among the three scenarios. However, our proposed framework with realistic representation of relays show

that the line outage occurs for the other scenarios, and not Scenario 2.

Observation 5.3. *With an increase in occurrence of hidden failures within protection system, the extent of cascading outages reduces.*

To elaborate this observation, we consider Scenario 3 where simultaneous targeted attack is performed on Target IDs 9,30. We study the impact of two different probabilities (0.2% and 65%) of hidden failure occurrence in transmission line protection relays. Figure 5.23 compares the evolution of net load and generation in the region for the two cases as the cascading outages propagate over few seconds after the attack. The inset figures show the drop in load immediately after the physical attack occurs. A high probability of hidden failure occurrence results in the outage of a number of transmission lines immediately after the attack, leading to disconnection of loads from the power grid. This does not happen for low hidden failure occurrence probability. The disconnection of significant amount of load helps in maintaining the load-generation balance of the disturbed power grid. Therefore, the operating generators do not suffer from rotor angle and voltage instabilities.

A 65% probability of occurrence of hidden failures is an extremely pessimistic estimate even for the worst maintained power grids, and hence the premise of our analysis might be an unlikely to happen in practice. Yet, our results show an important observation that an immediate disconnection of large loads from the power grid or operating major load centers as self sustaining microgrid networks can avoid a system wide collapse and a possible widespread blackout. In our analysis, the immediate disconnection of large loads occur as a result of relay misoperations due to hidden failures. In practice these can be done by power system operators in real time, even without the occurrence of hidden failures. This necessitates development of communication aided protection systems so that the relays exchange signals among each other and preferably update protection strategies Schrom et al. 2021. Wide area measurement systems (WAMS) and wide area protection systems

(WAPS) which use phasor measurement units (PMUs) will play a key role in such communication aided protection strategies. Few examples of such strategies include majority voting scheme based relay decisions and system protection schemes (SPS) which are specific to a particular system Phadke and Thorpe 2008; Horowitz and Phadke 1995.

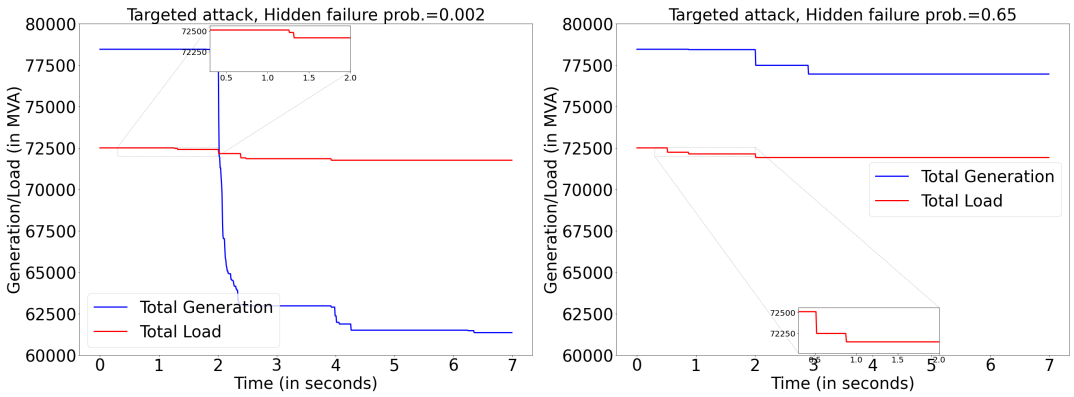

Figure 5.23: Plots showing change in total system load and generation in the aftermath of a targeted attack on a strategically selected substation node with low (left) and high (right) probability of hidden failure occurrence in relays. For a high probability of hidden failure, a number of transmission line outages occur following the targeted attack which causes a large amount of load getting isolated in the network, which eventually facilitates in maintaining load-generation balance.

5.7 Comparison of AC and DC Cascading Models

In this section, the traditional DC power flow based steady-state analysis is compared with the proposed cascading failure model. For the DC power flow analysis, the admittance matrix and power injections at each bus in the power system are evaluated. With the usual assumption of flat voltage profile at each bus and neglecting reactive power, the bus voltage angles are estimated. A transmission line is tripped if the electrical angular separation between the ends is more than 70 degrees as used in NERC 2012. The same set of process is executed until there are no more outages. A node with no edges connected to it is considered as a tripped node. Fig. 5.24 depicts the impact of targeted attack on high degree

nodes in the network. In comparison to the proposed model, the DC steady-state analysis underestimates the number of node outages. Furthermore, it does not identify a power system collapse due to transient instability.

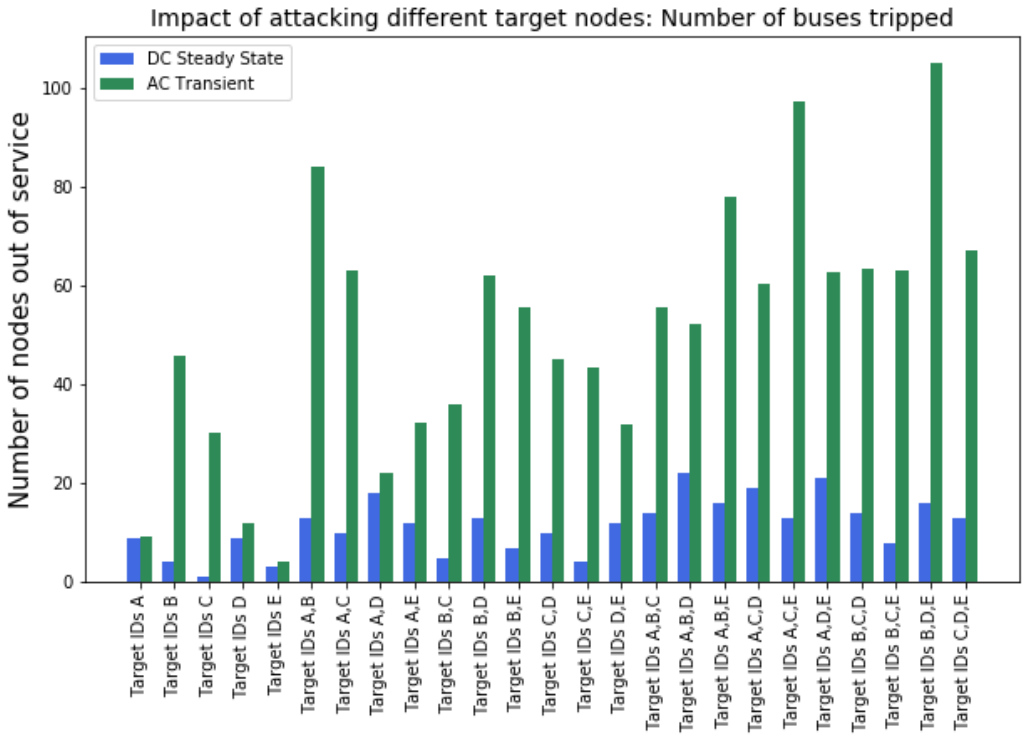

Figure 5.24: We observe the impact of targeted attack on high degree nodes using DC steady-state analysis. DC power flow based steady-state analysis underestimates the actual impact of a physical attack.

5.8 Concluding Remarks

We built a framework to assess vulnerability of the power grid to cascading outages when subjected to a severe disturbance. The framework includes a detailed realistic representation of generation, load and transmission lines of the power grid and uses AC power flow based transient stability analysis to study the cascading outages. The operation of protec-

tion relays is considered as the principal determinant of cascading outages. Therefore, we use detailed models of traditional electro-mechanical/digital relays, and consider occurrences of probable hidden failures in them. Our generic framework allows us to perform vulnerability analysis of a given power grid initiated by any severe event, such as hurricane, earthquake, forest fires or a targeted cyber/physical attack. We show the necessity of realistic representation of protection systems to capture power grid dynamics as accurately as possible.

Particularly, we have analyzed the impact of a targeted adversarial physical attack on the power grid of Washington DC. We compare a large scale physical attack (*Type 1* attack) initiated due to detonation of a tactical device at a large geographic region with a strategically targeted simultaneous attack on selected substations located far apart (*Type 2* attack). Our results show that albeit severe societal damage encompassing a large region, a *Type 1* attack on Washington DC area does not result in cascading power outages. On the contrary, a strategic simultaneous attack on selected substations can result in widespread cascading outages in and around Washington DC region. Though physically less severe than the former, this attack can impact a larger population through the resulting cascading power outage.

Bibliography

Federer, H. (1969). *Geometric Measure Theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. New York: Springer-Verlag. ISBN: 978-3540606567.

Raphael, A. F. and J. L. Bentley (1974). "Quad trees a data structure for retrieval on composite keys". In: *Acta Informatica* 4, pp. 1–9.

McCormick, G. P. (Dec. 1976). "Computability of global solutions to factorable nonconvex programs: Part I - Convex underestimating problems". In: *Mathematical Programming* 10.1, pp. 147–175.

Bender, E.A. and E.R. Canfield (1978). "The asymptotic number of labeled graphs with given degree sequences". In: *Journal of Combinatorial Theory, Series A* 24.3, pp. 296–307.

Preparata, F. P. and M. I. Shamos (1985). *Computational Geometry: An Introduction*. Berlin, Heidelberg: Springer-Verlag.

Bergen, A. R. (1986). *Power System Analysis*. Prentice Hall Series in Electrical and Computer Engineering.

Bak, P., C. Tang, and K. Wiesenfeld (July 1988). "Self-organized criticality". In: *Phys. Rev. A* 38 (1), pp. 364–374.

NRC, US (Mar. 1988). *Final Commission Policy Statement on Maintenance at Nuclear Power Plants*.

Billinton, R. and W. Li (1994). "Basic Concepts of Power System Reliability Evaluation". In: *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods*. Boston, MA: Springer US, pp. 9–31.

Grainger, J. J. and W. D. Stevenson (1994). *Power System Analysis*. McGraw-Hill.

Kundur, P. S. (1994). *Power System Stability and Control*. McGraw Hill Education.

Eppstein, D. (1995). "Subgraph Isomorphism in Planar Graphs and Related Problems". In: *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '95. San Francisco, California, USA: Society for Industrial and Applied Mathematics, pp. 632–640. ISBN: 0898713498.

Horowitz, S. H. and A. G. Phadke (1995). *Power System Relaying*. Research Studies Press, 2nd Edition, Taunton, UK.

Tamronglak, S. et al. (Apr. 1996). "Anatomy of Power System Blackouts: Preventive Relaying Strategies". In: *IEEE Transactions on Power Delivery* 11.2, pp. 708–715.

Sauer, P. W. and M. A. Pai (1998). *Power System Dynamics and Stability*. Prentice Hall.

The White House (1998). *White Paper: The Clinton Administration's Policy on Critical Infrastructure Protection: Presidential Decision Directive*. https://clintonwhitehouse4.archives.gov/WH/EOP/NSC/html/documents/NSCDoc3.html.

Thorp, J. S. et al. (Feb. 1998). "Anatomy of power system disturbances: Importance Sampling". In: *International Journal of Electrical Power & Energy Systems* 20.2, pp. 147–152.

Bae, K. and J. S. Thorp (Jan. 1999). "A Stochastic Study of Hidden Failures in Power System Protection". In: *Journal of Decision Support Systems* 24.3, pp. 259–268.

Kosterev, D. N., C. W. Taylor, and W. A. Mittelstadt (Aug. 1999). "Model validation for the August 10, 1996 WSCC system outage". In: *IEEE Transactions on Power Systems* 14.3, pp. 967–979.

MITRE (1999a). *CVE Database of FTP Vulnerability*. https://www.cvedetails.com/vendor/2/FTP.html. Last accessed 2 Dec 2020.

— (1999b). *CVE Details of Apache HTTP Vulnerability*. https://www.cvedetails.com/vendor/45/Apache.html. Last accessed 2 Dec 2020.

Carreras, B. A., D. E. Newman, et al. (Jan. 2000). “Initial Evidence for Self-organized Criticality in Electric Power System Blackouts”. In: *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*.

Dobson, I., B. A. Carreras, et al. (Jan. 2001). “An Initial Model for Complex Dynamics in Electric Power System Blackouts”. In: *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pp. 710–718.

Carreras, B. A., V. E. Lynch, et al. (Sept. 2002a). “Critical Points and Transitions in an Electric Power Transmission Model for Cascading Failure Blackouts”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 12.4, pp. 985–994.

— (Jan. 2002b). “Dynamics, Criticality and Self-organization in a Model for Blackouts in Power Transmission Systems”. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. Big Island,HI.

Dobson, I., J. Chen, et al. (Jan. 2002). “Examining Criticality of Blackouts in Power System Models with Cascading Events”. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. Big Island,HI.

Force, US-Canada Power System Outage Task (Apr. 2004). *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*.

Powell, L. (2004). *Power System Load Flow Analysis*. McGraw Hill Professional.

Amin, S. M. and B. F. Wollenberg (Sept. 2005). “Toward a Smart Grid: Power Delivery for the 21st Century”. In: *IEEE Power and Energy Magazine* 3.5, pp. 34–41.

Chen, J., J. S. Thorp, and I. Dobson (May 2005). “Cascading Dynamics and Mitigation Assessment in Power System Disturbances via a Hidden Failure Model”. In: *International Journal of Electrical Power & Energy Systems* 27.4, pp. 318–326.

Milano, F. (2005). “An open source power system analysis toolbox”. In: *IEEE Transactions on Power Systems* 20.3, pp. 1199–1206.

CAN/CSA-C22.2 No.257-06. Interconnecting Inverter-Based Micro-Distributed Resources To Distribution Systems (2006).

Cardillo, A. et al. (June 2006). "Structural properties of planar graphs of urban street patterns". In: *Phys. Rev. E* 73 (6), p. 066107.

Mackiewicz, R. E. (Oct. 2006). "Overview of IEC 61850 and Benefits". In: *2006 IEEE PES Power Systems Conference and Exposition*, pp. 623–630.

McQueen, M. A. et al. (2006). "Time-to-Compromise Model for Cyber Risk Reduction Estimation". In: *Quality of Protection: Advances in Information Security*. Ed. by D. Gollmann, F. Massacci, and A. Yautsiukhin. 1st ed. Boston, MA: Springer US, pp. 49–64.

Mell, P., K. Scarfone, and S. Romanosky (Nov. 2006). "Common Vulnerability Scoring System". In: *IEEE Security Privacy* 4.6, pp. 85–89.

Pourbeik, P., P. S. Kundur, and C. W. Taylor (Sept. 2006). "The Anatomy of a Power Grid Blackout - Root Causes and Dynamics of Recent Major Blackouts". In: *IEEE Power and Energy Magazine* 4.5, pp. 22–29.

CIGRE (Apr. 2007). "Security for Information Systems and Intranets for Electric Power Systems". In: *ELECTRA Tech. Brochure* 231.317, pp. 70–81.

Morgan, S. P. and K. R. Vixie (2007). "L^1TV computes the flat norm for boundaries". In: *Abstract and Applied Analysis* 2007, pp. 45153 1–14.

NIST (2007a). *National Software Reference Library*. http://www.nsrl.nist.gov. Last accessed 2 Dec 2020.

— (2007b). *National Vulnerability Database Version 2.2*. https://nvd.nist.gov/vuln-metrics/cvss/v2-calculator?name=CVE-2007-1001. Last accessed 2 Dec 2020.

Pepyne, David L. (June 2007). "Topology and cascading line outages in power grids". In: *Journal of Systems Science and Systems Engineering* 16, pp. 202–221.

Morgan, F. (2008). *Geometric Measure Theory: A Beginner's Guide*. fourth. Cambridge, MA, USA: Academic Press.

Panzieri, S. and R. Setola (May 2008). "Failure Propagation in Critical Interdependent Infrastructures". In: *International Journal of Modeling, Identification and Control* 3.1, pp. 69–78.

Phadke, A. G. and J. S. Thorpe (2008). *Synchronized Phasor Measurements and Their Applications*. Springer.

Ten, C., C. Liu, and G. Manimaran (Nov. 2008). "Vulnerability Assessment of Cybersecurity for SCADA Systems". In: *IEEE Transactions on Power Systems* 23.4, pp. 1836–1846.

PSRC, IEEE (2009). *Application of Overreaching Distance Relays*.

Putrus, G. A. et al. (2009). "Impact of Electric Vehicles on Power Distribution Networks". In: *2009 IEEE Vehicle Power and Propulsion Conference*, pp. 827–831.

Shao, S., M. Pipattanasomporn, and S. Rahman (2009). "Challenges of PHEV Penetration to the Residential Distribution Network". In: *2009 IEEE Power & Energy Society General Meeting*, pp. 1–8.

Stamp, J., A. McIntyre, and B. Ricardson (Mar. 2009). "Reliability Impacts from Cyber Attack on Electric Power Systems". In: *2009 IEEE/PES Power Systems Conference and Exposition*, pp. 1–8.

Wadhwa, C. L. (2009). *Electrical Power Systems*. New Age International.

Buldyrev, S. V. et al. (Apr. 2010). "Catastrophic cascade of failures in interdependent networks". In: *Nature* 464.1, pp. 1025–1028.

Clement-Nyns, K., E. Haesen, and J. Driesen (2010). "The Impact of Charging Plug-In Hybrid Electric Vehicles on a Residential Distribution Grid". In: *IEEE Transactions on Power Systems* 25.1, pp. 371–380.

Decker, I. C. et al. (2010). “System wide model validation of the Brazilian Interconnected Power System”. In: *IEEE PES General Meeting*. Minneapolis, MN, USA: IEEE, pp. 1–8.

Hines, P., S. Blumsack, et al. (2010). “The Topological and Electrical Structure of Power Grids”. In: *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10.

Hines, P., E. Cotilla-Sanchez, and S. Blumsack (2010). “Do topological models provide good information about electricity infrastructure vulnerability?” In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.3, p. 033122.

Parshani, R., S. V. Buldyrev, and S. Havlin (July 2010). “Interdependent Networks: Reducing the Coupling Strength Leads to a Change from a First to Second Order Percolation Transition”. In: *Physical Review Letters* 105.4, pp. 048701 1–4.

Pinar, A. et al. (2010). “Optimization Strategies for the Vulnerability Analysis of the Electric Power Grid”. In: *SIAM Journal on Optimization* 20.4, pp. 1786–1810.

Si, H. (2010). “Constrained Delaunay tetrahedral mesh generation and refinement”. In: *Finite Elements in Analysis and Design* 46 (1-2), pp. 33–46. ISSN: 0168-874X.

Wang, Z., A. Scaglione, and R. J. Thomas (2010). “Generating Statistically Correct Random Topologies for Testing Smart Grid Communication and Control Networks”. In: *IEEE Transactions on Smart Grid* 1.1, pp. 28–39.

Zamani, A., T. Sidhu, and A. Yazdani (July 2010). “A strategy for protection coordination in radial distribution networks with distributed generators”. In: *IEEE Power & Energy Society General Meeting*. Minneapolis, MN, USA: IEEE, pp. 1–8.

Boyd, S. et al. (Jan. 2011). “Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers”. In: *Foundation & Trends in Machine Learning* 3.1, pp. 1–122.

Chen, T. M., J. C. Sanchez-Aarnoutse, and J. Buford (Dec. 2011). “Petri Net Modeling of Cyber-Physical Attacks on Smart Grid”. In: *IEEE Transactions on Smart Grid* 2.4, pp. 741–749.

Chertkov, M., F. Pan, and M. G. Stepanov (Dec. 2011). “Predicting Failures in Power Grids: The Case of Static Overloads”. In: *IEEE Transactions on Smart Grid* 2.1, pp. 162–172.

Domingo, C. M. et al. (Feb. 2011). “A Reference Network Model for Large-Scale Distribution Planning With Automatic Street Map Generation”. In: *IEEE Transactions on Power Systems* 26.1, pp. 190–197.

Farkas, C., K. I. Szabó, and L. Prikler (2011). “Impact Assessment of Electric Vehicle Charging on a LV Distribution System”. In: *Proceedings of the 2011 3rd International Youth Conference on Energetics (IYCE)*, pp. 1–8.

Huang, X. et al. (June 2011). “Robustness of interdependent networks under targeted attack”. In: *Physical Review E* 83.6, pp. 065101 1–4.

Rotering, N. et al. (July 2011). “Medium-Voltage Network Planning with Optimized Power Factor Control of Distributed Generators”. In: *2011 IEEE Power and Energy Society General Meeting*. Detroit, MI, USA: IEEE, pp. 1–8.

Thompson, M. J. (Apr. 2011). “Percentage restrained differential, percentage of what?” In: *2011 64th Annual Conference for Protective Relay Engineers*, pp. 278–289.

Barrett, C. et al. (Oct. 2012). “Human Initiated Cascading Failures in Societal Infrastructures”. In: *Public Library of Science* 7.10, pp. 1–20.

Bernstein, A. et al. (Jan. 2012). “Sensitivity Analysis of the Power Grid Vulnerability to Large-scale Cascading Failures”. In: *SIGMETRICS Performance Evaluation Review* 40.3, pp. 33–37.

Brummitt, C. D., R. M. D’Souza, and E. A. Leicht (Mar. 2012). “Suppressing Cascades of Load in Interdependent Networks”. In: *Proceedings of the National Academy of Sciences* 109.12, E680–E689.

Cao, Y., S. Tang, et al. (2012). “An Optimized EV Charging Model Considering TOU Price and SOC Curve”. In: *IEEE Transactions on Smart Grid* 3.1, pp. 388–393.

Gan, L., U. Topcu, and S. H. Low (2012). “Stochastic distributed protocol for electric vehicle charging with discrete charging rate”. In: *2012 IEEE Power & Energy Society General Meeting*, pp. 1–8.

Kersting, W. H. (Jan. 2012). *Distribution System Modeling and Analysis*. Abingdon: CRC Press.

Liu, C. et al. (Jan. 2012). “Intruders in the Grid”. In: *IEEE Power and Energy Magazine* 10.1, pp. 58–66.

M. Govindarasu and C. Liu, S. Sridhar and (2012). “Risk Analysis of Coordinated Cyber Attacks on Power Grid”. In: *Control and Optimization Methods for Electric Smart Grids*. Ed. by A. Chakrabortty and M. D. Ili. 1st ed. New York, NY: Springer New York, pp. 275–294.

NERC (Apr. 2012). *NERC Reliability Standard TPL-001-1*.

Romero, J. J. (Oct. 2012). “Blackouts illuminate India’s power problems”. In: *IEEE Spectrum* 49.10, pp. 11–12.

Son, S. et al. (Jan. 2012). “Percolation theory on interdependent networks based on epidemic spreading”. In: *Europhysics Letters* 97.1, p. 16006.

Stefanov, A. and C. Liu (Oct. 2012). “ICT Modeling for Integrated Simulation of Cyber-Physical Power Systems”. In: *2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe)*, pp. 1–8.

Barrett, C. L. et al. (Nov. 2013). “Effects of Hypothetical Improvised Nuclear Detonation on the Electrical Infrastructure”. In: *International ETG-Congress 2013; Symposium 1: Security in Critical Infrastructures Today*, pp. 1–7.

Brummitt, C. D., P. D. H. Hines, et al. (2013). “Transdisciplinary electric power grid science”. In: *Proceedings of the National Academy of Sciences* 110.30, pp. 12159–12159.

Gonzalez-Sotres, L. et al. (Oct. 2013). “Large-Scale MV/LV Transformer Substation Planning Considering Network Costs and Flexible Area Decomposition”. In: *IEEE Transactions on Power Delivery* 28.4, pp. 2245–2253.

NERC (Apr. 2013). *Misoperations Report*.

Sharif, I., B. Krishnamoorthy, and K. R. Vixie (2013). “Simplicial flat norm with scale”. In: *Journal of Computational Geometry* 4.1, pp. 133–159.

Stefanov, A., C. Liu, et al. (Dec. 2013). “SCADA Modeling for Performance and Vulnerability Assessment of Integrated Cyberphysical Systems”. In: *International Transactions on Electrical Energy Systems* 25.3, pp. 498–519.

Wi, Y., J. Lee, and S. Joo (2013). “Electric Vehicle Charging Method for Smart Homes/buildings with a Photovoltaic System”. In: *IEEE Transactions on Consumer Electronics* 59.2, pp. 323–328.

Ahmed, M., B. T. Fasy, and C. Wenk (2014). “Local Persistent Homology Based Distance between Maps”. In: *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. SIGSPATIAL ’14. Dallas, Texas: ACM, pp. 43–52. ISBN: 9781450331319.

Bernstein, A. et al. (2014). “Power grid vulnerability to geographically correlated failures - Analysis and control implications”. In: *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*. Toronto, ON, Canada: IEEE, pp. 2634–2642.

D’Souza, R. M., C. D. Brummitt, and E. A. Leicht (2014). “Modeling Interdependent Networks as Random Graphs: Connectivity and Systemic Risk”. In: *Networks of Networks: The Last Frontier of Complexity*. Ed. by Gregorio D’Agostino and Antonio Scala. Springer International Publishing, pp. 73–94.

Dall’Anese, E. et al. (2014). “Decentralized Optimal Dispatch of Photovoltaic Inverters in Residential Distribution Systems”. In: *IEEE Transactions on Energy Conversion* 29.4, pp. 957–967.

IEEE Test Feeders (2014). https://cmte.ieee.org/pes-testfeeders/resources/. Last accessed 20 Oct 2021.

NERC (Dec. 2014). *NERC Staff Analysis of System Protection Misoperations*.

Onyeji, I., M. Bazilian, and C. Bronk (2014). "Cyber Security and Critical Energy Infrastructure". In: *The Electricity Journal* 27.2, pp. 52–60. ISSN: 1040–6190.

Pahwa, S., C. Scoglio, and A. Scala (Jan. 2014). "Abruptness of Cascade Failures in Power Grids". In: *Scientific Reports* 4.1, pp. 3694 1–9.

Soltan, S., D. Mazauric, and G. Zussman (June 2014). "Cascading Failures in Power Grids: Analysis and Algorithms". In: *Proceedings of the 5th International Conference on Future Energy Systems*. e-Energy '14. Cambridge, United Kingdom: ACM, pp. 195–206.

Wang, L. et al. (Jan. 2014). "k-Zero Day Safety: A Network Security Metric for Measuring the Risk of Unknown Vulnerabilities". In: *IEEE Transactions on Dependable and Secure Computing* 11.1, pp. 30–44.

Ahmed, M., B. T. Fasy, K. S. Hickmann, et al. (July 2015). "A Path-Based Distance for Street Map Comparison". In: *ACM Trans. Spatial Algorithms Syst.* 1.1.

Bolognani, S. and F. Dörfler (Sept. 2015). "Fast power system analysis via implicit linearization of the power flow manifold". In: *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. Monticello, IL, USA: IEEE, pp. 402–409.

Georgilakis, P. S. and N. D. Hatziargyriou (2015). "A review of power distribution planning in the modern power systems era: Models, methods and future research". In: *Electric Power Systems Research* 121, pp. 89–100.

Heidari, S., M. Fotuhi-Firuzabad, and S. Kazemi (2015). "Power Distribution Network Expansion Planning Considering Distribution Automation". In: *IEEE Transactions on Power Systems* 30.3, pp. 1261–1269.

Liu, M. et al. (June 2015). “Residential Electrical Vehicle Charging Strategies: The Good, The Bad and The Ugly”. In: *Journal of Modern Power Systems and Clean Energy* 3.2, pp. 190–202.

Liu, R. et al. (Sept. 2015). “Analyzing the Cyber-Physical Impact of Cyber Events on the Power Grid”. In: *IEEE Transactions on Smart Grid* 6.5, pp. 2444–2453.

NERC (Dec. 2015). *Analysis of System Protection Misoperations*.

Park, K. (2015). *Park’s Textbook of Preventive and Social Medicine, Twenty-third edition*. Bhanot Publishers.

Xu, H. (2015). *An Algorithm for Comparing Similarity Between Two Trees*. https://arxiv.org/abs/1508.03381. Last accessed 26 Sept 2022.

Zhang, Y., L. Wang, et al. (July 2015). “Power System Reliability Evaluation With SCADA Cybersecurity Considerations”. In: *IEEE Transactions on Smart Grid* 6.4, pp. 1707–1721.

Gegner, K. M. et al. (Feb. 2016). “A methodology for the creation of geographically realistic synthetic power flow models”. In: *2016 IEEE Power and Energy Conference at Illinois (PECI)*. Urbana, IL, USA: IEEE, pp. 1–6.

Kadavil, R., T. M. Hansen, and S. Suryanarayanan (July 2016). “An Algorithmic Approach for Creating Diverse Stochastic Feeder Datasets for Power Systems Co-simulations”. In: *2016 IEEE Power & Energy Society General Meeting (PESGM)*. Boston, MA, USA: IEEE, pp. 1–5.

NASEM (2016). *Analytic Research Foundations for the Next-Generation Electric Grid*. Washington, DC: The National Academies Press.

NERC (May 2016). *State of Reliability 2016*.

Pultarova, T. (Feb. 2016). “Cyber Security - Ukraine Grid Hack is Wake-up Call for Network Operators [News Briefing]”. In: *Engineering Technology* 11.1, pp. 12–13.

Soltan, S. and G. Zussman (2016). “Generation of synthetic spatially embedded power grid networks”. In: *2016 IEEE Power & Energy Society General Meeting (PESGM)*. Boston, MA, USA: IEEE, pp. 1–5.

Song, J. et al. (May 2016). “Dynamic Modeling of Cascading Failure in Power Systems”. In: *IEEE Transactions on Power Systems* 31.3, pp. 2085–2095.

Tang, W., S. Bi, and Y. J. Zhang (2016). “Online Charging Scheduling Algorithms of Electric Vehicles in Smart Grid: An Overview”. In: *IEEE Communications Magazine* 54.12, pp. 76–83.

Xie, J., A. Stefanov, and C. Liu (Mar. 2016). “Physical and Cyber Security in a Smart Grid Environment”. In: *Wiley Interdisciplinary Reviews: Energy and Environment* 5.5, pp. 519–542.

You, S. et al. (2016). “Co-optimizing generation and transmission expansion with wind power in large-scale power gridsImplementation in the US Eastern Interconnection”. In: *Electric Power Systems Research* 133, pp. 209–218.

Zhang, Y. and O. Yaan (June 2016). “Optimizing the robustness of electrical power systems against cascading failures”. In: *Scientific Reports* 6, pp. 27635 1–15.

Aziz, T. and N. Ketjoy (2017). “PV Penetration Limits in Low Voltage Networks and Voltage Variations”. In: *IEEE Access* 5, pp. 16784–16792.

Birchfield, A. B. et al. (Mar. 2017). “Statistical Considerations in the Creation of Realistic Synthetic Power Grids for Geomagnetic Disturbance Studies”. In: *IEEE Transactions on Power Systems* 32.2, pp. 1502–1510.

Brovelli, M. A. et al. (2017). “Towards an Automated Comparison of OpenStreetMap with Authoritative Road Datasets”. In: *Transactions in GIS* 21.2, pp. 191–206.

Gary, A. C. K. and U. N. Prananto (Oct. 2017). “Cyber Security in the Energy World”. In: *2017 Asian Conference on Energy, Power and Transportation Electrification (ACEPT)*, pp. 1–5.

Postigo, F. et al. (Nov. 2017). "A Review of Power Distribution Test Feeders in the United States and the Need for Synthetic Representative Networks". In: *Energies* 10.11, pp. 1896 1–14.

Schweitzer, E. et al. (June 2017). "Automated Generation Algorithm for Synthetic Medium Voltage Radial Distribution Systems". In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 7.2, pp. 271–284.

Subbiah, R. et al. (Feb. 2017). "Energy Demand Model for Residential Sector: A First Principles Approach". In: *IEEE Transactions on Sustainable Energy* 8.3, pp. 1215–1224.

Bai, Y. et al. (2018). *SimGNN: A Neural Network Approach to Fast Graph Similarity Computation*. https://arxiv.org/abs/1808.05689. Last accessed 26 Sept 2022.

Gonçalves, I., A. Gomes, and C. H. Antunes (2018). "Optimizing Residential Energy Resources with an Improved Multi-Objective Genetic Algorithm Based on Greedy Mutations". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. GECCO '18. Kyoto, Japan: ACM, pp. 1246–1253.

Khonji, M., S. C. Chau, and K. Elbassioni (2018). "Approximation Scheduling Algorithms for Electric Vehicle Charging with Discrete Charging Options". In: *Proceedings of the Ninth International Conference on Future Energy Systems*. e-Energy '18. Karlsruhe, Germany: ACM, pp. 579–585.

Meyur, R., A. Vullikanti, et al. (Dec. 2018). "Cascading Effects of Targeted Attacks on the Power Grid". In: *Complex Networks and Their Applications VII*. Springer, pp. 155–167.

Resener, M. et al. (Aug. 2018). "Optimization techniques applied to planning of electric power distribution systems: a bibliographic survey". In: *Energy Systems* 9.3, pp. 473–509.

Schäfer, B. et al. (May 2018). "Dynamically induced cascading failures in power grids". In: *Nature Communications* 9.1.

Thorve, S. et al. (Dec. 2018). “Simulating Residential Energy Demand in Urban and Rural Areas”. In: *2018 Winter Simulation Conference (WSC)*. Gothenburg, Sweden: IEEE, pp. 548–559.

Trpovski, A., D. Recalde, and T. Hamacher (Sept. 2018). “Synthetic Distribution Grid Generation Using Power System Planning: Case Study of Singapore”. In: *2018 53rd International Universities Power Engineering Conference (UPEC)*. Glasgow, UK: IEEE, pp. 1–6.

Wang, W. et al. (Nov. 2018). “An Approach for Cascading Effects within Critical Infrastructure Systems”. In: *Physica A: Statistical Mechanics and its Applications* 510.1, pp. 164–177.

Zhao, Y., Y. Chen, and B. Keel (Sept. 2018). “Optimal Scheduling of Home Energy Management System With Plug-In Electric Vehicles Using Model Predictive Control”. In: *Dynamic Systems and Control Conference*.

Atat, R. et al. (Oct. 2019). “Stochastic Geometry-Based Model for Dynamic Allocation of Metering Equipment in Spatio-Temporal Expanding Power Grids”. In: *IEEE Transactions on Smart Grid (Early Access)* 11.3, pp. 1–12.

Cimini, G. et al. (Jan. 2019). “The Statistical Physics of Real-World Networks”. In: *Nature Reviews Physics* 1.1, pp. 58–71.

Dey, A. K., Y. R. Gel, and H. V. Poor (2019). “What network motifs tell us about resilience and reliability of complex networks”. In: *Proceedings of the National Academy of Sciences* 116.39, pp. 19368–19373.

Dumas, M., B. Kc, and C. I. Cunliff (Aug. 2019). *Extreme Weather and Climate Vulnerabilities of the Electric Grid: A Summary of Environmental Sensitivity Quantification Methods*. https://info.ornl.gov/sites/publications/Files/Pub128663.pdf.

EPRI Test Circuits (2019). https://sourceforge.net/p/electricdss/code/HEAD/tree/trunk/Distrib/EPRITestCircuits/. Last accessed 20 Oct 2021.

Homeland Security, U.S. Dept. of (2019). *Electric Substations*. https://hifld-geoplatform.opendata.arcgis.com/datasets/electric-substations. Last accessed 20 Apr 2021.

Quiroga, D., E. Sauma, and D. Pozo (2019). "Power system expansion planning under global and local emission mitigation policies". In: *Applied Energy* 239, pp. 1250–1264.

Saha, S. S. et al. (2019). "A Framework for Generating Synthetic Distribution Feeders using OpenStreetMap". In: *2019 North American Power Symposium (NAPS)*. Wichita, KS, USA: IEEE, pp. 1–6.

Singh, M. K., V. Kekatos, and C. C. Liu (2019). "Optimal Distribution System Restoration with Microgrids and Distributed Generators". In: *2019 IEEE Power & Energy Society General Meeting (PESGM)*. Atlanta, GA, USA: IEEE, pp. 1–5.

Tantardini, M. et al. (Nov. 2019). "Comparing methods for comparing networks". In: *Scientific Reports* 9.1, p. 17557.

Weiss, M. and M. Weiss (May 2019). "An assessment of threats to the American power grid". In: *Energy, Sustainability and Society* 9.1.

Amin, B. M. et al. (2020). "Cyber attacks in smart grid dynamic impacts, analyses and recommendations". In: *IET Cyber-Physical Systems: Theory & Applications* 5.4, pp. 321–329.

ANSI (2020). *ANSI C84.1-2020: Electric Power Systems Voltage Ratings (60 Hz)*. https://blog.ansi.org/2020/10/ansi-c84-1-2020-electric-voltage-ratings-60. Last accessed on 13 Feb 2022.

Biswas, S., E. Bernabeu, and D. Picarelli (2020). "Proactive Islanding of the Power Grid to Mitigate High-Impact Low-Frequency Events". In: *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. Washington, DC, USA: IEEE, pp. 1–5.

Byeon, G. et al. (Apr. 2020). "Communication-Constrained Expansion Planning for Resilient Distribution Systems". In: *INFORMS Journal on Computing* 32.4, pp. 968–985.

Fatima, S., V. Püvi, and M. Lehtonen (2020). "Review on the PV Hosting Capacity in Distribution Networks". In: *Energies* 13.18.

Klemenjak, C. et al. (Apr. 2020). "A synthetic energy dataset for non-intrusive load monitoring in households". In: *Scientific Data* 7.1, p. 108.

Krishnan, V. et al. (Sept. 2020). "Validation of Synthetic U.S. Electric Power Distribution System Data Sets". In: *IEEE Transactions on Smart Grid* 11.5, pp. 4477–4489.

Lee, S. and D. Choi (Apr. 2020). "Energy Management of Smart Home with Home Appliances, Energy Storage System and Electric Vehicle: A Hierarchical Deep Reinforcement Learning Approach". In: *Sensors (Basel, Switzerland)* 20.7, p. 2157.

Lei, Shunbo et al. (2020). "Radiality Constraints for Resilient Reconfiguration of Distribution Systems: Formulation and Application to Microgrid Formation". In: *IEEE Transactions on Smart Grid* 11.5, pp. 3944–3956.

Li, H. et al. (2020). "Building Highly Detailed Synthetic Electric Grid Data Sets for Combined Transmission and Distribution Systems". In: *IEEE Open Access Journal of Power and Energy* 7, pp. 478–488.

Mateo, C., F. Postigo, F. de Cuadra, et al. (2020). "Building Large-Scale U.S. Synthetic Electric Distribution System Models". In: *IEEE Transactions on Smart Grid* 11.6, pp. 5301–5313.

Mateo, C., F. Postigo, F. De Cuadra, et al. (2020). *RNM-US Catalog of Equipment*. https://dx.doi.org/10.21227/4vne-pd37.

Meyur, R., M. Marathe, et al. (Dec. 2020). "Creating Realistic Power Distribution Networks using Interdependent Road Infrastructure". In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA: IEEE, pp. 1226–1235.

Morer, I. et al. (Apr. 2020). “Comparing Spatial Networks: A One-size-fits-all Efficiency-driven Approach”. In: *Phys. Rev. E* 101.4, pp. 042301 1–11.

Ok, S. (2020). “A graph similarity for deep learning”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Vancouver, BC, Canada: Curran Associates, Inc., pp. 1–12.

Richler, J. (July 2020). “Tell me something I don’t know”. In: *Nature Energy* 5.7, pp. 492–492.

Valdez, L. D. et al. (May 2020). “Cascading failures in complex networks”. In: *Journal of Complex Networks* 8.2.

Wang, Z., G. Chen, et al. (2020). “Cascading risk assessment in power-communication interdependent networks”. In: *Physica A: Statistical Mechanics and its Applications* 540, p. 120496.

Beers, L. M. (Nov. 2021). *Thousands of Victorian homes still without power a week after winds and storm*. https://7news.com.au/weather/melbourne-weather/thousands-of-victorian-homes-still-without-power-a-week-after-winds-and-storm-c-4426169.

Bidel, A., T. Schelo, and T. Hamacher (2021). “Synthetic Distribution Grid Generation Based on High Resolution Spatial Data”. In: *2021 IEEE International Conference on Environment and Electrical Engineering and 2021 IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe)*. Bari, Italy: IEEE, pp. 1–6.

Bistline, J. E. T. and G. J. Blanford (Dec. 2021). “The role of the power sector in net-zero energy systems”. In: *Energy and Climate Change* 2.

Blonsky, M., P. Munankarmi, and S. P. Balamurugan (2021). “Incorporating Residential Smart Electric Vehicle Charging in Home Energy Management Systems”. In: *2021 IEEE Green Technologies Conference (GreenTech)*, pp. 187–194.

Busby, J. W. et al. (2021). "Cascading risks: Understanding the 2021 winter blackout in Texas". In: *Energy Research & Social Science* 77, p. 102106.

Dominion Energy (Jan. 2021). *Off-Peak Plan - Time of Use Rate*. https://www.dominionenergy.com/virginia/rates-and-tariffs/off-peak-plan. Last accessed on 13 Feb 2022.

Fan, X. et al. (2021). "Model Validation Study for Central American Regional Electrical Interconnected System". In: *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*. Washington, DC, USA: IEEE, pp. 1–5.

Gaete-Morales, C. et al. (June 2021). "An open tool for creating battery-electric vehicle time series from empirical data, EMOBPY". In: *Scientific Data* 8.1, p. 152.

Guo, L. et al. (2021). "Line Failure Localization of Power Networks Part I: Non-Cut Outages". In: *IEEE Transactions on Power Systems* 36.5, pp. 4140–4151.

Hartmann, B. and V. Sugár (Mar. 2021). "Searching for small-world and scale-free behaviour in long-term historical data of a real-world power grid". In: *Scientific Reports* 11.1.

Hatziargyriou, N. et al. (2021). "Definition and Classification of Power System Stability Revisited & Extended". In: *IEEE Transactions on Power Systems* 36.4, pp. 3271–3281.

Joshi, S. et al. (Oct. 2021). "High resolution global spatiotemporal assessment of rooftop solar photovoltaics potential for renewable electricity generation". In: *Nature Communications* 12.1, p. 5738.

King, C.W., J.D. Rhodes, and J. Zarnikau (July 2021). *The Timeline and Events of the February 2021 Texas Electric Grid Blackouts*. https://energy.utexas.edu/sites/default/files/UTAustin%20%282021%29%20EventsFebruary2021TexasBlackout%2020210714.pdf.

Liang, M. et al. (2021). "FeederGAN: Synthetic Feeder Generation via Deep Graph Adversarial Nets". In: *IEEE Transactions on Smart Grid* 12.2, pp. 1163–1173.

Liu, Y. et al. (2021). “Searching for Critical Power System Cascading Failures With Graph Convolutional Network”. In: *IEEE Transactions on Control of Network Systems* 8.3, pp. 1304–1313.

Meyur, R. (2021). *Synthetic Distribution*. https://github.com/rounak-meyur/synthetic-distribution/tree/master/output. Last accessed 5 Apr 2022.

NERC (Aug. 2021). *State of Reliability: An Assessment of 2020 Bulk Power System Performance*. https://www.nerc.com/pa/RAPA/PA/Performance%20Analysis%20DL/NERC_SOR_2021.pdf.

Open Street Maps (2021). www.openstreetmap.org. Last accessed 20 Apr 2021.

Popovich, N. D. et al. (Nov. 2021). “Economic, environmental and grid-resilience benefits of converting diesel trains to battery-electric”. In: *Nature Energy* 6.11, pp. 1017–1025.

Republican Policy Committee (July 2021). *Infrastructure Cybersecurity: The US Electric Grid*. https://www.rpc.senate.gov/policy-papers/infrastructure-cybersecurity-the-us-electric-grid.

Riba, P. et al. (2021). “Learning graph edit distance by graph neural networks”. In: *Pattern Recognition* 120, p. 108132.

Schrom, E. et al. (2021). *Challenges in cybersecurity: Lessons from biological defense systems*. Last accessed 15 Aug 2022.

Sepulveda, N. A. et al. (May 2021). “The design space for long-duration energy storage in decarbonized power systems”. In: *Nature Energy* 6.5, pp. 506–516.

Shuvo, S. S. and Y. Yilmaz (2021). “CIBECS: Consumer Input Based Electric Vehicle Charge Scheduling for a Residential Home”. In: *2021 North American Power Symposium (NAPS)*, pp. 1–6.

Tong, K., A. S. Nagpure, and A. Ramaswami (2021). “All urban areas energy use data across 640 districts in India for the year 2011”. In: *Scientific Data* 8.

Wei, W. et al. (Jan. 2021). "Personal Vehicle Electrification and Charging Solutions for High-Energy Days". In: *Nature Energy* 6.1, pp. 105–114.

Cao, Y., H. Wang, et al. (2022). "Smart Online Charging Algorithm for Electric Vehicles via Customized ActorCritic Learning". In: *IEEE Internet of Things Journal* 9.1, pp. 684–694.

ESRI (2022). *Georeferencing a raster to a vector*. https://desktop.arcgis.com/en/arcmap/latest/manage-data/raster-and-images/georeferencing-a-raster-to-a-vector.html.

Feuerstein, J. (May 2022). *Deadly thunderstorm complex cuts power to nearly a million in Canada*. https://www.washingtonpost.com/weather/2022/05/22/canada-storm-derecho-ontario-quebec/.

Majhi, S. and C. Wenk (2022). *Distance Measures for Geometric Graphs*. https://arxiv.org/abs/2209.12869. Last accessed 31 Oct 2022.

Paaβen, B. (2022). *Revisiting the tree edit distance and its backtracing: A tutorial*. https://arxiv.org/abs/1805.06869. Last accessed 26 Sept 2022.

Singh, M. K., S. Taheri, et al. (July 2022). "Joint grid topology reconfiguration and design of Watt-VAR curves for DERs". In: *IEEE Power & Energy Society General Meeting*. Denver, CO, USA: IEEE.