Data analysis of cascading outages using historical data to mitigate blackouts

by

Kai Zhou

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Electrical Engineering (Electric Power and Energy Systems)

Minor: Statistics

Program of Study Committee:
Ian Dobson, Major Professor
Zhaoyu Wang, Major Professor
Arka P. Ghosh
Aditya Ramamoorthy
Jarad Niemi

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

I would first like to sincerely thank my advisors Dr. Ian Dobson and Dr. Zhaoyu Wang. To Dr. Ian Dobson, thank you for your guidance, encouragement, patience, and support during my research and writing this thesis. Your countless hours of professional mentorship enable my continued learning and growth at Iowa State University. You are one of those that genuinely love helping others and sharing his life experience and wisdom as a researcher. To Dr. Zhaoyu Wang, thank you for your support during my research and for bringing me to Iowa State so that I learned from and met with many ingenious people.

I would also like to gratefully acknowledge funding from NSF grants 1609080 and 1735354, making my research possible. I would also like to thank BPA for making the outage data public.

To my committee members, Dr. Arka P. Ghosh, Dr. Jarad Niemi, and Dr. Aditya Ramamoorthy, thank you for helping and teaching me at different stages of the P.h.D program. And from your courses, I learned skills and fundamental knowledge about probability and statistics for my research.

To all my collaborators, thank you for your assistance on the research projects so that the papers were published with high quality.

To my friends and peers, thank you for making my time a wonderful experience at Iowa State University.

To my parents and uncles, thank you for love, encouragement, and frequent calls to inquire about my health and progress.

ABSTRACT

Cascading in an electric power transmission system is a sequence of dependent outages that successively weakens the transmission system. It is caused by initial outages and then propagates as a series of dependent outages. Cascading is one of the main causes of blackouts in high voltage transmission networks. Utilities are routinely collecting outage data. This work proposes statistical methods that are applied to real outage data. It aims to understand the propagation of cascading outages for blackout mitigation and resilience evaluation and study the distribution of initial outages for reliability analysis and contingency selection.

A Markovian influence graph model is formed from historical outage data. This Markovian influence graph defines a Markov chain and generalizes the previous influence graph by including multiple line outages as Markov chain states. It describes the transition probabilities between generations of cascading outages. The Markovian influence graph reproduces the distribution of the cascade size in the data and estimates the probabilities of small, medium, and large cascades. The key advantage of the Markovian influence graph is that it allows the mitigation effects to be analyzed and readily tested, which is not available from the historical data. The asymptotic property of the Markov chain indicates the critical lines that are most involved in the propagation of large cascades. Upgrading these critical lines will reduce the probability of large cascades.

Extreme events can damage power system components and then cause cascading outages. Methods are needed to evaluate the cascading phase of resilience. The Markovian influence graph can simulate cascading line outages that follow initial outages from extreme events by an improved sampling method. It efficiently produces large cascade samples. Thus, we can better estimate the large cascades that are rare but significant for cascade resilience.

The Markovian influence graph is validated by two tests. As mitigation results are not easily extracted from historical outage data, simulation is indicated to further test the Markovian

influence graph. The test forms the Markovian influence graph from simulated cascades before mitigation, calculates the mitigation effect using the influence graph, and compares this computed mitigation effect with simulated mitigation effect. It uses several different cascading models on several different power systems. Moreover, the Markovian influence graph assumes current line outages depend only on preceding line outages. This assumption is tested by comparing the influence graph with the kinetic Monte Carlo (KMC) cascading simulation which is also a Markov chain but depends all line outages before current line outages.

Transmission line outage rates are foundation for many reliability calculations. However, line outages are rare, occurring only about once per year. A Bayesian hierarchical model is proposed to mitigate the limited data problem. This Bayesian hierarchical model leverages line dependencies to better estimate individual transmission line outage rates. The Bayesian estimator produces more accurate estimates of individual line outage rates and the uncertainty of these estimates. Better estimates of individual line outage rates using the Bayesian hierarchical model benefit the reliability calculation. Three applications are illustrated: detect lines with increased outage rates, quantify outage rates for specific causes, and discuss the effect of outage rate uncertainty on a simple availability calculation.

Multiple contingencies show different patterns in the graph representation of a power grid. The analysis of the historical outage data reveals that multiple contingencies occur frequently in contingency motifs, which are subgraphs that occur significantly more frequently than random subgraphs in the given power network. Based on this finding, this study proposes a probabilistic model to estimate the probability of multiple contingencies and a corresponding contingency selection scheme. The contingency selection scheme is much more efficient than randomly selecting contingencies. Moreover, the analysis reveals that the diameter of contingency subgraphs, the maximum network distance between any two lines in the subgraphs, follows a Zipf distribution.

CHAPTER 1. INTRODUCTION

1.1 Background

Cascading in an electric power transmission system is a sequence of dependent outages that successively weakens the transmission system. [2]. It is triggered by initial events and then propagates. One useful way to study cascading phenomena on the power transmission system is to record the sequences of lines outages. Some cascades of line outages, especially the longer ones, will result in a blackout (significant amounts of load shed), whereas others do not result in load shedding and can be regarded as precursors to a blackout. These blackouts are infrequent but high-impact events that occur often enough to pose a substantial risk to society [3, 4]. From 2008 to 2018 in the US, 10 million or more people were affected due to the power outages, as surveyed in [5].

The cascading outage is one of the major causes of blackouts. Ekisheva in [6] analyzed the historical outage data in the North American bulk power system, and found that groups of outages overlapping in time were initialized by many causes and propagated because of power system conditions, such as overloading, voltage problems, and bad weather.

Utilities in many countries are routinely collecting outage data. The Bonneville Power Administration (BPA) has made its outage data publicly available [7]. The North American Electric Reliability Corporation (NERC) has been collecting transmission element outages since 2008 through the Transmission Availability Data System (TADS) [6]. [8] summarised outage data systems developed in different countries. These outage data collecting entities include but are not limited to the Mid-Atlantic Power Pool since 1997, the Western Electricity Coordinating Council (WECC) since 2006, the Canadian Electricity Association's Equipment Reliability Information System since 1980, the ENTSO-E Regional Group Nordic in Europe since 1990, and the Idaho Power Company since 1991.

Cascading outages involve various mechanisms. [5] summarizes causes of recorded severe power outages around the world. Also, by inspecting the outage data recorded by BPA, dozens of outage causes are recorded (Table 7.3), and the main causes are lightning, terminal equipment failure, and unknown causes from inside and outside of the BPA area.

1.2 Motivation

Model-based simulation is widely used in studying cascading outages. However, this simulation method is practically limited to approximating a subset of cascading mechanisms, and the cascading model in the simulation has often not been well benchmarked and validated for estimating the blackout risk [9, 10]. Historical outage data encompass all the information about cascading mechanisms during the observed period. The power industry has always analyzed specific blackouts and taken steps to mitigate cascading. However, and especially for the largest blackouts of the highest risk, the challenges of evaluating and mitigating the cascading risk in a quantitative way remain. This motivates us to analyze real outage data for the cascading risk by data-driven methods.

A perennial problem of using real outage data is that outages are rare events, and the outage data is limited. On average, each transmission line outages once per year. Effective methods are needed to address the sparse outage data problem.

Components in the power system are not equally important, and some of them are more critical than others in terms of cascading risks. The power system is a complex network comprising hundreds and thousands of components such as transmission lines, transformers, buses, and others. Not all components have equal importance. Indeed, only a small set of system components contributes to a large blackout [11]. This implies that we can upgrade a small number of critical components to reduce cascading risks by a large proportion. Also, it is practical to upgrade an only limited number of components within a budget. Therefore, identifying critical components is significant to mitigate cascading blackouts.

Cascading line outages are comprised of initial outages and propagating outages. Transmission lines in initial outages and propagating outages have different influences on cascading outages. We should identify critical lines in initial outages and propagating outages, respectively.

The significance of studying initial line outages is twofold. For one thing, line outage rates are foundation for many reliability calculations; for another thing, the analysis of initial outages benefits the power system security analysis, especially contingency selection.

1.3 Research Objective

The research objective is to develop statistical methods that are applied to real outage data to understand the propagation of cascading outages and the distribution of initial outages. Specifically,

- use observed transmission line outage data to make a Markovian influence graph that describes the probabilities of transitions between generations of cascading line outages. This generalized influence graph is used to reproduce the distribution of cascade sizes in outage data and evaluate the mitigation effect. The asymptotic property of the Markovian influence graph is used to identify the critical lines that are most involved in the propagation of large cascades.

- test two aspects of the Markovian influence graph: the large cascade mitigation and the model assumption.

- have better estimates of outage rates for individual transmission lines by proposing a Bayesian hierarchical model. The Bayesian hierarchical model incorporates prior information of transmission lines and leverages transmission line dependencies to mitigate the limited data problem. The Bayesian estimator produces the uncertainty of outage rates.

- explore the applications of the Markovian influence graph driven by historical data and the Bayesian estimates of individual transmission line outage rates.

- study the spatial characteristics of initial outages and form a systematic contingency selection scheme.

CHAPTER 2. LITERATURE REVIEW

2.1 Influence graphs for power system cascading outages

This section reviews the previous literature on influence graphs for power grid cascading outages and related topics. There is increasing interest in graphs to represent cascading outages, in which the graph describes the interaction between outaged components and is not the power grid topology. These graphs of interactions have differences in how they are formed and have different names, such as the influence graph, the interaction graph, the correlation network, and the cascading faults graph. The idea of a graph of interactions can be traced back to [12] which has a stochastic process at each graph node that interacts with different strengths along the graph edges joining that node to other nodes. Rahnamay-Naeini [13] generalizes the model of interacting and cascading nodes in [12] to include interactions within and between two interdependent networks. This type of interacting particle system model has some nice properties allowing analysis, but remains a somewhat abstract model for power system cascading because it is not known how to estimate the model parameters from data.

Influence graphs in their present form were introduced by Hines and Dobson [14], and further developed by Qi, Hines, and Dobson [15, 16]. These influence graphs describe the statistics of cascading data with networks whose nodes represent outages of single transmission lines and whose directed edges represent probabilistic interactions between successive line outages. The more probable edges correspond to the interactions between line outages that appear more frequently in the data. Cascades in the influence graph start with initial line outages at the nodes and spread probabilistically along the directed graph edges. Once the influence graph is formed from the simulated cascading data, it can be used to identify critical components and test mitigation of blackouts by upgrading the most critical components [15–17].

As well as outages of single lines, cascading data typically includes multiple line outages that occur nearly simultaneously. When the states are single line outages, these multiple simultaneous outages cause problems in obtaining well-defined Markov chain transitions between states. For example, if the outage of two lines causes an outage in the next generation, it is hard to tell which line caused the subsequent outage or whether the two lines caused the subsequent outage together. To address this, [16] assigns an equal share to the two lines. The resulting influence graph is then approximated to enable analysis. Qi [15] assumes that the subsequent outage is caused by the most frequent line outage. Improving on this assumption, Qi [18] considers the causal relationships among successive outages as hidden variables and uses an expectation maximization algorithm to estimate the interactions underlying the multiple outage data. This work solves this problem in a novel way by defining an additional state for each multiple line outage. Thus the new influence graph generalizes the interaction between single lines to multiple line outages, so we do not need to make assumptions or approximations when calculating the interactions between two single lines. This enables a Markov chain to be cleanly and clearly defined.

Considering the different types of graphs of interactions more generally, there are three methods of quantifying interactions between components which are the edges in the graph of interactions. First, as explained in the preceding paragraph, in [14–16], the edge corresponds to the conditional probability of a single line outage given that the previous line has outaged. Second, in [19–21], the edge weight is calculated based on the line flow changes due to a single line outage applied to the base case using a DC load flow (In contrast to [14–16] and this work, this implies that the edge weights do not change during the cascade.). In Merrill [20], the edge weight is obtained from the line outage distribution factors. In Zhang [19] and Ma [21], the directed edge weights are obtained from both the line flow changes and the remaining margin in the line the power is transferred to. Then Zhang [19] combines the directed edges to give undirected edges. On the other hand, Ma [21] retains the directed edges and also represents hidden failures by additional nodes. Third, in Yang [22], the edge corresponds to the correlation between any two lines. In [23], Carreras constructs a synchronization matrix from simulation data from the OPA

model to identify the lines with higher overloading probabilities. Other papers [17, 18, 24–26] form their graph of interactions similarly to the above methods. This work bases the influence graph edges on conditional probabilities. However, the edges are different than the edges in [14–16] as they directly correspond to transition probabilities in a rigorously defined Markov chain.

Influence graphs describing the interactions between successive cascading outages were developed using simulated data (Zhou [17] is an exception, but [17] differs from this work because it applies the methods of [16] to utility data). But even for simulated cascade data, there remain challenges in extracting good statistics for the influence graph from limited data. Hines, Dobson and Qi [14–16] estimate the conditional probabilities of transitions with empirical probabilities. This work mitigates the limited historical cascading data by using a Bayesian method and carefully combining the sparser data of the later stages of cascading in a sophisticated way.

Various measures are proposed for the identification of critical components based on the influence graph. [15, 16, 21, 27] form their specific measures based on their own influence/interaction graph. Ma [21] uses a modified page-rank algorithm to find critical lines. Nakarmi [24] forms the influence graph using methods of both [16] and [22], and proposes a community-based measure to identify critical components. [24] compares its measure with other centrality measures based on network theory, and concludes that its method performs better than other methods in most cases. In this work, the new influence graph is a rigorous Markov chain, and the identification of critical lines is based on the asymptotic quasi-stationary distribution. The quasi-stationary distribution has a clear interpretation of specifying the probabilities that each of the lines is involved in large cascades.

The graph of interactions also provides useful information about mitigation actions in power system operation. Ju [25] extends the interaction graph to a multi-layer graph, in which the three layers reflect the number of line outages, load shed, and electrical distance of the cascade spread, respectively. This multi-layer graph is suggested to mitigate cascades in system operation by providing the critical lines at different states of cascades. Chen [26] proposes a dynamic interaction graph to better support online mitigation actions than a static interaction graph.

During the propagation of a specific cascade, this dynamic interaction graph removes the interactions involving already outaged lines, and optimal power flow controls the power flow on the critical lines indicated by the dynamic interaction graph. The dynamic interaction graph model reduces the risk of large cascades more than the static interaction graph.

As expected, the graph of interactions and any conclusions drawn depend on the outage data from which the graph is formed. If the outage data is simulated, the selection of initial system states matters. For example, Nakarami [24] shows that different system states lead to different influence graphs. This work forms our influence graph from fourteen years of public outage data of a specific area, so that our influence graph reflects the initial faults and states encountered over that period of time in that power system area. The textbook [28] includes material on both influence and interaction graphs.

Another related line of research is fault chains. A fault chain as described in [29] is one cascading sequence of line outages. Each initial line fault gives a fault chain of lines most stressed at each step until outage or instability. Usually only the most stressed or most likely next line outage is selected to form fault chains. By taking each line in the system as the initial outage of each fault chain, Wei [27] obtains a set of fault chains using a branch loading index to select the most stressed next line to outage. Each fault chain is expressed as a subgraph whose nodes are transmission lines, and directed edges are branch loading assessment indexes, and the union of the subgraphs forms a cascading faults graph. The edge weights depend on the sum of the branch loading indices, each scaled by the length of the fault they are in. Then critical lines are identified according to the in- or out-degree of the cascading faults graph. Luo [30] also forms a cascading faults graph with weights depending on load loss in the chain, and then uses hypertext-induced topic search to select critical lines. The edge weights of [27, 30] differ from those in influence graphs because they are not based on conditional probabilities. Li and Wu [31] combine simulated fault chains into a network and use reinforcement learning to explore, evaluate, and find chains most critical to load shed. In further work, Li and Wu [32] combine simulated fault chains into a state-failure network from which expected load shed can be computed for each state and failure

by propagating load shed backwards accounting for the transition probabilities of the edges. The transition probabilities are estimated similarly to an influence graph by the relative frequency of that transition at that stage of the data. However, in contrast to the practice in influence graphs, the state transition data for the later stages is not combined together to get better estimates. Moreover, fault chains differ from this work in only considering single line outages one after another.

There are also approaches to modeling cascading with continuous-time Markov processes. Wang [33] drives line loadings with generator and load power fluctuations to determine overloads and outages that change the Markov state and hence simulate the cascading.
Rahnamay-Naeini [34] constructs, using simulated cascading data and fitted functional forms, a Markov process with states highly aggregated into 3 quantities, namely the number of failed lines, the maximum of the capacities of all of the preceding failed lines, and a cascade stopping index. The aggregated Markov process can model the time evolution of the cascade and the distribution of cascade size. In further work, Rahnamay-Naeini reduces the aggregated model to a discrete time Markov chain and generalizes it to model cyber and power interdependent network cascading interactions in [35] and to model operator actions interacting with cascading in [36].

For another, independent perspective on the literature, Nakarmi's review paper [37] surveys various methods of constructing interaction graphs and the reliability analysis based on interaction graphs.

2.2 Simulation of cascading outages

There is a large literature on model-based simulation of cascading (reviewed in [2, 38]), and substantial literature on influence or interaction graphs and fault chains driven by simulated data that is discussed above (also reviewed in [37, 39]). To our knowledge, the only previous work on influence graphs driven by real data is [17, 39, 40].

The need for higher-level statistical simulation of cascading arose from broader studies of the multiple phases of resilience. For example, Romero [41] optimized investments to improve

resilience to earthquakes, and discussed but did not model the cascading phase of resilience. Recently Kelly-Gorham [42, 43] proposed a high-level statistical method driven largely by observed statistics called CRISP to quantify power transmission system overall resilience in all its phases. CRISP models the cascading phase of resilience by sampling from a probability distribution of the total number of lines out based on historical data. Then, in [43], given the number of lines out, the lines outaged in a cascade are chosen in accordance with an observed probability distribution [1] of network distance between cascaded line outages.

2.3 Estimation of individual transmission line outage rates

Bayesian approaches encode uncertainty in uncertain parameters such as outage rates as random variables. The Bayesian analysis aims to estimate a probability distribution for the uncertain parameters by incorporating all of our knowledge and accurately reflecting the uncertainty. Bayes' theorem is used to combine data with prior distributions that describe the initial knowledge of uncertainty. The prior distributions are updated with available data to give a posterior distribution that describes the uncertainty in the parameter values given all the available data. The mean or mode of the posterior distribution can be used to give a point estimate of the parameter. For further detail explaining Bayesian methods we suggest [44] as an introduction and [45] as a reference.

Bayesian methods are ideal for problems with limited data (such as the estimation of outage rates), where it is necessary to use all the information available. Studies in ecology and social science have shown that when data are limited, Bayesian methods have less bias and are more robust than frequentist methods that consider parameters as fixed values [46, 47]. When lots of data are available, the data outweighs any effect of the prior distributions and a Bayesian method is less advantageous.

There is previous research predicting outage rates using Bayesian methods. Li [48], Iešmantis [49], and Moradkhani [50] present three Bayesian hierarchical models. All three hierarchical models have a Poisson distribution for outage counts, but how the outages are

counted and lower levels of the model are different. Li [48] develops a hierarchical model to predict outage counts in a substation district given weather conditions, in which the log of the outage rate is a linear combination of weather factors. Iešmantis [49] presents a Poisson-Gamma random field model to estimate 230 kV transmission lines outage rates in a specified rectangular cell. The grid cells are introduced to model spatial dependence by constructing a correlation matrix in the Gamma field. The hierarchical model in Moradkhani [50] estimates failure rates of individual overhead distribution feeders, which are assumed to be independent of each other. To have an analytical form for the posterior distribution, conjugate priors are used, which results in a Gamma posterior distribution. Bayesian networks are also applied to estimate outage rates. Zhou [51] proposes a simple Bayesian network to predict weather-related outage rates given lightning and wind conditions over the whole system. Zhou compares the Bayesian network with a Poisson regression model and concludes that the Bayesian network is preferable. Yang [52] gives interval estimates of outage rates of individual transmission lines given weather conditions using a credal network with imprecise priors, which is an extension of Bayesian networks. Dunn [53] formulates a Bayesian hierarchical model for the total outage counts in a system. All components share the same failure rate derived from a fragility curve. In contrast to all the references above, this work estimates outage rates of individual transmission lines using a Bayesian hierarchical model considering line dependencies.

Transmission line outages are correlated with each other in several ways. Lines in the power grid interconnect at substations, and some faults or substation arrangements may trip several lines simultaneously. Multiple line outages also occur due to protection schemes such as control protection groups and remedial action schemes. Moreover, lines in the same area experience similar weather conditions. There is some previous work on these dependencies. Li [48] uses the network adjacency matrix to model district dependencies. Similarly, Dokic [54] uses the weighted adjacency matrix to model substation dependencies. The difference between them is that [48] models the dependencies as a covariance matrix from the Bayesian perspective, while [54] uses an embedding method by learning vector representations of dependencies from a frequentist

perspective. Iešmantas [49] models geographical dependencies between the outage rate per kilometer of 230 kV lines by making a rectangular grid of the area. Portions of lines in the same rectangle are assumed to have the same geographical influence, and the correlation between lines in different rectangles is assumed and modeled in the Gamma field. The main conclusion of [49] is that geographical correlation between line outage rates is present but weak. However, our method captures partial similarities between lines, including proximity, length, and rated voltage as a layer in the Bayesian hierarchical model.

Many researchers focus on predicting outage probabilities in a short term according to the weather condition [48, 51, 52, 54–57]. [50, 52, 57] consider the data deficiency when building the outage rate model.

2.4 Application of network motifs in power systems

Network motifs are first introduced by U. Alon, R. Milo and their group in gene regulation networks [58, 59]. They are recurrent and statistically significant subgraphs of a network. The network motif is widely used in gene regulation networks in systems biology and successfully used in ecological, sociological, and epidemiological networks [60]. There are also several studies on the utility of the network motif in the power system context. Ren et. al. have proposed using the network motif as an indicator of the cascading outage risk [61]. They show that cascading outages exhibit three phases as the load level increases, and the phases correspond to the decrease of the frequency of network motifs. The frequency of motifs reflects the connectivity of power grid, hence, it can be a warning sign of the cascading outage risk [61]. Other researchers have studied the network motif as an indicator of power grid robustness and reliability [62–65]. The work uses techniques in network science. Specifically, they carry out attacks on the power grid by removing nodes according to some order and monitor changes of network motif properties such as concentration, z-score and lifetime. Then, they determine the robustness and reliability of the network based on the idea that the robust network tends to preserve longer its motif-based measurements.

Previous work on network motif applications in power systems uses the definition of motifs by Milo. Milo defines network motifs as connected subgraphs in a network that occur in significant higher number than in randomized networks. However, this definition is not well fitted to contingency selection because the power network is fixed and known, and multiple contingencies could also be disconnected subgraphs [1]. Specifically, we represent multiple contingencies as subgraphs of the power network. Some subgraphs appear significantly more frequently than their random occurrence in the specific power network. We define them as contingency motifs. Furthermore, some subgraphs are disconnected subgraphs, and to what extent their components are separated varies from subgraphs to subgraphs and follows a certain distribution.

[1] A subgraph is disconnected if at least two nodes are not connected by a sequence of lines

CHAPTER 3. A MARKOVIAN INFLUENCE GRAPH FORMED FROM OUTAGE DATA

This chapter forms a Markovian influence graph from historical outage data. This Markovian influence graph is a rigorously defined Markov chain. The transition probabilities describe the influences between generations in cascades. This Markov chain reproduces the distribution of cascade sizes and estimates the probabilities of small, medium, and large cascades. The asymptotic property of the Markov chain indicates critical lines in propagation of cascades. The mitigation effect of upgrading these critical lines can be readily tested by the Markovian influence graph.

This chapter is developed with assistance from Arka P. Ghosh, the Department of Statistics, Iowa State University, and Alexander Roitershtein, Texas A&M University. The material in this chapter is published in [39].

3.1 Introduction

There are two main approaches to evaluating cascading risk: simulation and analyzing historical utility data. Cascading simulations can predict some likely and plausible cascading sequences [2, 66]. However, only a subset of cascading mechanisms can be approximated, and simulations are only starting to be benchmarked and validated for estimating blackout risk [9, 10]. Historical outage data can be used to estimate blackout risk [4] and detailed outage data can be used to identify critical lines [67]. However it is clear that proposed mitigation cannot be tested and evaluated with historical data. This work processes historical line outage data to form a Markovian influence graph that statistically describes the interactions between the observed outages. The Markovian influence graph can quantify the probability of different sizes of cascades,

identify critical lines, and assess the impact of mitigation on the probability of different sizes of cascades.

3.2 Forming the Markovian influence graph from historical outage data

We use detailed historical line outage data consisting of records of individual automatic transmission line outages that specify the lines outaged and the outage times to the nearest minute. We emphasize that this data is routinely recorded by utilities worldwide, for example in the North American Transmission Availability Data System.

The first step in building an influence graph is to take many cascading sequences of transmission line outages and divide each cascade[1] into generations of outages as detailed in [68]. Each cascade starts with initial line outages in generation 0, and continues with subsequent generations of line outages 1,2,3,... until the cascade stops. Each generation of line outages is a set of line outages that occur together on a fast time scale of less than one minute. Often there is only one line outage in a generation, but protection actions can act quickly to cause several line outages in the same generation. (Sometimes in a cascading sequence an outaged line recloses and outages in a subsequent generation. In contrast to [17, 68], here we neglect the repeats of these outages.)

The influence graph represents cascading as a Markov chain $X_0, X_1, ...$, in which X_k is the set of line outages in generation k of the cascade. We first illustrate the formation of the influence graph from artificial cascading data with the simple example of four observed cascades involving three lines shown in Fig. 3.1. The first cascade has line 1 outaged in generation 0, line 3 outaged in generation 1, line 2 outaged in generation 2, and then the cascade stops with no lines (indicated by the empty set {}) outaged in generation 3. All cascades eventually stop by transitioning to and remaining in the state {} for all future generations. The five states observed in the data are {}, {line 1}, {line 2}, {line 3}, and {line 1, line 3}, where this last state is lines 1

[1] The grouping of line outages into cascades uses the simple method of [68]: The grouping is done by looking at the gaps in start time between successive line outages. If successive outages have a gap of one hour or more, then the outage after the gap starts a new cascade. More elaborate methods of grouping real line outages into cascades could be developed and applied.

cascade number	generation 0 X_0	generation 1 X_1	generation 2 X_2	generation 3 X_3
1	{line 1}	{line 3}	{line 2}	{}
2	{line 2}	{line 1, line 3}	{}	{}
3	{line 3}	{line 1}	{}	{}
4	{line 1}	{}	{}	{}

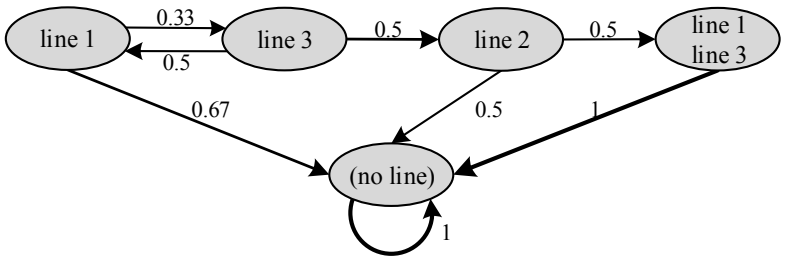


Figure 3.1: Simple example forming influence graph from artificial data (the influence graph formed from real utility data is shown in Fig. 3.2).

and 3 outaging together in the same generation, as in generation 1 of cascade 2. Introducing the state {line 1, line 3} with two line outages avoids the problems in previous work in accounting for transitions to and from the simultaneous outages of line 1 and line 3.

We can estimate the probabilities of transitioning from state i to state j in the next generation by counting the number of those transitions in all the cascades and dividing by the number of occurrences of state i. For example, the probability of transitioning from state {line 1} to state {line 3} is 1/3 and the probability of transitioning from state {line 2} to state {line 1, line 3} is 1/2. The probability of transitioning from state {line 1} to {}; that is, stopping after the single outage of line 1, is 2/3. The probabilities of the edges out of each state sum to 1. By working out all the transition probabilities, we can make the network graph of the Markov chain as shown in Fig. 3.1. The transitions between states with higher probability are shown with thicker lines. In this generalized influence graph, nodes are sets of line outages and edges indicate transitions or interactions between sets of line outages in successive generations of cascading. The influence graph is different than the physical grid network and cascades are generated in the influence graph by moving along successive edges, selecting them according to their transition probabilities.

In the general case, there are many states s_0, s_1, ... , and we describe the transitions between them. Let $\boldsymbol{P}_k$ be the Markov chain transition matrix for generation k. The $\boldsymbol{P}_k$ matrix entry $P_k[i,j]$ is the conditional probability that the set of outaged lines is s_j in generation $k+1$, given that the set of outaged lines is s_i in generation k; that is,

$$P_k[i,j] = \mathrm{P}[X_{k+1} = s_j \mid X_k = s_i]. \tag{3.1}$$

The key task of forming the Markov chain is to estimate the transition probabilities in the matrix $\boldsymbol{P}_k$ from the cascading data. If one supposed that $\boldsymbol{P}_k$ does not depend on k, a straightforward way to do this would first construct a counting matrix $\boldsymbol{N}$ whose entry $N[i,j]$ is the number of transitions from s_i to s_j among all generations in all cascades. Then $\boldsymbol{P}_k$ would be estimated as

$$P_k[i,j] = \frac{N[i,j]}{\sum_j N[i,j]}. \tag{3.2}$$

However, we find that $\boldsymbol{P}_k$ must depend on k in order to reproduce the increasing propagation of outages observed in the data [68]. On the other hand, there is not enough data to accurately estimate $\boldsymbol{P}_k$ individually for each $k > 0$. Our solution to this problem involves both grouping together data for higher generations and having $\boldsymbol{P}_k$ varying with k, as well as using empirical Bayesian methods to improve the required estimates of cascade stopping probabilities. The detailed explanation of this solution is postponed to section 3.6, and until section 3.6 we assume that $\boldsymbol{P}_k$ has already been estimated for each generation k from the utility data. Forming the Markov chain transition matrix from the data in this way makes the Markovian assumption that the statistics of the lines outaged in a generation only depend on the lines outaged in the preceding generation. This is a pragmatic assumption that yields a tractable data-driven probabilistic model of cascading.

One way to visualize the influence graph interactions between line outages in $\boldsymbol{P}_k$ is to restrict attention to the interactions between single line states, and show these as the red network in Fig. 3.2. The gray network is the actual grid topology, and the gray transmission lines are joined by a red line of the thickness proportional to the probability of being in successive generations, if

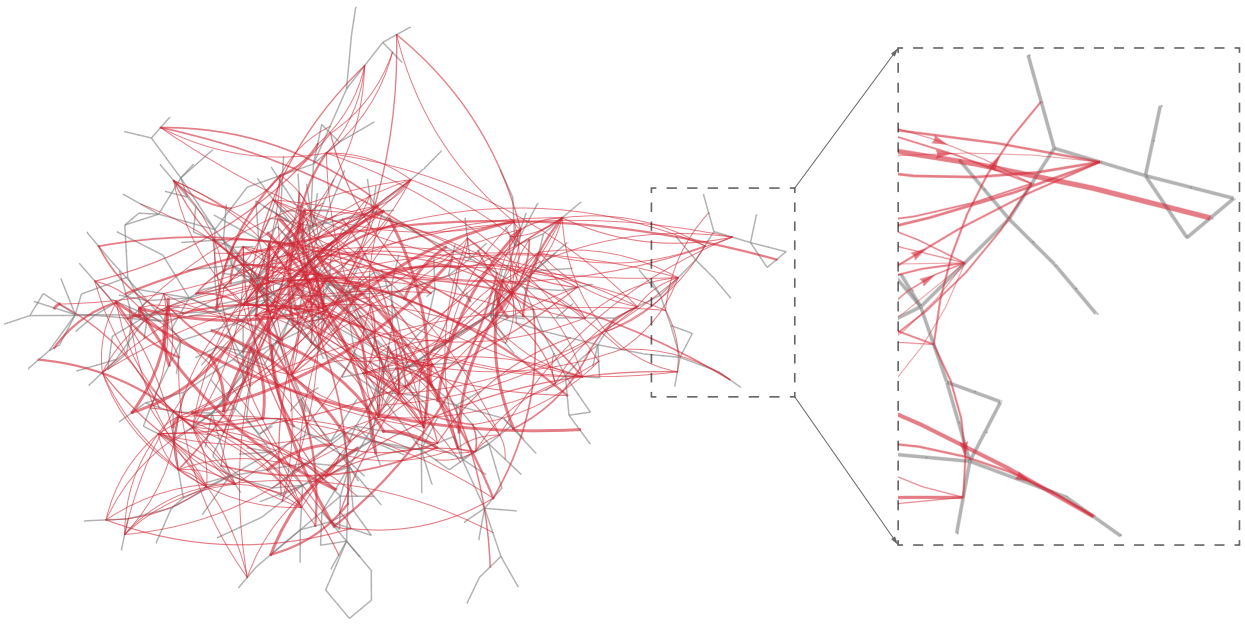


Figure 3.2: The gray network is the system network and the red network is the influence graph showing the main influences between lines. The red edge thickness indicates the strength of the influence.

that probability is sufficiently large. The interactions in Fig. 3.2 reflect a very wide range of mechanisms. The longer-range mechanisms include redistributions of power due to line and generator outages, remedial action schemes, and bad weather across the grid.

Let the row vector $\boldsymbol{\pi}_k$ be the probability distribution of states in generation k. The $\boldsymbol{\pi}_k$ entry $\pi_k[i]$ is the probability that the set of outaged lines is s_i in generation k; that is,

$$\pi_k[i] = \mathrm{P}[X_k = s_i]. \tag{3.3}$$

Then the propagation of sets of line outages from generation k to generation $k+1$ is given by

$$\boldsymbol{\pi}_{k+1} = \boldsymbol{\pi}_k \boldsymbol{P}_k \tag{3.4}$$

and, using (3.4), the distribution of states in generation k depends on the initial distribution of states $\boldsymbol{\pi}_0$ according to

$$\boldsymbol{\pi}_k = \boldsymbol{\pi}_0 \boldsymbol{P}_0 \boldsymbol{P}_1 ... \boldsymbol{P}_{k-2} \boldsymbol{P}_{k-1}. \tag{3.5}$$

3.3 Illustrative historical outage data

While our method applies generally to the detailed outage data routinely collected by utilities, we illustrate our method with a specific publicly available data set, which is the automatic transmission line outages recorded by a large North American utility over 14 years starting in 1999 [7]. We group the 9,741 line outages into 6,687 cascades [68]. Most of the cascades (87%) have one generation because initial outages often do not propagate further. There are 614 lines and the observed cascades have 1094 subsets of these lines that form the 1094 states s_0, s_1, ... , s_{1093}. Among these 1094 states, 50% have multi-line outages. And among these multi-line outage states, about 20% are comprised of lines sharing no common buses. While in theory there are 2^{614} subsets of 614 lines, giving an impractically large number of states, we find in practice with our data that the number of states is less than twice the number of lines. Note that our statistical modeling approximates the power grid as unchanging over the time span of the data [1]. In practice a utility would have the records of changes to partially mitigate the effects of this approximation.

3.4 Computing the distribution of cascade sizes and its confidence interval

We compute the distribution of cascade sizes from the Markov chain and check that it reproduces the empirical distribution of cascade sizes, and estimate its confidence interval with a bootstrap.

We can measure the cascade size by its number of generations. Define the survival function of the number of generations in a cascade as

$$S(k) = \mathrm{P}[\text{number of cascade generations} > k] \tag{3.6}$$

$\pi_k[0]$ is the probability that a cascade is in state $s_0 = \{\}$ in generation k and also the probability that the cascade stops at or before generation k. Hence, $1 - \pi_k[0]$ is the probability that a cascade

has at least k generations. That is,

$$\begin{aligned} S(k) &= 1 - \pi_k[0] = \boldsymbol{\pi}_k(\mathbf{1} - \boldsymbol{e}_0) \\ &= \boldsymbol{\pi}_0 \boldsymbol{P}_0 \boldsymbol{P}_1 ... \boldsymbol{P}_{k-2} \boldsymbol{P}_{k-1} (\mathbf{1} - \boldsymbol{e}_0), \end{aligned} \tag{3.7}$$

where $\mathbf{1}$ is the column vector $(1, 1, 1, ..., 1)'$, and $\boldsymbol{e}_0$ is the column vector $(1, 0, 0, 0, ..., 0)'$. The initial state distribution $\boldsymbol{\pi}_0$ can be estimated directly from the cascading data.

Then we can confirm that the influence graph reproduces the statistics of the cascade size in the cascading data by comparing the survival function $S(k)$ computed from (3.7) with the empirical survival function computed directly from the cascading data as shown in Fig. 3.3. The Markov chain reproduces the statistics of the cascade size closely, with a Pearson χ^2 goodness-of-fit test p-value of 0.99.

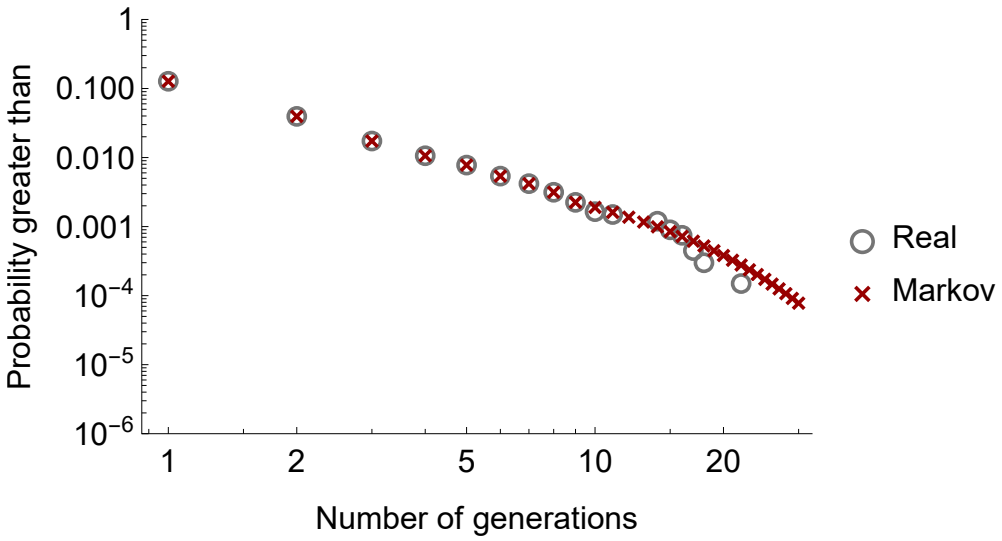


Figure 3.3: Survival functions of the number of generations from real data and from the Markov chain.

We use bootstrap resampling [69] to estimate the variance of our estimates of probabilities of cascade sizes. A bootstrap sample resamples the observed cascades with replacement, reconstructs the Markov chain, and recomputes the probabilities of cascade sizes. Note that each bootstrap resampling amounts to a different selection of the cascades observed in the data. The variance of the probabilities of cascade sizes is then obtained as the empirical variance of the

bootstrap samples. We use 500 bootstrap samples to ensure a sufficiently accurate estimate of the variance of the probabilities.

The risk of a given size of blackout is estimated as risk = (estimated probability $\hat{p}$ of that size of blackout) $\times$ (cost of that size of blackout). Knowing the multiplicative uncertainty in $\hat{p}$ is useful. For example, if we know $\hat{p}$ varies within a factor of 2, then this contributes a factor of 2 to the uncertainty of the risk. Therefore, it is appropriate to use a multiplicative form of confidence interval for $\hat{p}$ specified by a parameter κ. A 95% multiplicative confidence interval for an estimated probability $\hat{p}$ means that the probability p satisfies $\mathrm{P}[\hat{p}/\kappa \leq p \leq \hat{p}\,\kappa] = 0.95$. The confidence interval for the estimated survival function is shown in Fig. 3.4. Since larger cascades are rarer than small cascades, the variation increases as the number of generations increases.

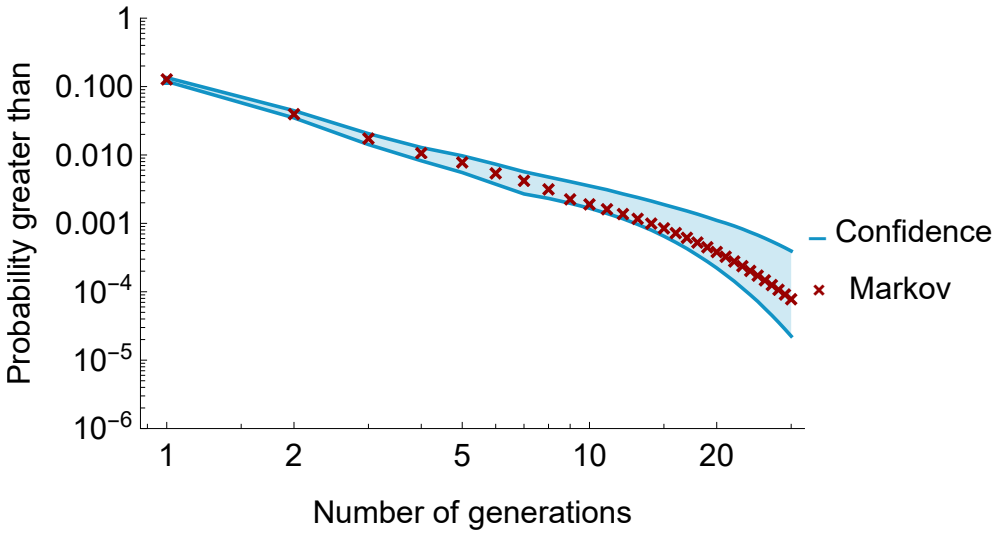


Figure 3.4: Survival function of cascade sizes. Red crosses are from Markov chain, and blue lines indicate the 95% confidence interval estimated by bootstrap.

To apply and communicate the probability distribution of cascade size, it is convenient to combine sizes together to get the probabilities of small, medium, and large cascades, where a small cascade has 1 or 2 generations, a medium cascade has 3 to 9 generations, and a large cascade has 10 or more generations. (The respective probabilities are calculated as $1 - S(2)$, $S(2) - S(9)$, and $S(9)$). The 95% confidence intervals of the estimated probabilities of small, medium, and large cascades are shown in Table 3.1. The probability of large cascades is estimated

within a factor of 1.5, which is adequate for the purposes of estimating large cascade risk, since the cost of large cascades is so poorly known: estimates of the direct costs of cascading blackouts vary by more than a factor of 2.

Table 3.1: 95% Confidence intervals using bootstrap

cascade size	probability	κ
small (1 or 2 generations)	0.9606	1.005
medium (3 to 9 generations)	0.0372	1.132
large (10 or more generations)	0.0022	1.440

We now discuss tracking cascades by their number of generations. The number of generations is the same concept as the number of tiers in commercial cascading software [38]. Basic to cascading analysis is the grouping of line outages into successive generations within each cascade. This grouping is usually done by outage timing as in this work, or by simulation loops naturally producing generations of outages. This influence graph is structured in terms of these generations, so that propagation is determined by the probability of a next generation (i.e. the cascade not stopping at the current generation), and cascade size is measured by number of cascade generations. In contrast, some previous papers [16, 17, 67, 68] are structured in terms of the line outages in the generations, so that, according to the branching process model [68], each line outage in each generation propagates independently to form line outages in the next generation. Then the propagation is determined by the number of line outages per line outage in the previous generation, and it is natural to use the total number of lines outaged as a measure of cascade size. While it is not yet clear which approach is better, there may be some advantages to tracking cascades by generations rather than line outages. Generations are simpler and more general than line outages, and can more easily encompass other outages significant in cascading such as transformer outages. Also, it may be that the statistics of the number of generations is more simply described, as in the Zipf distribution observed in utility data in [70].

3.5 Critical lines and cascade mitigation

3.5.1 The transmission lines involved in large cascades

The lines eventually most involved in large cascades can be calculated from the asymptotic properties of the Markov chain. While all cascades eventually stop, we can consider at each generation those propagating cascades that are not stopped at that generation. The probability distribution of states involved in these propagating cascades converges to a probability distribution $\boldsymbol{d}_\infty$, which is called the quasi-stationary distribution. $\boldsymbol{d}_\infty$ can be computed directly from the transition matrices (as explained in Appendix A, $\boldsymbol{d}_\infty$ is the left eigenvector corresponding to the dominant eigenvalue of the transition submatrix $\bar{\boldsymbol{Q}}_{1+}$). That is, except for a transient that dies out after some initial generations, the participation of states in the cascading that continues past these initial generations is well approximated by $\boldsymbol{d}_\infty$. Thus the high probability states corresponding to the highest probability entries in $\boldsymbol{d}_\infty$ are the critical states most involved in the latter portion of large cascades. Since $\boldsymbol{d}_\infty$ does not depend on the initial outages, the Markov chain is supplying information about the eventual cascading for all initial outages.

We now find the critical lines corresponding to these critical states by projecting the states onto the lines in those states. Let $\boldsymbol{\ell}_k$ be the row vector whose entry $\ell_k[j]$ is the probability that line j outages in generation k. Then

$$\ell_k[j] = \sum_{i:j\in s_i} \pi_k[i] \quad \text{or} \quad \boldsymbol{\ell}_k = \boldsymbol{\pi}_k \boldsymbol{R}, \tag{3.8}$$

where the matrix $\boldsymbol{R}$ projects states to lines according to

$$R[i,j] = \begin{cases} 1; & \text{line } j \in s_i \\ 0; & \text{line } j \notin s_i \end{cases} \tag{3.9}$$

Then the probability distribution of lines eventually involved in the propagating cascades that are not stopped is $\boldsymbol{c}_\infty = \boldsymbol{d}_\infty \boldsymbol{R}$ and the critical lines most involved in the latter portion of large cascades correspond to the highest probability entries in $\boldsymbol{c}_\infty$. Fig. 3.5 shows the probabilities in $\boldsymbol{c}_\infty$ in order of decreasing probability. We identify the top ten lines as critical and as candidates for upgrading to decrease the probability of large cascades.

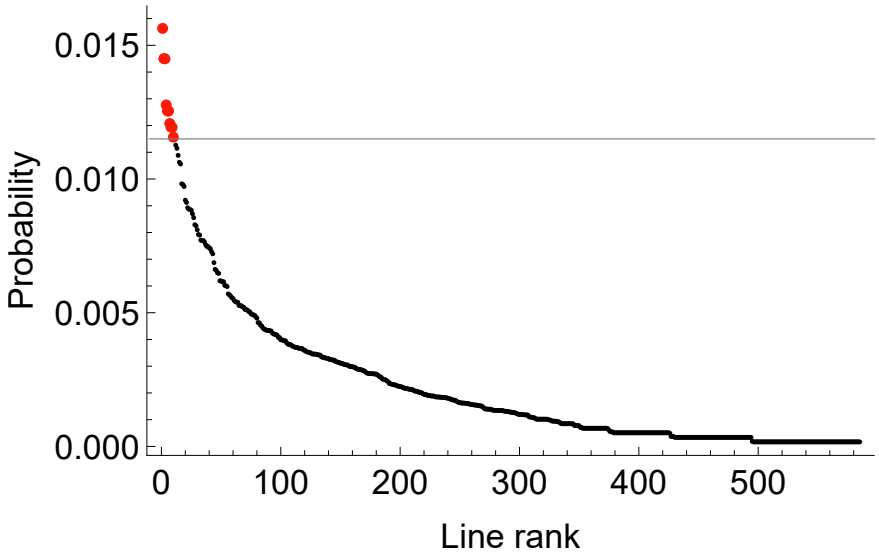


Figure 3.5: Quasi-stationary distribution of transmission lines eventually involved in propagating cascades. Red dots are ten critical lines.

3.5.2 Modeling and testing mitigation in the Markov chain

A transmission line is less likely to fail due to other line outages after the line is upgraded, its protection is improved, or its operating limit is increased. These mitigations have the effect of decreasing the probability of transition to states containing the upgraded line, and are an adjustment of the columns of the transition matrix corresponding to these states. The mitigation is represented in the Markov chain by reducing the probability of transition to the state s containing the upgraded line by $(r/|s|)\%$, where $|s|$ is the number of lines in the state. The reduction is $r\%$ if the state contains only the upgraded line, and the reduction is less if the state contains multiple lines.

We demonstrate using the Markov chain to quantify the impact of mitigation by upgrading ten lines critical for large cascades identified in section 3.5.1 with $r = 80\%$. The effect of this mitigation on cascade probabilities is shown in Fig. 3.6. It shows that upgrading the critical lines reduces the probability of large cascades by 45%, while the probability of medium cascades is slightly decreased and the probability of small cascades is slightly increased.

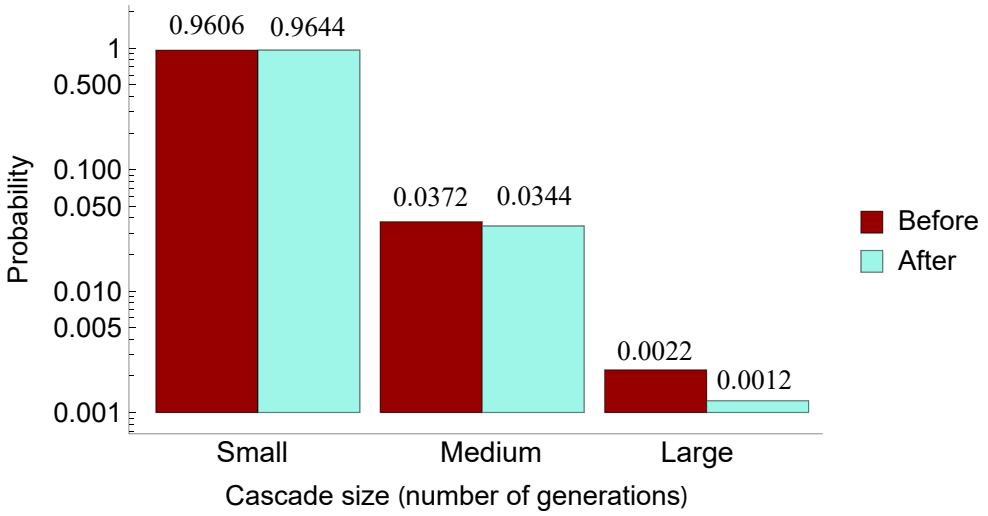


Figure 3.6: Cascade size distribution before (red) and after (light green) mitigating lines critical in propagating large cascades.

To show the effectiveness of the method of identifying critical lines, we compare the mitigation effect of upgrading critical lines and upgrading ten random lines. Randomly upgrading ten lines only decreases the probability of large cascades by 11% on average.

So far we have only considered upgrading the lines critical for propagating large cascades. Now, in order to discuss this mitigation of large cascades in a larger context, we briefly consider and contrast a different mitigation tactic of upgrading lines that are critical for initial outages. Since initial outages are caused by external causes such as storm, lightning, or misoperation, they often have different mechanisms and different mitigations than for propagating outages. A straightforward method to identify lines critical for initial outages selects ten lines in the data with the highest frequencies of initial outage [17]. Upgrading these ten lines will reduce their initial outage frequencies and hence reduce the overall cascade frequency. In the Markov chain, this upgrading is represented by reducing in the first generation the frequency of states s that contain the critical lines for initial outages by $r/|s|\%$, where $r = 80\%$. The main effect is that by reducing the initial outage frequencies of the critical lines by 80%, we reduce the frequency of all cascades by 19%. In addition, this mitigation will change the probabilities of states $\boldsymbol{\pi}_0$ after

renormalizing the frequencies of states. It turns out for our case that there is no overlap between critical lines for initial outages and for propagation.

Changing the initial state distribution $\boldsymbol{\pi}_0$ has no effect on the distribution of cascade sizes in the long-term. However, it directly reduces the frequency of all cascades. In contrast, mitigating the lines critical for propagating large cascades reduces the probability of large cascades relative to all cascades but has no effect on the frequency of all cascades. (Note that Fig. 3.6 shows the distribution of cascade sizes assuming that there is a cascade, but gives no information about the frequency of all cascades.)

In practice, a given mitigation measure can affect both the initial outages and the propagation of outages into large cascades. The combined mitigation effects can also be represented in the influence graph by changing both the initial state distribution and the transition matrix, but here it is convenient to discuss them separately.

This work aims to select the lines critical for large cascades and quantify the impact on cascade probability of generic upgrades to these lines. Once the critical lines are selected, an engineering process of much wider scope is required to determine the possible approaches to upgrade each of the lines, quantify the benefits other than reducing large cascades and balance the costs and feasibilities of the upgrading approaches against the total benefits of upgrading. One part of this process is that for each line, the percentage reduction in outage probability for the best approach to line upgrade is estimated and the Markov chain is used to quantify the corresponding reduction in large, medium, and small cascade probabilities. However, cascade mitigation is only one of the many factors to be considered in justifying upgrade. Evaluating and costing specific upgrading approaches for specific lines requires utility expertise, including details of the line construction and right of way, maintenance history, and operation.

3.6 Estimating the transition matrix

The Markov chain has an absorbing first state $s_0 = \{\}$, indicating no lines outaged as the cascade stops and after the cascade stops. Therefore the transition matrix has the structure

$$\boldsymbol{P}_k = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline \boldsymbol{u}_k & & \boldsymbol{Q}_k & \end{array}\right] \tag{3.10}$$

where $\boldsymbol{u}_k$ is a column vector of stopping probabilities; that is, $u_k[i] = P_k[i, 0]$. $\boldsymbol{Q}_k$ is a submatrix of transition probabilities between transient states which contains the non-stopping probabilities. The first row of $\boldsymbol{P}_k$ is always $\boldsymbol{e}_0'$, so the transition probabilities to be estimated are $\boldsymbol{u}_k$ and $\boldsymbol{Q}_k$ for each generation k. The rows and columns of $\boldsymbol{P}_k$ are indexed from 0 to $|\mathcal{S}| - 1$ and the rows and columns of $\boldsymbol{Q}_k$ are indexed from 1 to $|\mathcal{S}| - 1$, where $|\mathcal{S}|$ is the number of states.

As summarized in section 3.2 after (3.1), we need to both group together multiple generations to get sufficient data and account for variation with generation k. The statistics of the transition from generation 0 to generation 1 are different than the statistics of the transitions between the subsequent generations. For example, stopping probabilities for generation 0 are usually larger than stopping probabilities for subsequent generations [17]. Also, the data for the subsequent generations is sparser. Therefore, we construct from counts of the number of transitions from generation 0 to generation 1 a probability transition matrix $\bar{\boldsymbol{P}}_0$, and construct from the total counts of the number of transitions from all the subsequent generations a probability transition matrix $\bar{\boldsymbol{P}}_{1+}$. Specifically, we first use the right-hand side of (3.2) to construct two corresponding empirical transition matrices, and then we update stopping probabilities by the empirical Bayes method and adjust non-stopping probabilities to obtain $\bar{\boldsymbol{P}}_0$ and $\bar{\boldsymbol{P}}_{1+}$. Finally, we adjust $\bar{\boldsymbol{P}}_0$ and $\bar{\boldsymbol{P}}_{1+}$ to match the observed propagation rates to obtain $\boldsymbol{P}_k$ for each generation k.

3.6.1 Bayesian update of stopping probabilities

The empirical stopping probabilities are improved by an empirical Bayes method [71, 72] to help mitigate the sparse data for some of these probabilities. Since the method is applied to both $\bar{\boldsymbol{P}}_0$ and $\bar{\boldsymbol{P}}_{1+}$, we simplify notation by writing $\bar{\boldsymbol{P}}$ for either $\bar{\boldsymbol{P}}_0$ or $\bar{\boldsymbol{P}}_{1+}$.

The matrix of empirical probabilities obtained from the transition counts $N[i,j]$ is

$$\bar{P}^{\text{counts}}[i,j] = \frac{N[i,j]}{\sum_j N[i,j]} \tag{3.11}$$

We construct $\bar{\boldsymbol{P}}$ from $\bar{\boldsymbol{P}}^{\text{counts}}$ in two steps. First, Bayesian updating is used to better estimate stopping probabilities and form a matrix $\bar{\boldsymbol{P}}^{\text{bayes}}$. Second, the non-stopping probabilities in $\bar{\boldsymbol{P}}^{\text{bayes}}$ are adjusted to form the matrix $\bar{\boldsymbol{P}}$ to account for the fact that some independent outages are grouped into cascading outages when we group outage data into cascades.

We need to estimate the probability of the cascade stopping at the next generation for each state encountered in the cascade. For some of the states, the stopping counts are low, and cannot give good estimates of the stopping probability. However, by pooling the data for all the states we can get a good estimate of the mean probability of stopping over all the states. We use this mean probability to adjust the sparse counts in a conservative way. In particular, we form a prior that maximizes its entropy subject to the mean of the prior being the mean of the pooled data. This maximum entropy prior can be interpreted as the prior distribution that makes the least possible further assumptions about the data [73] [74].

Finding a maximum entropy prior Assuming the stopping counts are independent with a common probability, the stopping counts follow a binomial distribution. Its conjugate prior distribution is the beta distribution, whose parameters are estimated using the maximum entropy method.

Let stopping counts C_i be the observed number of transitions from state s_i to s_0 $(i = 1, ..., |\mathcal{S}| - 1)$. Then $C_i = N[i, 0]$. Let $n_i = \sum_{j=0}^{|\mathcal{S}|-1} N[i,j]$ be the row sum of the counting matrix $\boldsymbol{N}$. The stopping counts C_i follow a binomial distribution with parameter U_i, with

probability mass function

$$f_{C_i|U_i}(c_i|u_i) = \frac{n_i!}{c_i!(n_i - c_i)!} u_i^{c_i}(1 - u_i)^{n_i - c_i} \tag{3.12}$$

The conjugate prior distribution for the binomial distribution is the beta distribution. Accordingly, we use the beta distribution with hyperparameters β_1, β_2 for the stopping probability U_i:

$$f_{U_i}(u_i) = B(\beta_1, \beta_2) u_i^{\beta_1 - 1}(1 - u_i)^{\beta_2 - 1} \tag{3.13}$$

where $B(\beta_1, \beta_2) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}$. Alternative parameters for the beta distribution are its precision $m = \beta_1 + \beta_2$ and its mean $\mu = \frac{\beta_1}{\beta_1 + \beta_2}$. The entropy of the beta distribution is

$$\begin{aligned} \mathrm{Ent}(m, \mu) = &\ln B(m\mu, m(1 - \mu)) - (m\mu - 1)\psi(m\mu) \\ &- (m(1 - \mu) - 1)\psi(m(1 - \mu)) + (m - 2)\psi(m) \end{aligned} \tag{3.14}$$

where $\psi(x)$ is the digamma function.

We want to estimate hyperparameters β_1, β_2 to make the beta distribution have maximum entropy subject to the mean being the average stopping probability of the pooled data $\hat{u} = (\sum_{i=1}^{|\mathcal{S}|-1} c_i)/(\sum_{i=1}^{|\mathcal{S}|-1} n_i)$. Then we can obtain hyperparameters β_1, β_2 by finding the $m > 0$ that maximizes $\mathrm{Ent}(m, \hat{u})$ and evaluating $\beta_1 = m\hat{u}$ and $\beta_2 = m(1 - \hat{u})$. The hyperparameters used for $\bar{\boldsymbol{P}}_0^{\mathrm{bayes}}$ are $(\beta_1, \beta_2) = (2.18, 0.32)$, and the hyperparameters for $\bar{\boldsymbol{P}}_{1+}^{\mathrm{bayes}}$ are $(\beta_1, \beta_2) = (1.10, 0.93)$.

Updating the observed data using the prior The posterior distribution of the stopping probability U_i is a beta distribution with parameters $c_i + \beta_1, n_i - c_i + \beta_2$. We use the mean of the posterior distribution as a point estimate of the stopping probability:

$$\bar{P}^{\mathrm{bayes}}[i, 0] = \mathrm{E}(U_i | C_i = c_i) = \frac{c_i + \beta_1}{n_i + \beta_1 + \beta_2} \tag{3.15}$$

Fig. 3.7 shows a comparison between the empirical stopping probabilities and the updated stopping probabilities. Black dots are the empirical probabilities sorted in ascending order (if two

probabilities are equal, they are sorted according to the total counts observed). Red dots are the updated stopping probabilities. As expected, the empirical probabilities with the fewest counts move towards the mean the most when updated. As the counts increase, the effect of the prior decreases and the updated probabilities tend to the empirical probabilities.

Equation (3.15) forms the first column of $\bar{\boldsymbol{P}}^{\text{bayes}}$. Then the nonstopping probabilities in the rest of the columns of the $\bar{\boldsymbol{P}}^{\text{counts}}$ matrix are scaled so that they sum to one minus the stopping probabilities of (3.15) to complete the matrix $\bar{\boldsymbol{P}}^{\text{bayes}}$:

$$\bar{P}^{\text{bayes}}[i,j] = \frac{1-\bar{P}^{\text{bayes}}[i,0]}{\sum_{r=1}^{|\mathcal{S}|-1}\bar{P}^{\text{counts}}[i,r]}\bar{P}^{\text{counts}}[i,j],\ \ j>0 \tag{3.16}$$

This Bayesian updating is applied to form $\bar{\boldsymbol{P}}_0^{\text{bayes}}$ for the first transition and $\bar{\boldsymbol{P}}_{1+}^{\text{bayes}}$ for the subsequent transitions.

3.6.2 Adjust nonstopping probabilities for independent outages

The method explained in section 3.2 that groups outages into cascades has an estimated 6% chance that it groups independent outages into cascading outages [1]. These 6% of outages occur independently while the cascading of other outages proceeds and do not arise from interactions with other outages. The empirical data for the nonstopping probabilities includes these 6% of outages, and we want to correct this. Therefore, the non-stopping probabilities are modified by shrinking the probabilities in transition matrix by 6%, and sharing this equally among all the states. That is,

$$\bar{P}[i,j] = 0.94\bar{P}^{\text{bayes}}[i,j] + \frac{0.06}{|\mathcal{S}|-1}(1-\bar{P}^{\text{bayes}}[i,0]) \tag{3.17}$$

where $\bar{P}^{\text{bayes}}$ indicates the transition matrices after the Bayesian update of section 3.6.1. Notice that $\bar{P}$ is a probability matrix since $\sum_j \bar{P}(i,j) = 1$ for each i. A benefit is that this adjustment makes the submatrix $\boldsymbol{Q}_k$ have non-zero off-diagonal entries, making $\bar{P}$ irreducible.

3.6.3 Adjustments to match propagation

The average propagation ρ_k for generation k [68] is estimated from the data using

$$\begin{aligned}\hat{\rho}_k &= \frac{\text{Number of cascades with} > k+1 \text{ generations}}{\text{Number of cascades with} > k \text{ generations}} \\ &= \frac{S(k+1)}{S(k)} = \frac{\boldsymbol{\pi}_{k+1}(\mathbf{1}-\boldsymbol{e}_0)}{\boldsymbol{\pi}_k(\mathbf{1}-\boldsymbol{e}_0)}\end{aligned} \tag{3.18}$$

An important feature of the cascading data is that average propagation ρ_k increases with generation k as shown in Table 3.2. To do this, we need to form transition matrices for each of these generations that reproduce this propagation. We define a matrix $\boldsymbol{A}_k$ to adjust $\bar{\boldsymbol{P}}_0$ and

Table 3.2: Propagations of generations $k = 0$ to 17

k	0	1	2	3	4	5	6	7	8
$\hat{\rho}_k$	0.13	0.31	0.44	0.61	0.73	0.70	0.78	0.75	0.71
k	9	10	11	12	13	14	15	16	17
$\hat{\rho}_k$	0.73	0.91	1.00	1.00	0.80	0.75	0.83	0.60	0.67

$\bar{\boldsymbol{P}}_{1+}$ so that the propagation in $\boldsymbol{P}_k$ matches the empirical propagation for each generation up to generation 8. For generation 9 and above, the empirical propagation for each generation is too noisy to use individually and we combine those generations to obtain a constant transition matrix. That is, $\boldsymbol{P}_0 = \bar{\boldsymbol{P}}_0\boldsymbol{A}_0$, $\boldsymbol{P}_1 = \bar{\boldsymbol{P}}_{1+}\boldsymbol{A}_1$, ... , $\boldsymbol{P}_8 = \bar{\boldsymbol{P}}_{1+}\boldsymbol{A}_8$, $\boldsymbol{P}_{9+} = \bar{\boldsymbol{P}}_{1+}\boldsymbol{A}_{9+}$. Then the transition matrices for all the generations are $\boldsymbol{P}_0, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5, \boldsymbol{P}_6, \boldsymbol{P}_7, \boldsymbol{P}_8, \boldsymbol{P}_{9+}, \boldsymbol{P}_{9+}, \boldsymbol{P}_{9+}, ...$

The matrix $\boldsymbol{A}_k$ has the effect of transferring a fraction of probability from the transient to stopping transitions and has the following form:

$$\boldsymbol{A}_k = \begin{pmatrix} 1 & 0 & ... & 0 \\ a_k & 1-a_k & ... & 0 \\ \vdots & & \ddots & \\ a_k & 0 & ... & 1-a_k \end{pmatrix} \tag{3.19}$$

a_k is determined from the estimated propagation rate $\hat{\rho}_k$ as follows. Using (3.18), we have

$$\hat{\rho}_k = \frac{\boldsymbol{\pi}_k\bar{\boldsymbol{P}}\boldsymbol{A}_k(\mathbf{1}-\boldsymbol{e}_0)}{\boldsymbol{\pi}_k(\mathbf{1}-\boldsymbol{e}_0)} = (1-a_k)\frac{1-\boldsymbol{\pi}_k\bar{\boldsymbol{P}}\boldsymbol{e}_0}{1-\boldsymbol{\pi}_k\boldsymbol{e}_0} \tag{3.20}$$

and we solve (3.20) to obtain a_k for each generation k.

3.7 Conclusion and discussion

We process observed transmission line outage utility data to form a generalized influence graph and the associated Markov chain that statistically describe cascading outages in the data. Successive line outages, or, more precisely, successive sets of near simultaneous line outages in the cascading data correspond to transitions between nodes of the influence graph and transitions in the Markov chain. The more frequently occurring successive line outages in the cascading data give a stronger influence between nodes and higher transition probabilities. The generalized influence graph introduces additional states corresponding to multiple line outages that occur nearly simultaneously. This innovation adds a manageable number of additional states and solves some problems with previous influence graphs, making the formation of the Markov chain clearer and more rigorous.

One of the inherent challenges of cascading is the sparse data for large cascades. We have used several methods to partially alleviate this when estimating the Markov chain transition matrices, including combining data for several generations, conservatively improving estimates of stopping probabilities with an empirical Bayes method, accounting for independent outages during the cascade, and matching the observed propagation for each generation. The combined effect of these methods is to improve estimates of the Markov chain transition matrices. Although some individual elements of these transition matrices are nevertheless still poorly estimated, what matters is the variability of the results from the Markov chain, which are the probabilities of small, medium and large cascades. We assess the variability of these estimated probabilities with a bootstrap and find them to be estimated to a useful accuracy. This assessment of variability is necessary for getting useful estimates of large cascade probability because large cascades are rare, and probability estimates for rare events have the potential to be so wildly variable that they are useless.

The Markov chain only models the statistics of successive transitions in the observed data. Also, there is an inherent limitation of not being able to account for transitions and states not present in the observed data. That is, the common transitions and states and some of the rarer

transitions and states will be present in the data and will be represented in the Markov model, while the rarer transitions and states not present in the data will be neglected. However, the Markov chain can produce, in addition to the observed cascades, combinations of the observed transitions that are different than and much more extensive than the observed cascades. The Markov chain approximates the statistics of cascading rather than reproducing only the observed cascades.

We exploit the asymptotic properties of the Markov chain to calculate the transmission lines most involved in the propagation of larger cascades, and we show with the Markov chain that upgrading these lines can significantly reduce the probability of large cascades. Since a large cascade of line outages with many generations is very likely to shed substantial load, mitigating large cascades will also mitigate blackouts with large amounts of load shed.

A Markov chain driven by real data incorporates all the causes, mechanisms, and conditions of the cascading that occurred, but does not distinguish particular causes of the interactions. However, once the lines critical to large cascades have been identified with the influence graph, the causes related to outage of those particular lines can be identified by analyzing event logs and cause codes. Also, the overall impact on cascading of factors such as loading and weather can be studied by dividing the data into low and high loading or good and bad weather and forming influence graphs for each case.

While the Markov model is driven by historical data in this work, the Markov model is not limited to historical data. The Markov model could be driven by simulated cascades or a combination of simulated and historical cascades. Moreover, if the probabilities of specific cascading interactions between line outages are available, these probabilities could be combined into the entries of the Markov transition matrices. The Markov chain is applied here to cascading transmission line outages, but the formulation would apply generally to process real or simulated data for the cascading outage of components within or between networked infrastructures.

We show how to estimate the Markov chain from detailed outage data that is routinely collected by utilities. Being driven by observed data has some significant advantages of realism.

In particular, and in contrast with simulation approaches, no assumptions about the detailed mechanisms of cascading need to made. Since the Markov chain driven by utility data has different assumptions than simulation, we regard the Markov chain and simulation approaches as complementary. The Markov chain driven by observed data offers another way to find critical lines and to test proposed mitigations of cascading by predicting the effect of the mitigation on the probabilities of small, medium, and large cascades.

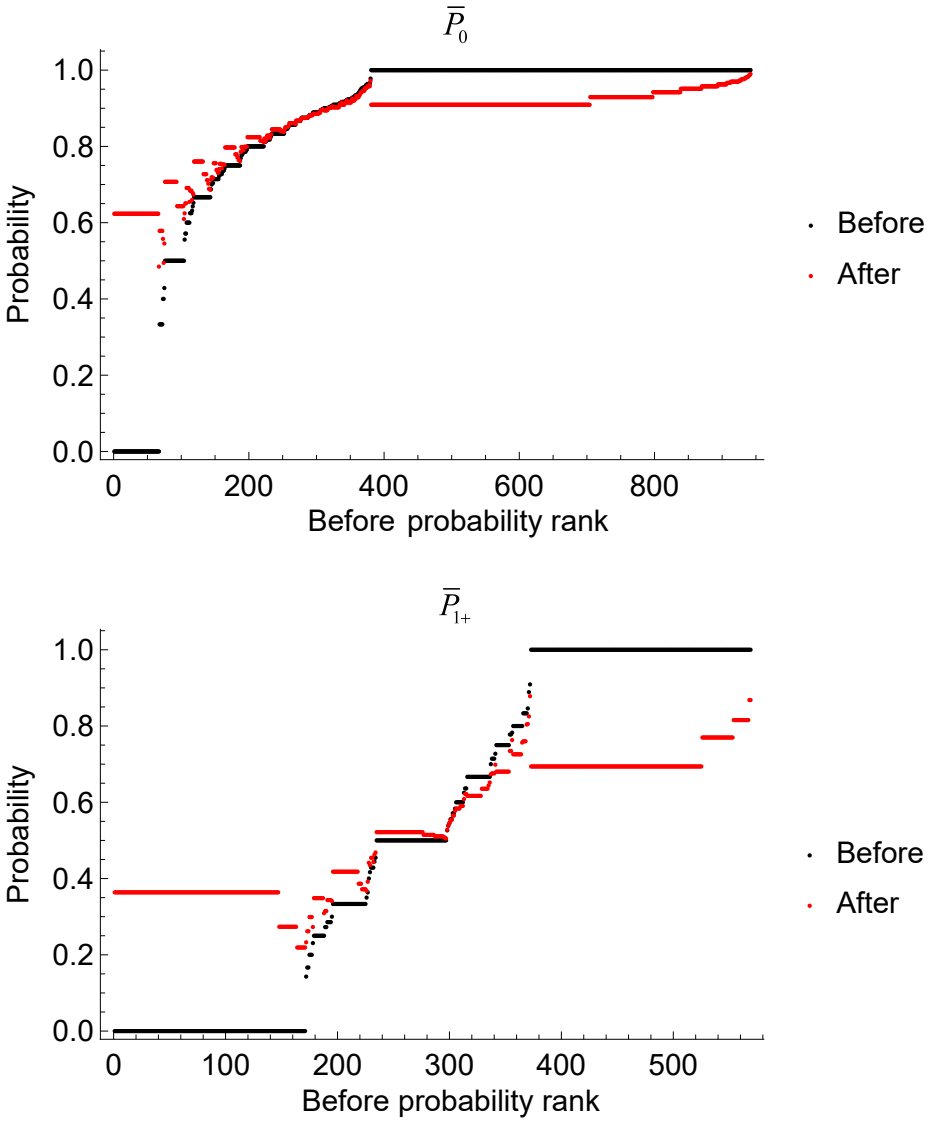


Figure 3.7: Stopping probabilities before and after Bayesian updating

CHAPTER 4. SIMULATING CASCADING RESILIENCE FROM HISTORICAL DATA USING THE MARKOVIAN INFLUENCE GRAPH

Extreme events can damage power system components and then cause cascading outages. Methods are needed to evaluate the cascading phase of resilience. The Markovian influence graph is a discrete Markov chain with variant transition matrices. It is straightforward to sample cascades from it starting with some initial outages. However, large cascades are rare and the Markovian influence graph produces limited large cascade samples. This chapter proposes an improveed sampling method to encompass the rare, large cascades that contribute greatly to the blackout risk. Then, the load shed distribution is estimated from the samples of cascaded lines and hence the risk of a widespread blackout caused by extreme events and cascading.

This chapter uses OPA cascading failure simulation results. We thank Benjamin Carreras of BACV Solutions, Oak Ridge TN, David E. Newman of University of Alaska-Fairbanks, and José-Miguel Reynolds-Baredo of Universidad Carlos III de Madrid for producing OPA simulation results. The material in this chapter is published in [75].

4.1 Introduction

The processed historical cascades are the observational bedrock for the study of cascading failure, since they occurred in practice. However, if one assumes some initial outages and seeks to predict the probabilistic extent of further cascading, ranging from no further outages to blackouts, the historical cascades are limiting: the particular initial outages may not occur in the historical cascades, and even if they do occur once or twice, it is only one or two samples of the possible cascading outcomes. To address this problem, we propose using the Markovian influence graph to describe the statistics of the historical outage data, and then sample from the Markovian influence graph to simulate the consequences of some assumed initial outages. This gives a

high-level and flexible statistical model of cascading that can be driven by standard utility data. In suggesting this approach, we are motivated by the resilience problem of estimating the cascading that can follow damage to the power transmission system in an extreme event [41–43].

Extreme events such as storm, fire, or earthquake can damage multiple power system components. Then further power system components can outage in a cascade. Usually the cascading only outages components without damaging them, but the cascading does make the blackout more widespread and impactful, and can seriously hinder the subsequent recovery from the event. There is considerable expertise modeling probabilistic power system component failure under extreme conditions of wind, flooding, icing, earthquake and fire. However, the cascading phase of resilience is much less well characterized. Given the initially damaged components, one can simulate the cascading using a model-based simulation. While useful, simulation only captures a limited subset of approximated cascading mechanisms. The alternative that we suggest and explore in this work uses a Markovian influence graph driven by historical utility data to generate samples of the cascaded transmission lines. Throughout this chapter we are interested in properly sampling from the largest cascades since these dominate the risk [4], because straightforward sampling does not work well for the larger cascades.

4.2 Sampling cascades with the influence graph

Let Y_0 be the set of initially failed lines that are damaged by the extreme event. We express Y_0 as a disjoint union of m Markov chain states:

$$Y_0 = x_0^{(1)} \cup x_0^{(2)} \cup ... \cup x_0^{(m)} \tag{4.1}$$

Consider the state $x_0^{(r)}$ in Y_0 with $1 \leq r \leq m$. Let the rth Markov chain starting from state $x_0^{(r)}$ but subsequently avoiding any initially failed states be $X_0^{(r)}, X_1^{(r)}, ...$ That is, $\mathrm{P}[X_0^{(r)} = x_0^{(r)}] = 1$ and $\mathrm{P}[X_k^{(r)} \in Y_0] = 0$ for $k > 0$. (The transition matrix for the Markov chain $X_0^{(r)}, X_1^{(r)}, ...$ is easily obtained by preventing transitions to states in Y_0 by deleting the columns of the transition matrix corresponding to states in Y_0 and renormalizing.)

We write $|x|$ for the number of line outages in state x. The number of lines out in the rth chain is

$$N^{(r)} = \sum_{k=0}^{\infty} |X_k^{(r)}| \tag{4.2}$$

and the total number of lines out is

$$N = \sum_{r=1}^{m} N^{(r)} \tag{4.3}$$

Note that (4.2) and (4.3) neglect any repeats of lines out within or between chains.

4.2.1 Simulating the influence graph

We first describe a straightforward but inferior way to do the simulation. For the rth chain we need to simulate $X_0^{(r)}, X_1^{(r)}, ...$ from its starting state $x_0^{(r)}$ until it stops. That is, the simulation produces a series of states $x_0^{(r)}, x_1^{(r)}, x_2^{(r)}, ...$ until it stops by transitioning to the empty state $\{\}$. Suppose state $x_j^{(r)}$ is produced at step j. Then the next state is produced as follows: Let $e_j^{(r)}$ be the row vector with a one at the index of state $x_j^{(r)}$ and zeros elsewhere. Let P_k be the transition matrix from generation k to $k+1$. Then $e_j^{(r)} P_k$ is a probability distribution over the states not in Y_0 [1]. Sample from this probability distribution to obtain the state $x_{k+1}^{(r)}$. Thus $x_0^{(r)}, x_1^{(r)}, x_2^{(r)}, ...$ are produced.

The problem with this straightforward way to do the simulation is that it will mainly sample the frequent short cascades with few line outages, so that a huge number of samples is needed to accurately estimate the longer cascades. An advantage of the influence graph is that it can be easily modified to sample more uniformly over the range of the possible cascades by manipulating the cascade stopping probabilities. Instead of allowing chains to stop by themselves, the stopping is inhibited until a maximum number of cascade generations $g_{\max}$ is simulated, and then the chain stops. At each generation before $g_{\max}$, the line outages are recorded, and, although the chain does not stop, the probability that the state would have stopped is recorded. This gives many samples

[1] During the simulation, however, we allow lines not in Y_0 to outage again in successive generations except the next generation.

of the number of line outages for each of the generations $0, 1, 2, .., g_{\max}$, and these samples range from a small to a large number of line outages. And the probability of stopping at each intermediate length cascade can be calculated.

We now give the details of this improved simulation. Suppose the rth chain is simulated and is at state $x_k^{(r)}$ at generation $k < g_{\max}$. When the simulation samples from the probability distribution to obtain the next state $x_{k+1}^{(r)}$, it is easy to prohibit the choice $x_{k+1}^{(r)} = \{\,\}$ that would stop the chain. (This is equivalent to zeroing the probability of transition to $\{\,\}$ and renormalizing the probabilities of the other transitions.) It is also straightforward to record $x_k^{(r)}$ (which contains the lines outaged in generation k), and the probability $\sigma_k^{(r)} = \mathrm{P}[\text{transition from } x_k^{(r)} \text{ to } \{\,\}]$ that the chain stops when the state is $x_k^{(r)}$. $\sigma_k^{(r)}$ is the entry in the first column of the transition matrix P_k corresponding to $x_k^{(r)}$. The probability that the rth chain has exactly k generations is

$$q_k^{(r)} = (1 - \sigma_0^{(r)})(1 - \sigma_1^{(r)})...(1 - \sigma_{k-1}^{(r)})\sigma_k^{(r)} \tag{4.4}$$

More precisely, we have simulated (realized) one particular sequence of states $x_1^{(r)}, x_2^{(r)}, ..., x_k^{(r)}$ that avoid stopping. Now, conditioned on the states that do happen occurring in this sequence, we compute in the Markov chain that does not avoid stopping the probability of stopping at generation k with (4.4).

We indicate the first simulation of the rth chain by the superscript $(r; 1)$. We perform the first simulation of the rth chain up to generation $g_{\max}$ and extract results for each generation $k \leq g_{\max}$. For generation k, the total number of lines out is

$$n_k^{(r;1)} = \sum_{j=0}^{k} |x_j^{(r;1)}| \tag{4.5}$$

and the probability of $n_k^{(r;1)}$ lines out is equal to $q_k^{(r;1)}$, since the number of lines out increases at each non-stopping generation. Repeating the simulation of the rth chain t times for the same initial state $x_0^{(r)}$ gives different sequences $x_0^{(r;s)}, x_1^{(r;s)}, x_2^{(r;s)}, ...$ for $s = 1, 2, , ..., t$, generating many samples of the number of line outages $n_k^{(r;s)}$ and their probabilities $q_k^{(r;s)}$ for $s = 1, 2, ..., t$ and $k = 0, 1, ..., g_{\max}$. All these results are combined to give the distribution of the number of line

outages $N^{(r)}$ in the simulations of the rth chain:

$$\mathrm{P}[N^{(r)} = v] = \frac{1}{t}\sum_{k=0}^{g_{\max}}\sum_{s=1}^{t} I[n_k^{(r;s)} = v]q_k^{(r;s)} \tag{4.6}$$

where the indicator function $I[\cdot]$ limits the sums in (4.6) to the results giving v line outages. Thus (4.6) is the average of all the probabilities corresponding to the t possible occurrences of v line outages in the simulations of the rth chain.

Then, according to (4.3) and assuming the chains are independent, we evaluate the distribution of the total number of lines out N by convolving the distributions $N^{(1)}, N^{(2)}, ..., N^{(m)}$. The convolution is done by multiplying probability generating functions:

$$\mathrm{E}[z^N] = \mathrm{E}[z^{(\sum_{r=1}^{m} N^{(r)})}] = \prod_{r=1}^{m} \mathrm{E}[z^{N^{(r)}}] \tag{4.7}$$

The coefficient of z^v in $\mathrm{E}[z^N]$ is the probability $\mathrm{P}[N = v]$.

4.3 Probability distribution of load shed

The load shedding of a cascade is denoted as L. We want to estimate f_L, the probability distribution of load shed. We do this by conditioning on the number of line outages.

The number of line outages N ranges from ℓ_0 to $\ell_{\max}$, where ℓ_0 is number of lines in Y_0. We partition the range of N into K disjoint bins $B_1, B_2, ..., B_K$ so that

$$\{\ell_0, \ell_0 + 1, \ell_0 + 2, ..., \ell_{\max}\} = B_1 \cup B_2 \cup .... \cup B_K \tag{4.8}$$

We use the following subsections to obtain $f_{L|N\in B_\kappa}$, the distribution of load shed given that the number of lines out are in bin B_κ. The bins (4.8) are chosen large enough so that there is sufficient data in each bin to be able to approximate $f_{L|N\in B_\kappa}$.

From the distribution of N provided in section 4.2, we can easily evaluate the bin probabilities:

$$b_\kappa = \mathrm{P}[N \in B_\kappa] = \sum_{v\in B_\kappa} \mathrm{P}[N = v] \tag{4.9}$$

The idea is to evaluate the distribution of load shed f_L by conditioning on the number of lines in the bins:

$$f_L = \sum_{\kappa=1}^{K} b_\kappa f_{L|N\in B_\kappa} \tag{4.10}$$

We now use the OPA simulation to approximate $f_{L|N\in B_\kappa}$.

4.3.1 Load shed given the number of lines out

Given that the number of lines N outaged after cascading are in bin B_κ, we want to obtain the distribution of load shed $f_{L|N\in B_\kappa}$. We use a probability distribution of load shed because we are trying to estimate the risk of a future extreme event, and the power system loading condition, generator dispatch and maintenance status for a future event are uncertain and variable. This variability will produces different load sheds for the same line outages, or the same number of line outages.

The OPA model [76–79] has been validated to approximate well the observed bulk statistics of blackouts of WECC [80, 81]. Here, noting that our historical data is from part of WECC, we use the OPA results on a 1553 bus model of WECC to generate the conditional distributions of load shed $f_{L|N\in B_\kappa}$, $\kappa = 1, 2, 3, ..., 12$. Note that OPA is a long-term simulation that samples from a variety of grid loading conditions. The OPA results consist of 58 903 cascades. Each cascade yields the load shed and the number of lines out.

The OPA results are easily sorted into the bins $B_1, B_2, ..., B_{12}$ according to the number of lines outaged in each cascade. Bin B_κ has κ line outages for $1 \leq \kappa \leq 11$, and bin B_{12} has 12 or more line outages. Each bin has at least 83 data points. The empirical distribution for load shed in bin B_κ is fitted with the lognormal distribution $f_{L|N\in B_\kappa}$. Figure 4.1 shows three of these fits. The data points that have a fraction of load shed less than 0.01 are excluded. The mean μ and standard deviation σ of the lognormal distributions for the 12 bins are shown in Table 4.1. The mean and standard deviation increase as the number of line outages in cascades increase, as expected. Moreover, the Kolmogorov-Smirnov test for each bin's fitting has a p-value at least 0.1, so these fits are statistically significant.

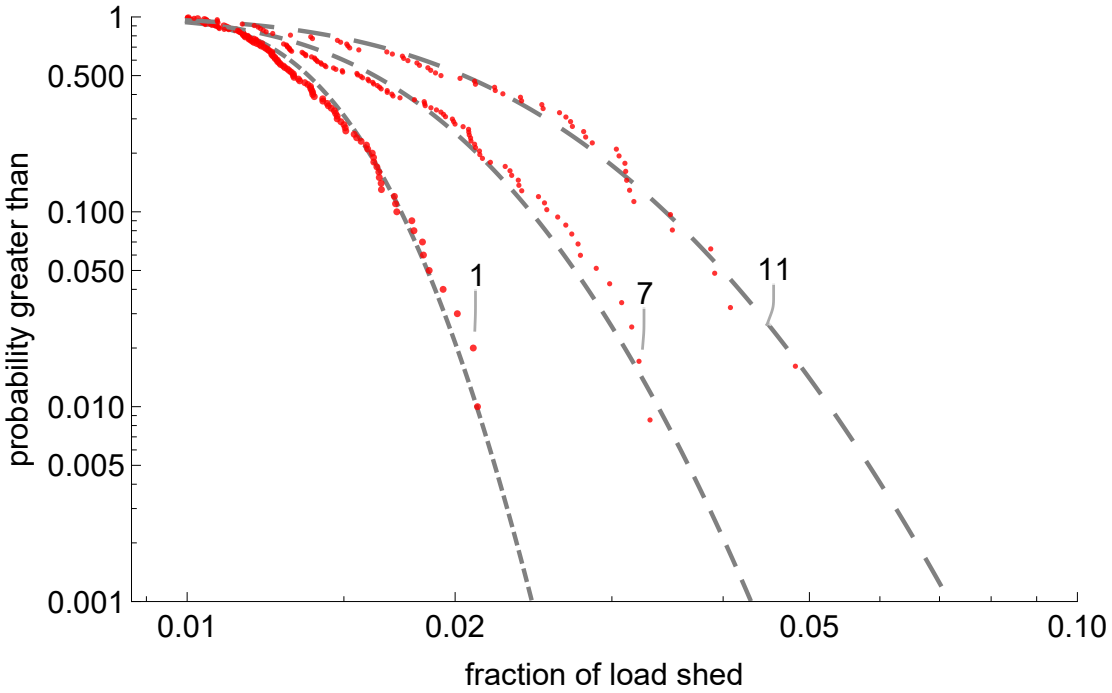


Figure 4.1: The survival functions of the distribution of fraction of load shed for cascades with 1, 7, or 11 line outages.

Table 4.1: Parameters of the lognormal distributions of the load shed given the number of line outages in the κth bin

κ	1	2	3	4	5	6	7	8	9	10	11	12
μ	-4.29	-4.26	-4.27	-4.22	-4.21	-4.18	-4.12	-4.18	-4.05	-3.99	-3.89	-3.75
σ	0.19	0.22	0.23	0.26	0.30	0.32	0.33	0.31	0.32	0.33	0.41	0.42

4.4 Results

4.4.1 Simulation of line outages

We use the improved sampling of section 4.2.1 to sample cascades from the Markovian influence graph formed from the utility data. Specifically, starting with assumed 3 initial outages, we simulate 100 cascades up to $g_{\max} = 100$ generations. Since the simulation also records data for each cascade stopping at any generation before 100 generations, this is equivalent to simulating 10 000 cascades, in which 100 cascades have 1 generation, 100 have 2 generations, and so on.

To contrast the improved sampling with straightforward sampling, we also simulate 10 000 cascades with the same initial outages using the straightforward sampling method of simply simulating until the cascade stops, with no special control of the stopping. The two simulations have close execution times. Figure 4.2 shows that the survival functions match except for some variability in the tail due to limited samples from the straightforward sampling. With the same simulation time, the improved sampling has two benefits: it has smaller standard deviations and generates more large cascade samples. For example, the standard deviation of the probability that cascades have more than 30 line outages is 0.00004 for improved sampling, and 0.0004 for straightforward sampling. As the number of line outages increases, this advantage is even more significant. The straightforward sampling focuses on the small cascades and does not sample enough large cascades to accurately estimate the large cascades. In contrast, the improved sampling samples uniformly across a full range of cascade sizes to better estimate a longer tail. The Markovian influence graph flexibly allows this improved sampling, addressing the straightforward sampling problem common in the literature of inherently undersampling large cascades.

Although in this work we only estimate the distribution of the number of lines out, there is a wealth of detailed information in the simulated cascades that could be useful.

4.4.2 Distribution of load shed

After estimating distribution of the number of line outages N, we proceed to estimate the distribution of load shed using the method described in Section 4.3. Subsection 4.3.1 calculates the conditional lognormal load shed distributions $f_{L|N\in B_1}, f_{L|N\in B_2}, ..., f_{L|N\in B_{12}}$. The distribution of N gives the bin weights $b_1, b_2, ..., b_{12}$ according to (4.9). Then the distribution of load shed f_L is the mixture of $f_{L|N\in B_1}, f_{L|N\in B_2}, ..., f_{L|N\in B_{12}}$ weighted by $b_1, b_2, ..., b_{12}$ as described by (4.10).

Figure 4.3 shows the survival function of the distribution of load shed f_L given 3 initial outages (red solid curve). In Figure 4.3, we also vary the number of initial outages to simulate

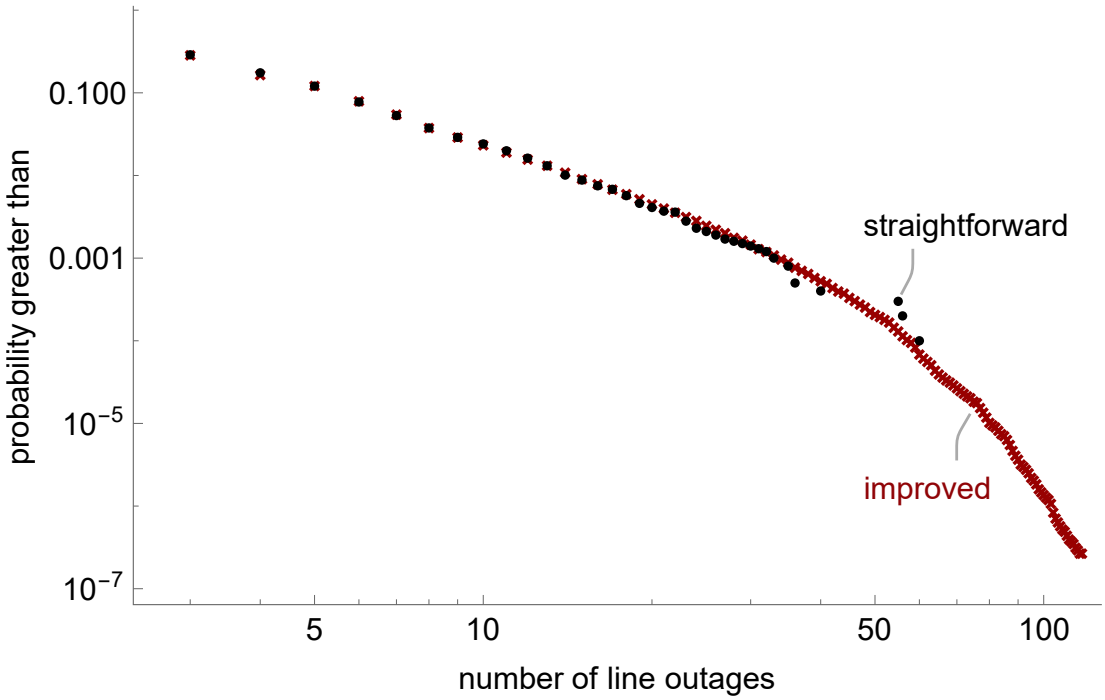


Figure 4.2: The survival function of the number of line outages N given 3 initial outages using the improved sampling (red crosses) and the straightforward sampling (black dots).

different initial line damage scenarios. As the number of initial outages increases, the probability of large load shed increases.

4.5 Comparing simulation driven by historical data with model-based simulation

This section describes and contrasts the strengths and weaknesses of model-based simulation and simulation of the Markovian influence graph driven by historical data.

Realism: A major limitation is that model-based simulations are practically constrained to approximate a limited subset of cascading mechanisms. The Markovian influence graph driven by historical data uses the statistics of real cascades, which encompass all the cascading mechanisms encountered in the historical period. It produces many cascades not observed in the real cascades. However, the Markovian influence graph does not describe the pairwise interactions between outages that could happen but that did not happen in the historical period. The power grid

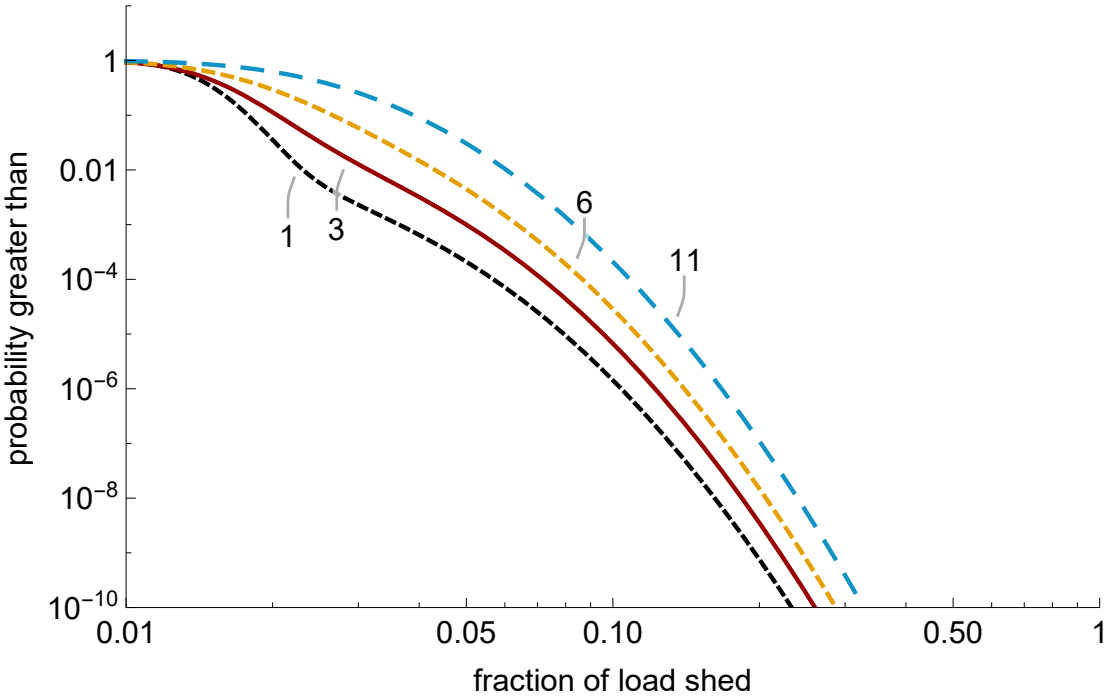


Figure 4.3: Survival functions of load shed with 1, 3, 6, or 11 initial outages.

slowly changes over the historical period as components age and upgrades to the grid and operational procedures are made. The Markovian influence graph pools together all the interactions in the power system over the historical period. For example, if an interaction was mitigated half way during the historical period, it still contributes a possible interaction to the Markovian influence graph.

Validation: When used to predict cascades, model-based simulation can produce cascades that are often judged to be credible, but most model-based simulations are not yet validated against historical cascade data. (One of the exceptions is the OPA simulation used in subsection 4.3.1, which is validated with WECC data in [80, 81].) An appropriate validation is reproducing the form of cascade statistics, and there is notable progress towards this goal [9, 10]. On the other hand, the Markovian influence graph describes the historical statistics of successive line outages, and reproduces the statistics of numbers of lines out [39], so important aspects of validation are inherent or already checked. There are additional assumptions in the processing of the historical data and the influence graph formulation, and some of these issues are discussed below. However,

the more subtle validations of the Markovian influence graph would seem to require more elaborate approaches to statistical validation.

Sampling of grid conditions: Another requirement, which is not always satisfied in the literature, is that model-based simulations should sample appropriately from a range of grid operating conditions [82]. Historical data inherently samples all the actual grid conditions encountered over the observed time, and this is often an appropriate sampling.

Sampling of cascades: Predicted cascading is inherently probabilistic due to the many interactions and protection actions that involve thresholding in an uncertain environment. Note that even "deterministic" model-based simulations can sample from cascade possibilities by randomizing the grid conditions. As regards sampling technique, the Markovian influence graph easily allows computing the rarer but riskier long cascades while tracking the outcomes and probabilities of all the truncations of the long cascades, as explained in subsection 4.2.1. A corresponding advantage in computing the largest cascades can be achieved for model-based simulation using splitting [83] or other methods.

Markov assumption: The Markovian influence graph only describes the statistics of successive Markov states in the historical cascades. Each Markov state is a specific line outage or set of line outages. The issue is the extent to which one can assume that knowing the state in a cascade generation is sufficient to approximate the statistics of which state is in the next generation. This is a pragmatic but fairly strong assumption.

Limited data: The Markovian influence graph is formed from historical data, which is limited in extent, especially for the higher cascade generations. This limitation can be partially mitigated [39], but not eliminated. In practice the higher generations are combined together in some ways to get sufficient data. Model-based simulations can, if not too detailed, produce larger amounts of cascading data.

Commonality between cascades: The Markovian influence graph describes the statistics of all types of cascades, but some of these may not be the cascades of interest. That is, there is an assumption that the same set of probabilistic cascading interactions tend to occur for all cascades.

In particular, statistical patterns in small cascades are to some extent extrapolated to large cascades. It is certainly possible to restrict the historical data to the subset of cascades of interest if the subset is large enough, but there is the tradeoff that as data set becomes smaller, estimation becomes more uncertain. In model-based simulation it seems easier to restrict the cascades simulated, but the challenges of validation for the restricted subset of cascades remain.

4.6 Conclusion

This chapter suggests a new form of cascading simulation driven by the detailed transmission line outage data that is routinely collected by utilities. This historical outage data is first processed into cascades and generations within cascades, and then used to form the Markovian influence graph that describes the statistics of outages in successive cascade generations as a Markov chain. Some initial line outages are assumed, and in this chapter these are the lines damaged by some extreme events, such as weather, fire, icing, or earthquake. Our immediate aim is to simulate and quantify the cascading of line outages after the initial damage. The Markovian influence graph is sampled to produce the simulated cascades. The simulated cascades are statistically similar to but more variable than the cascades in the historical data. The Markovian influence graph easily allows improved sampling that is more uniform across all sizes of cascades, and this gives better estimates of the large cascades that are rare but significant for cascade risk.

The Markovian influence graph produces cascades of specific line outages but no direct estimates of load shed. We show one way to estimate load shed by using a model-based simulation, OPA, to evaluate the probability distribution of load shed conditioned on the number of line outages. The distribution of load shed is then a weighted sum of these conditional distributions, with the weights determined by the line outage statistics produced by the Markovian influence graph. Other methods of estimating load shed can be developed and compared in future work. The combined result of the Markovian influence graph cascading simulation and the load shed estimation is the probability distribution of load shed for choices of specific initial lines damaged by the extreme event.

CHAPTER 5. TESTING THE MARKOVIAN INFLUENCE GRAPH

5.1 Testing large cascade mitigation by the Markovian influence graph on simulations

Chapter 3 forms a Markovian influence graph that models cascades statistically from historical outage data and mitigates long cascades by upgrading several critical lines [39]. The cascade length is defined as the number of cascade generations. When the mitigation is expressed in the Markovian influence graph, it does roughly halve the frequency of long cascades. Thus the mitigation is self-consistent, but one obviously cannot do any further testing on the real system. And load shed has not been considered because this information is not available in the historical outage data.

Simulation is indicated to further test the mitigation. This section aims to test the Markovian influence graph by applying the Markovian influence graph to simulated cascades, calculating the mitigation, and testing the effectiveness of the mitigation by simulation. It uses several different cascading simulations [15,16,81] on several different power systems. Using several simulations and systems is more thorough and tends to mitigate arguments against the modeling in individual simulations. The simulation of the IEEE 118-bus system is prepared by Junjian Qi, Stevens Institute of Technology, the simulation of the Polish 2383-bus system is prepared by Paul D.H. Hines and Molly Rose Kelly-Gorham, University of Vermont, and the simulation of the WECC 1553-bus system is prepared by Benjamin A. Carreras, BACV Solutions Inc.

5.1.1 Cascading outage models

This section briefly summarizes the cascading models used in the study.

The cascade model used in Polish 2383-bus system case, which is proposed in [84], is based on the DC power flow and simulates cascading caused by overload. It incorporates an overcurrent

relay at each transmission line to determine if a line outages due to overload. As shown in Figure 5.8, the cascade model is initialized by calculating pre-contingency power flow on the $N-1$ secure Polish power system; then, a multiple contingency is applied by modifying system susceptance matrix to reflect the remove of failed lines; if no system failure, where system failure is defined as as a state in which at least 10% of the buses are no longer connected to the largest island, the cascading model re-dispatches generators and re-calculate power flow by considering possible islanding; next, the model updates the time-delayed overcurrent relay at each transmission line to determine whether and when a line trips because of overload; if some lines are outaged, the model updates the system susceptance matrix, and repeats the process until no line outages or a system failure is detected.

Closed-loop OPA, which is described in [85–87], models the complex dynamical evolution of a power system. It contains two timescales: a fast timescale modeling the cascading outages, which corresponds the inner loop in Figure 5.9; and a slow timescale modeling the evolution of the power system, which corresponds to the outer loop in Figure 5.9. The fast timescale has the same function as the cascade model for Polish 2383-bus system case; however, OPA uses optimal DC power flow to re-dispatch generators and determine overload of transmission lines. An overloaded line outages with a specified probability. In the slow timescale, the power system has a slowly increasing electricity demand, and the reliability of transmission lines are increased through updating after a blackout. Specifically, at the beginning of each day, the load is increased at a rate of 2% per year, and the max generation increases when the capacity margin decreases below a given critical level $\Delta P/P$. After a blackout, lines involved in it have their flow limit increased slightly by multiplying a parameter b.

Open-loop ACOPA is the inner loop of the flowchart in Figure 5.9, and it uses AC optimal power flow instead of DC optimal power flow [88]. ACOPA is "open-loop" because it uses a fixed power system and does not represent any evolution of the power system in respond to blackouts.

5.1.2 Procedure of testing the influence graph mitigation on simulated data

The steps of testing the influence graph mitigation on simulated data are:

1. Generate cascade data by simulation.

   Each of the simulations generates a large sample of cascades recording the specific lines outaged and load shed at each cascade generation.

2. Form the influence graph for each case and suggest critical lines to be mitigated.

   Form an influence graph from the simulated cascade data in terms of number of generations using the method in [39]. Use the influence graph to identify ten critical lines for mitigating long cascades as in [39]. Then upgrade these critical lines statistically in the influence graph, and estimate using the influence graph the mitigation amount in terms of reduction of long cascades. This list of critical lines and how much impact they can have on is what to be tested.

3. Upgrade critical lines in simulation and resimulating.

   After critical lines are upgraded, these lines are less likely to outage due to other outaged lines and it is represented in the simulations. To decrease the probability that this line outages due to other outaged lines, we increase the flow limit of a line [16], or decrease the probability that an overloaded line will outage [81] [15]. Then the case is resimulated with the mitigation in place.

4. Compare mitigation results of the influence graph and resimulation.

   Compare the distribution of cascade size in terms of the number of generations and load shed using the influence graph and resimulation after mitigation. This aims to find out to what extent the influence graph mitigation works in each simulation in terms of cascade length and load shed and risk reduction. The overall correlation between cascade length and load shed is also determined.

The outage data is organized as a matrix. Each row is a record of a line outage which includes X1=line ID, X2=cascade ID, X3=generation/iteration number, X4=(load shed)/(power demand)

in this iteration. If there is a multi-line outage, then X4 should be the same for each line in this multi-line outage.

An example of the outage data is:

Table 5.1: An example of outage data

line ID	cascade ID	generation	(load shed)/(power demand)
1	1	1	0
2	1	1	0
3	1	2	0.3
6	1	3	0.1
2	2	1	0.2
4	2	1	0.2
3	2	2	0

In this example, there are two cascades. The first cascade has line 1, 2, 3, 6 outaged. Line 1 and line 2 outaged in the first generation which are initial outages. There is no load shed in the first generation. Then line 3 outaged in the second generation, associated with a 30% percent load shed of the total power demand. Finally, line 6 outaged in the third generation. When line 6 outaged, 10% percent load shed occurred. The first cascade has 40% load shed in the end. The second cascade has line 2,3,4 outaged. Line 2 and line 4 outaged simultaneously. They caused 20% load shed. The second cascade has 20% load shed in the end.

5.1.3 IEEE 118-bus system

5.1.3.1 Statistics of simulated cascades

The IEEE 118-bus system has 118 buses and 186 lines. The open-loop ACOPA model simulates cascading outages and produces 20,000 cascades and 48,546 outages. All lines have outaged at least once. The largest cascade has 6 generations. The smallest cascade has 1 generation, and 60% of cascades only have one generation. The simulated data have single line outages and simultaneous outages with multiple lines. The proportion of distinct sets of simultaneous outages is 97%.

5.1.3.2 Influence graph identifies critical lines

This section forms the influence graph based on the simulated cascade data. The influence graph can capture the distribution of cascade size distribution, as shown in Figure 5.1. For the convenience of communication, this study groups cascades into small, medium, and large cascade such that the log probabilities are roughly sitting on a line.

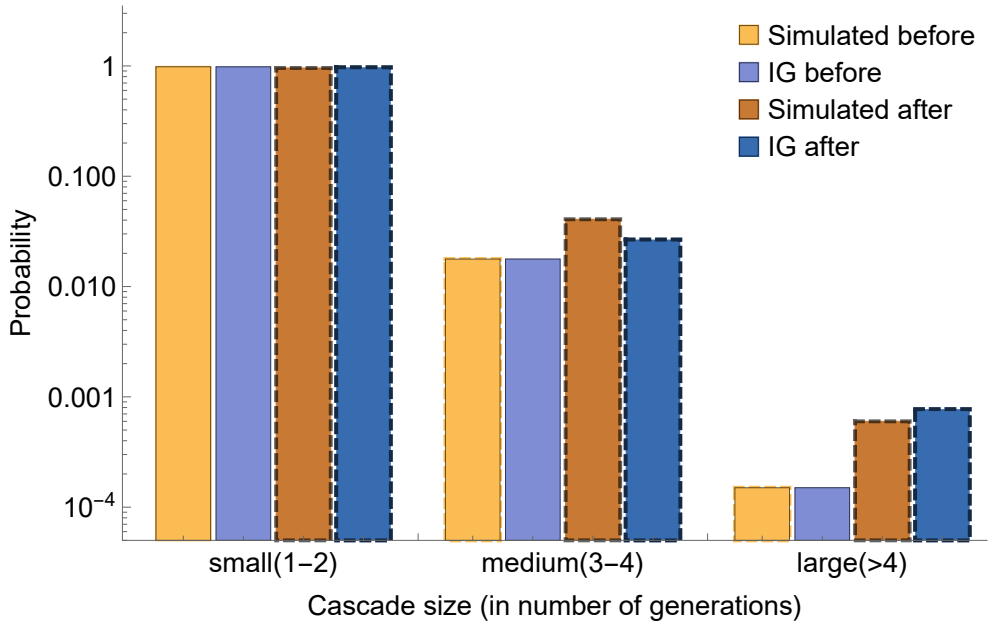


Figure 5.1: Comparing probabilities of small, medium and large cascades from the open-loop ACOPA simulation and the influence graph (IG) before and after mitigation (IEEE 118-bus system case).

The influence graph is an absorbing Markov chain. The transition matrix is

$$\boldsymbol{P} = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline \boldsymbol{u} & & \boldsymbol{Q} & \end{array}\right] \tag{5.1}$$

The submatrix $\boldsymbol{Q}$ corresponds to transient states. By the Perron-Frobenius theorem [89], it has a unique largest eigenvalue μ which is positive and simple. The corresponding eigenvector's elements are all real and positive. $\boldsymbol{P}$ has a unique largest eigenvalue 1, and the second largest

eigenvalue is the same as the largest eigenvalue of $\boldsymbol{Q}$. let λ_i be the i-th eigenvalue of $\boldsymbol{P}$, then $\lambda_1 = 1$, $\lambda_2 = \mu$, and $1 > \mu > \lambda_3 \geq \lambda_4 \geq \cdots$. Let $\boldsymbol{w}_i$ be the right eigenvector corresponding to λ_i, and $\boldsymbol{v}_i'$ be the left eigenvector. Define

$$\boldsymbol{W} = \begin{bmatrix} & & & \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_n \\ & & & \end{bmatrix} \tag{5.2}$$

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_n' \end{bmatrix} \tag{5.3}$$

Assume $\boldsymbol{P}$ is diagonalizable, then $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{V}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with eigenvalues as entries. Note $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ are normalized accordingly so that $\boldsymbol{W}\boldsymbol{V} = \boldsymbol{I}$. Then, the state distribution at step t is

$$\begin{aligned} \boldsymbol{\pi}_t &= \boldsymbol{\pi}_0 \boldsymbol{P}^t = \boldsymbol{\pi}_0 \boldsymbol{W} \boldsymbol{\Lambda}^t \boldsymbol{V} \\ &= \boldsymbol{\pi}_0 \left(\lambda_1^t \boldsymbol{w}_1 \boldsymbol{v}_1' + \lambda_2^t \boldsymbol{w}_2 \boldsymbol{v}_2' + \lambda_3^t \boldsymbol{w}_3 \boldsymbol{v}_3' + \cdots \right) \\ &= \boldsymbol{\pi}_0 \left(1 \, [1 \quad 1 \cdots 1]' \, [1 \quad 0 \cdots 0] + \mu^t \boldsymbol{w}_2 \boldsymbol{v}_2' + \lambda_3^t \boldsymbol{w}_3 \boldsymbol{v}_3' + \cdots \right) \\ &= [1 \quad 0 \cdots 0] + \boldsymbol{\pi}_0 \left(\mu^t \boldsymbol{w}_2 \boldsymbol{v}_2' + \lambda_3^t \boldsymbol{w}_3 \boldsymbol{v}_3' + \cdots \right) \end{aligned} \tag{5.4}$$

(5.4) shows that the Markov chain eventually goes to absorption. The convergence speed depends on the eigenvalue μ. We have shown in [39] that the state distribution conditional on the chain not going to absorption is asymptotically distributed according to the left eigenvector $\boldsymbol{v}_2'$ corresponding to the largest eigenvalue μ of $\boldsymbol{Q}$. The convergence speed depends on the difference between μ and λ_3. (5.4) also shows that the state distribution has some dependence on the initial state distribution $\boldsymbol{\pi}_0$ and the value of μ. However, note that the quasi-stationary distribution does not depend on the initial state distribution.

In the IEEE 118-bus system case, the first three eigenvalues of $\boldsymbol{Q}$ are $\lambda_1 = 1$, $\lambda_2 = \mu = 0.10471$, $\lambda_3 = -0.10466$. As μ is far from 1, this system would go to absorption very

fast. The difference between μ and λ_3 is less than 0.0001, so the system would take a long time before converging to the quasi-stationary distribution. That is, we are unlikely to observe only the quasi-stationary distribution if we sample this Markov chain.

Furthermore, the fourth eigenvalue λ_4 equals to 0.0591. There is a large enough gap between λ_3 and λ_4, hence, in a short time, the quasi-stationary distribution can be approximated by the second and third left eigenvectors.

We consider a cascade is long enough when the second component $\boldsymbol{\pi}_0\mu^t\boldsymbol{w}_2\boldsymbol{v}_2'$ is nine times greater than the third component $\boldsymbol{\pi}_0\lambda_3^t\boldsymbol{w}_3\boldsymbol{v}_3'$. The time at which this occurs can be approximated by (5.5).

$$\left(\frac{\mu}{|\lambda_3|}\right)^{t_l} \geq 10 \tag{5.5}$$

In IEEE 118-bus system case, $t_l \geq 4821$. Considering $\mu = 0.10471$, a cascade stops in several generations, so we never see long cascades.

Comparing to the BPA case, the top three eigenvalues are $\lambda_1 = 1, \lambda_2 = \mu = 0.502, \lambda_3 = 0.381$. Using the same criteria as in (5.5), after $t_l = 9$ generations, the conditional distribution converges to the quasi-stationary distribution.

The left eigenvector $\boldsymbol{v}_2'$ corresponding to the dominant eigenvalue μ of $\boldsymbol{Q}$ is the asymptotic probability distribution that lines involved in large cascades. We take the top ten lines as critical lines. In the IEEE 118-bus system, they are:

critical lines: $101, 46, 102, 100, 10, 72, 121, 124, 45, 142$

5.1.3.3 Upgrading critical lines in simulation and influence graph

Cascading outage risk is mitigated by upgrading critical lines. The simulation models the upgrading by reducing the probability that an overloaded line outages, which is called the triggering probability. This will affect the transition probabilities to the states that include critical lines. An example is used to discuss the relationship between triggering probabilities in OPA and transition probabilities in IG.

Suppose a state s_1 contains multiple line outages and a stop state $\emptyset$ represents the cascade stopping. We want to figure out the transition probability from s_1 to other states $s_2 = \{j\}$ and $s_3 = \{j, k\}$, which represent line j outaged and line j, k outaged in the same generation, respectively. We can represent this transition by the diagram in Figure 5.2. Whatever s_1 is, we can define a overload probability $p_{overload}$ and an outage probability p_{outage}. $p_{overload}$ is the probability that lines overload in a generation given another state in previous generation; p_{outage} is the probability that lines outage given these lines overloaded. Then, the transition probability is the product of the overload probability and the outage probability. In Figure 5.2, $p_{s_1,j} = p_{overload,j} \times p_{outage,j}$; assuming outage probabilities are independent for different lines, then $p_{s_1,jk} = p_{overload,jk} \times p_{outage,j} \times p_{outage,k}$.

In OPA simulation, modeling is represented by reducing the outage probability. Suppose j is a critical line, and k is not. Consider an extreme case, in which $p_{outage,j}$ is reduced to 0. Then $s_1 \to j$ becomes $s_1 \to \emptyset$, and $s_1 \to \{j, k\}$ becomes $s_1 \to \{k\}$.

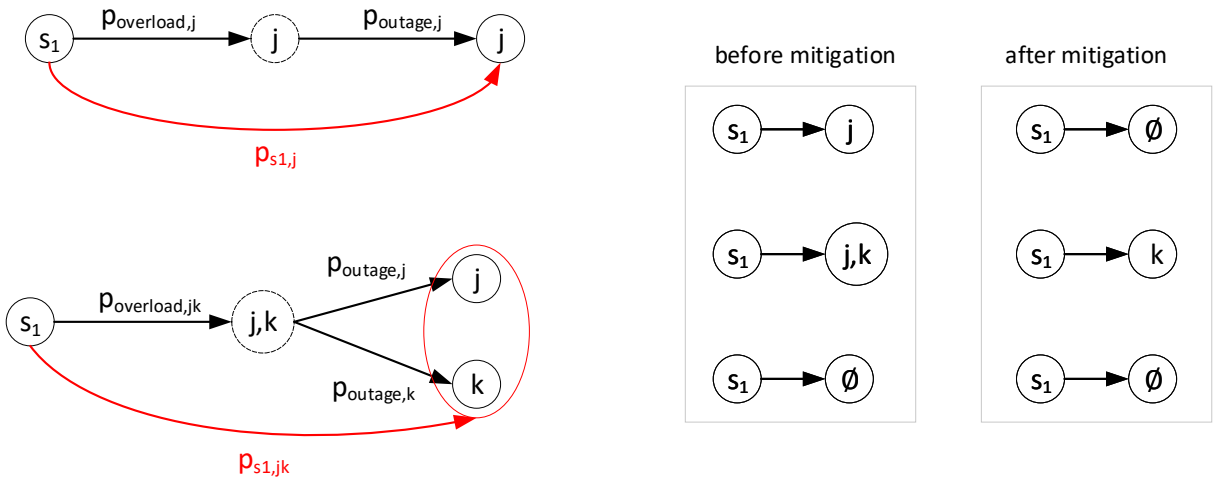


Figure 5.2: A simple diagram that illustrates the relation between triggering probabilities in OPA and transition probabilities in IG (the IEEE 118-bus system case).

Therefore, the relation between triggering probabilities p_{outage} in OPA and transition probabilities in IG is that reducing the triggering probability will reduce the transition probability to states including critical lines. The reduction of the probability to states only including critical

lines is equivalent to the increase of the transition probability to the stop state, while the reduction of the probability to states including critical lines and other lines is equivalent to the increase of the transition probability to states only including other lines.

The above discussion shows the general changing trend of transition probabilities when transmission lines are upgraded. However, it is still too complicated to quantify the change of each specific transition probability. This study uses a simple method by comparing overall changes of transition probabilities to critical lines before and after upgrading critical lines in simulation.

We first define the average aggregated transition probability to critical states γ. Critical states are those states in the influence graph that contain any critical lines. Then,

$$\gamma = \frac{\sum_{i=2}^{n} \sum_{j \in \text{critical states}}^{n} \boldsymbol{P_1}+[i,j]}{n-1} \tag{5.6}$$

where n is the number of states. As discussed in Chapter 3, two matrices, corresponding to the initial generation and dependent generations, respectively, are the foundation for constructing the transition matrices. Therefore, we inspect the reduction of transition probabilities in these two matrices. It turns out that the average aggregated transition probability to critical states is 0.24 before mitigation and 0.05 after mitigation, which is decreased by 78% for the first matrix; while there is no reduction for the second matrix.

Then, the study updates the transition matrices of the influence graph to represent the mitigation according to the aforementioned relation. Specifically, this study decreases the transition probabilities to critical lines by 78% in the matrix corresponding to the initial generation and adjusts the stopping probabilities and other non-stopping probabilities accordingly to make the matrix still a transition matrix.

Finally, we compare the computed cascade size distribution with the simulation result with the same initial outages as shown in Figure 5.1.

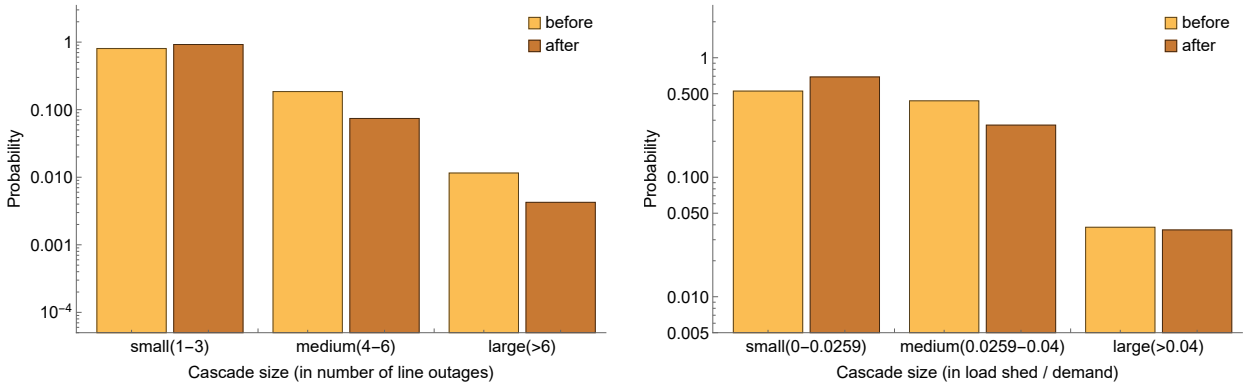
5.1.3.4 Mitigation effect

This study evaluates the mitigation effect by the reduction in cascade sizes after mitigation. Cascade sizes are measured by number of generations, number of line outages, and load shed.

Number of generations is a simple and mathematically convenient measurement. Number of line outages is a straightforward measurement. And load shed is a directly related to the cascading impact to customers. Figure 5.1 and 5.3 show the mitigation effect in three measurements.

The probability of large cascades in terms of number of line outages is decreased, and the number of cascade with more than one generation is reduced by 50%. . However, the probability of large cascades in terms of number of generations are increased, and there is almost no mitigation effect in terms of load shed. That the mitigation is not achieved in terms of number of generations and load shed is a result of the modeling of upgrading. Specifically, the upgrading is modeled by reducing the triggering probability, however, it cannot alleviate the stress of the overloaded system. For example, a generation has three line outages $\{46, 102, 154\}$ in a simulation. If lines 46 and 154 are upgraded by using a reduced triggering probability, in the new simulation, lines 46 and 154 do not outage in that generation. However, this cascade does not stop because the system condition is still stressful and line 46 is selected to be disconnected due to overloading in a subsequent generation.

However, it does not mean there is no mitigation. In terms of line outages, the probabilities of large and medium cascades are both reduced. And this mitigation also reduced the probability that initial outages propagate further.



(a) Cascade size measured in line outages.

(b) Cascade size measured in fraction of load shed (load shed / total demand).

Figure 5.3: Comparing empirical cascade size distribution before and after mitigation with the same initial outages in simulation (IEEE 118-bus system case).

5.1.4 Polish 2383-bus system

5.1.4.1 Statistics of simulated cascades

The Polish system has 2383 buses and 2896 lines. The simulated data has 7,692 cascades and 525,041 outages. 632 lines have outaged at least once. The largest cascade has 149 generations. The smallest cascade has 19 generations. The simulated data have single line outages as dependent outages and simultaneous outages as initial outages, in which the proportion of different sets of simultaneous outages is 50%.

5.1.4.2 Influence graph identifies critical lines

This section forms the influence graph based on the simulated cascade data. The influence graph can capture the distribution of cascade size distribution, as shown in 5.4. As we did in IEEE 118-bus system case, we group the cascades into small, medium and large cascades.

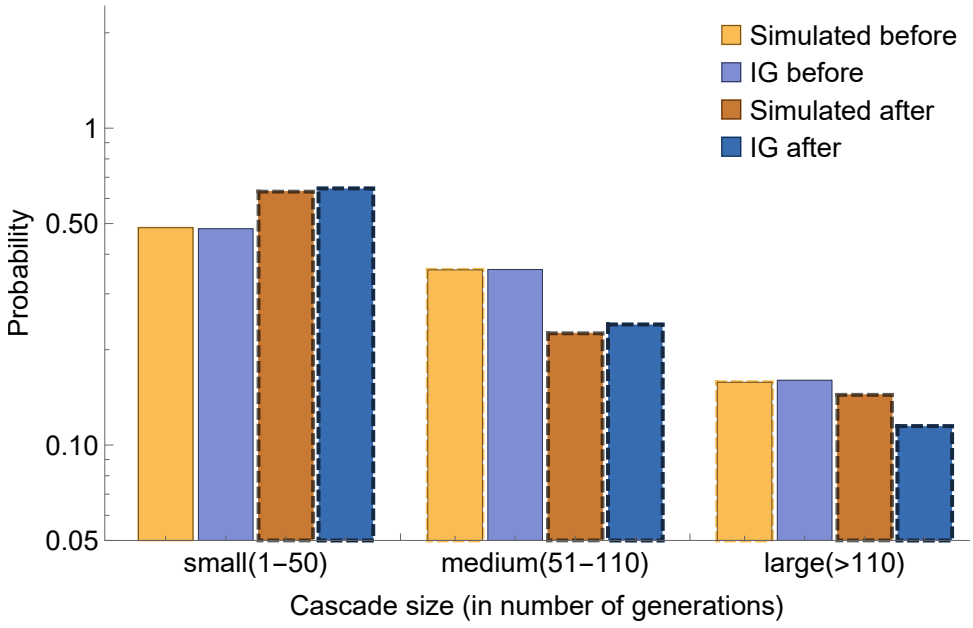


Figure 5.4: Comparing probabilities of small, medium and large cascades from simulation and the influence graph before and after mitigation (Polish 2383-bus system case).

Then, we analyze the eigenvalues and eigenvectors of matrix $\boldsymbol{Q}$ to identify the critical lines. The largest and the second largest eigenvalues of submatrix $\boldsymbol{Q}$ are 0.983 and 0.809, respectively.

There is a gap between the two eigenvalues, so the left eigenvector corresponding to the dominant eigenvalue is the asymptotic probability distribution that lines involved in large cascades. We take the top ten lines as critical lines. They are:

$$\text{critical lines:} \qquad 169, 2309, 1833, 617, 543, 168, 476, 455, 24, 2109$$

The Polish 2383-bus system case has $t_l \geq 12$ by formula 5.5. Therefore, simulated cascades are long cascades. This result is consistent with the purpose of the cascading model, as the Random Chemistry algorithm generates initial outages that lead to long cascades [84].

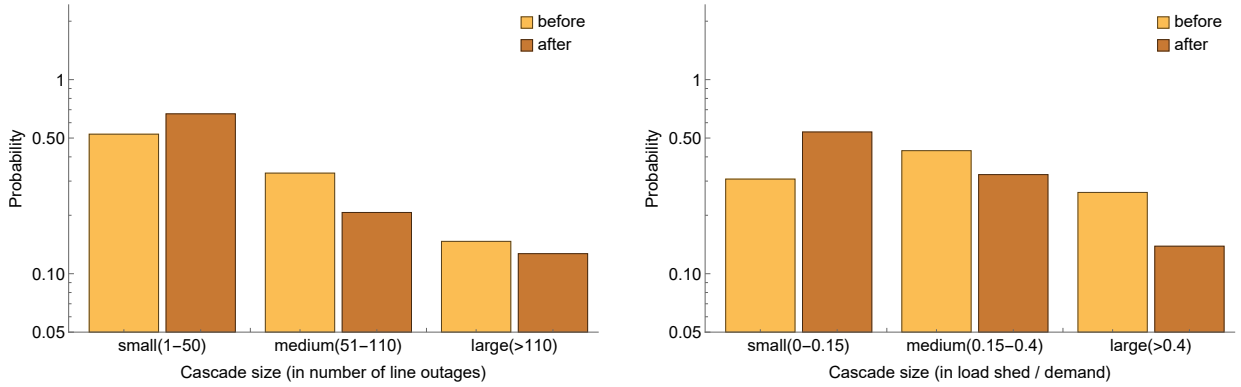
5.1.4.3 Upgrading critical lines in simulation and influence graph

Cascading outage risk is mitigated by upgrading critical lines. The simulation models the upgrading by doubling the line flow limits and rerunning with the same initial outages. This will reduce the transition probabilities to the states that include critical lines. When a line flow limit is increased, this line is less likely to overload. In some cases, this line outaged simultaneously with other lines before upgrading; while after upgrading, this line may not outage due to overloading. Thus, the transition probability to states including this line becomes smaller. This modeling of upgrading is different from that in the IEEE 118-bus system case: the Polish 2383-bus system system case reduces the probability that a line overloads, while the IEEE 118-bus system case does not reduce the probability of overloading but the probability that an overloaded line outages. It turns out that the average aggregated transition probability to critical states γ is decreased by 0.9% for the first matrix and 23% for the second matrix.

As in IEEE 118-bus system case, we decrease the transition probabilities to critical lines by 0.9% in the matrix corresponding to the initial generation and 23% for the second matrix, and adjust the stopping probabilities and other non-stopping probabilities accordingly to make the two matrices still transition matrices. Finally, we compare the computed cascade size distribution to the simulation result with the same initial outages as shown in Figure 5.4.

5.1.4.4 Mitigation effects

The mitigation effect is tested by simulation. Figure 5.4 shows the cascade size distribution before and after mitigation in terms of number of generations, and Figure 5.5 shows the mitigation effect in terms of the other two measurements.



(a) Cascade size measured in line outages.

(b) Cascade size measured in fraction of load shed (load shed / total demand).

Figure 5.5: Comparing empirical cascade size distribution before and after mitigation with the same initial outages in simulation (Polish 2383-bus system case).

The probabilities of the middle and large cascades are both reduced. The criteria of grouping cascades into three categories in terms of number of generations and number of line outages are the same because most of the dependent outages are single-line outages.

5.1.5 WECC 1553-bus system

The WECC 1553-bus system has 1553 buses and 2114 lines. The simulated data has 29,365 cascades and 44,877 outages. The largest cascade has 15 generations. The smallest cascade has 1 generation.

Figure 5.6 shows that the influence graph can reproduce the statistics of the cascading model. The influence graph indicates that the mitigation effect is small. Specifically, if we reduce the transition probabilities to critical lines, which is identified by the quasi-stationary distribution of the Markovian influence graph, the probability of large cascades is slightly decreased. This can be explained by the probability that a line involves in large cascades, which is the value of the

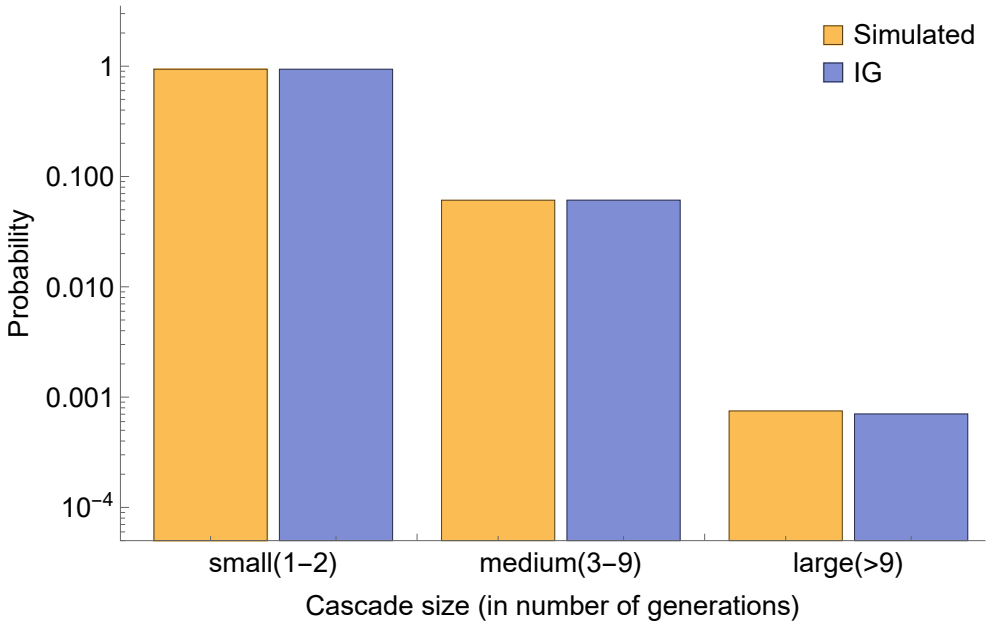


Figure 5.6: Survival functions of the number of generations from real data and from the Markov chain (WECC 1553-bus case).

quasi-stationary distribution of the Markovian influence graph. The probabilities for critical lines quantifies the criticality of these lines and determines the amount of mitigation. Figure 5.7 shows quasi-stationary distributions for two systems: the WECC 1553-bus system and the BPA system. The Markovian influence graph for the BPA system is formed from 14 years of historical outage data in Chapter 3. The BPA system is contained within WECC, but the netowrk models have some differences where they overlap, and the largest cascades for WECC will differ from the largest cascades restricted to BPA. Moreover, OPA computes the complex system "steady state" cascading after network evolution whereas the BPA system describes cascading over a historical period. The probabilities for BPA system have a wider range than WECC 1553-bus system. This shows that lines in WECC 1553-bus system are more similar in terms of cascading criticality to the system than in BPA system. Moreover, large probabilities in BPA system are much greater than that in WECC 1553-bus system, which shows that critical lines in WECC 1553-bus system is less critical to WECC system than critical lines in BPA system to their own system. The difference is due to the characteristic of the closed-loop OPA model for WECC 1553-bus system.

The closed-loop OPA involves a slow timescale describing the evolution of the grid, especially lines are upgraded after their failure. Therefore, lines in closed-loop OPA model tends to be similar in cascading criticality to the system.

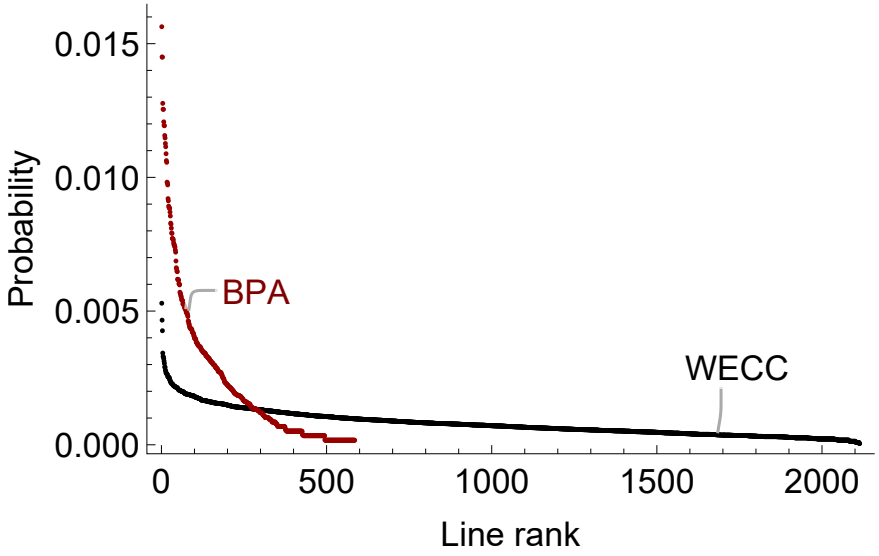


Figure 5.7: Quasi-stationary distributions for WECC 1553-bus system and BPA system.

5.1.6 Conclusion and discussion

This section tests the influence graph on three different cascade simulations. The simulations use different cascade models on three power systems. The results confirm that the Markovian influence graph reproduces the probabilities of small, medium, and large cascades. Moreover, the simple modeling method of upgrading in influence graph captures the mitigation effect in simulations. The upgrading in simulations includes reducing the overloaded line triggering probability and increasing line flow limit. Both of them have complex effect on propagation of cascading outages. The influence graph represents the upgrading by adjusting transition probabilities to upgraded lines. This simple method models the mitigation effect on the probability of different size cascades.

However, the mitigation effect is dependent on the modeling of cascades and the measurement of cascade sizes. The closed-loop OPA continuously increases line flow limits involved in previous

cascades, hence, the influence graph indicates that only a small mitigation amount can be achieved by upgrading critical lines.

Open-loop ACOPA simulating the IEEE 118-bus system models the upgrading by decreasing the triggering probability, which mimics blocking a zone 2 or zone 3 relay. The probability of large cascade size measured in number of generations becomes larger than before after mitigation. The reason is complicated. First, cascades with 5 or more generations are rare (less than 0.5% of total cascades). Second, close inspection of each cascade shows that the upgrading does not mitigate the power system stress in some cascades but makes multiple simultaneous outages become a sequence of single outages, which increases the number of generations of these cascades but actually does not make the situation severe. However, the overall mitigation is effective, and it is more obvious when the cascade size is measured in number of line outages.

The mitigation effect is more obvious when the cascade size is measured in number of line outages. However, the cascading failure simulator on the Polish 2383-bus system simulates long cascades; the mitigation effect is shown in medium and large cascades, and the load shed of large cascades is reduced most.

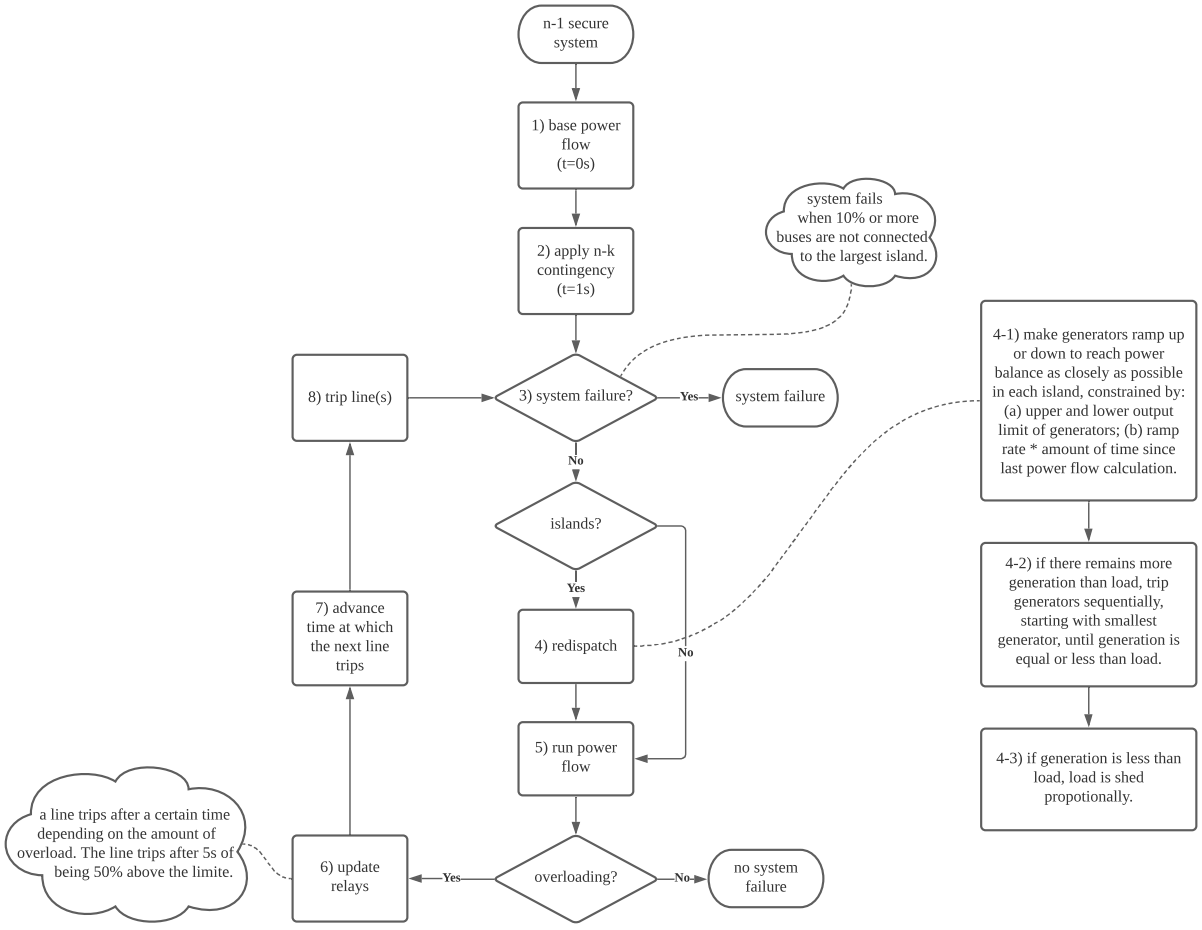


Figure 5.8: Flowchart of cascading failure simulator used in Polish 2383-bus system case.

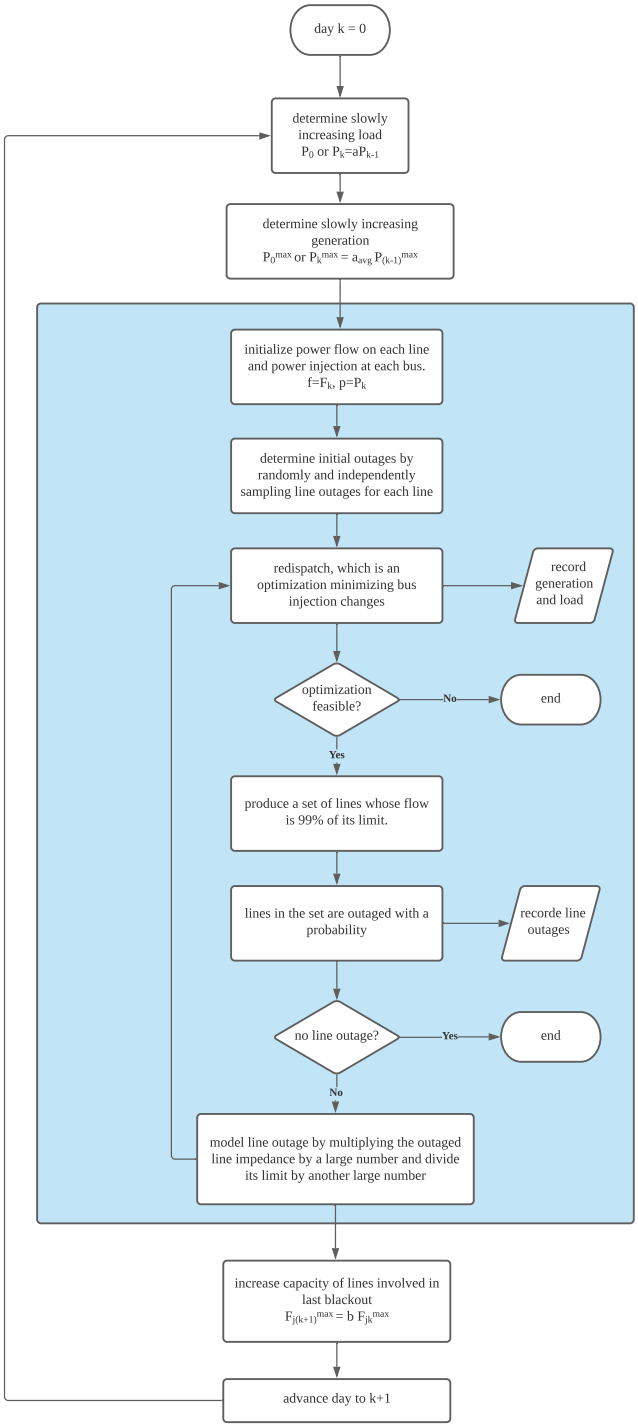


Figure 5.9: Flowchart of the closed-loop OPA model. The inner loop (in light blue area) corresponds to the open-loop OPA model.

5.2 Testing the assumption of the influence graph

A big assumption of the Markovian influence graph is that current line outages depend only on preceding line outages. Although all line outages determine the state of the power system, the history before the preceding line outage is lost in the influence graph. This chapter tests this assumption by comparing the influence graph with the kinetic monte carlo (KMC) cascading simulation proposed by Argonne National Lab [90], which is also a Markov chain but depends on all line outages before current line outages. The KMC simulation runs on the IEEE 118-bus system in this chapter and was prepared by Mihai Anitescu, Albert Lam, and Jake Roth at Argonne National Lab.

5.2.1 Overview of DAG generated by KMC

The KMC model describes line failure rates and organizes them into a directed acyclic graph (DAG). DAG enumerates all possible failure paths between the fully operational network topology (where no lines have failed) and the fully-degraded network topology (where all lines have failed). Each node of the DAG is a set of outaged lines, and the edge connects two nodes in which the end node has one and only one more line outage than the starting node. An example of the DAG is shown in Figure 5.10. Each edge is weighted with the transition rate from the starting node to the end node, which is also the failure rate of the new line in the end node given the starting node. In principle, the DAG enumerates all possible line outages combinations. That is, there are 2^n nodes in the DAG for a n-line system.

5.2.2 The relationship between DAG and the influence graph

The influence graph takes single line outages in one generation as nodes and the conditional probabilities of the end node given the starting node as edges. It defines a rigorous Markov chain and has nice properties. We can use the influence graph to compute the distribution of small, medium, and large cascades, with cascade size measured in terms of number of generations or number of lines. Moreover, a quasi-stationary distribution of the influence graph can identify the

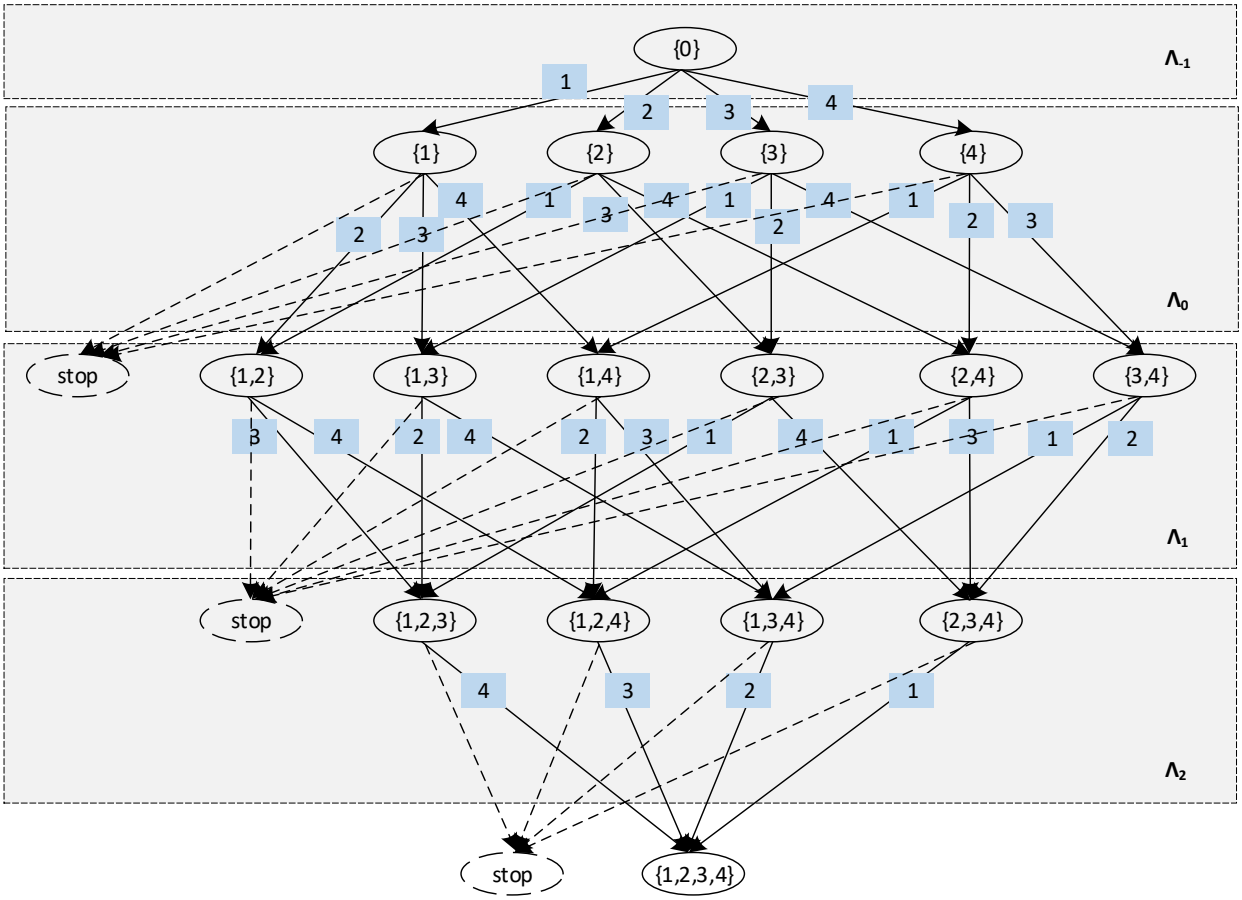


Figure 5.10: DAG of a four-line system. The light blue box is the outaged line in the current generation. Stop nodes represent the cascade stopping at the current state.

lines most critical in propagating long cascades. The influence graph can then evaluate the effect of upgrading these critical lines on the distribution of cascade size [39].

Figure 5.10 illustrates the relationship between the edges in DAG and the vertices in the influence graph using a 4-line system. The DAG is organized hierarchically. Ellipses are vertices, and there are $2^4 = 16$ vertices plus 1 stop vertex [1]. Layer Λ_{-1} represents the initial state of the system. Layer Λ_0 contains initial outages, which are single line outages. Layer Λ_1 contains cases when two lines have outaged. Layer Λ_2 contains cases when three lines have outages. And Layer Λ_3 presents the case when the whole system is done. The layers are joined by edges. An edge

[1] Figure 5.10 shows three stop vertices just in order to plot a clear graph. Otherwise, every vertex connects to the stop vertex, which causes many edges and vertices to overlap.

between two vertices represents the failure rate of the new outaged line in head vertex given the line outages in tail vertex. For example, the edge from vertex $\{1\}$ to vertex $\{1, 2\}$ represents the failure rate of line 2 given that line 1 has outaged. Dashed edges indicate the stop rates; that is, the rate at which cascades stop at given vertices.

There is a correspondence between edges of DAG and vertices of the influence graph. Vertices of the influence graph are light blue boxes in Figure 5.10. A light blue vertex is the new outaged line in the head ellipse given the tail ellipse. It also corresponds to the failure rate of that new outaged line given the tail vertex in DAG. According to edges connected to layer Λ_0 (including edges from Λ_{-1} to Λ_0 and from Λ_0 to Λ_1), we can derive the transition matrix from generation 0 to generation 1. The failure probability that is product of failure rate and time of DAG is equivalent the transition probability from generation 0 to generation 1 in the influence graph. It is complicated for transition matrix from generation 1 to generation 2 of the influence graph because failure probabilities are conditioned on two line outages and they are from two paths. For example, edge from vertex $\{1, 3\}$ to $\{1, 2, 3\}$ indicates the new outaged line is line 2, but there are two possibilities, which are line 1 outaged in previous step or line 2 outaged in previous step. If we want to find the transition probability from line 1 to line 2, we must distribute the probability of from vertex $\{1, 3\}$ to $\{1, 2, 3\}$ to the path that line 1 outaged in previous step. It is even more complicated for transition matrix from generation 2 to generation 3. A rigorous derivation is shown in Section 5.2.3.

The DAG defines a Markov chain between all the possible states 2^n of the power grid with n lines. While it might be feasible to somehow eliminate the most unlikely of these states, it would seem that there would remain many likely states to be considered, and this is unwieldy for power systems of practical size for cascading analysis (n of order hundreds or more).[2] This problem of the scaling of the number of states is probably the most important motivation for translating the DAG to an influence graph, which has $n + 1$ states (the n lines together with the empty set

[2]Note that small systems cannot accommodate more than a few generations of outages without hitting the edge of the system. If the small system is actually part of a larger interconnection, this means that the edge effects dominate and the cascading effects across the interconnection (which are the cascades of highest risk) cannot be addressed in the small system.

indicating stopping). The DAG states (nodes) record all the lines that have previously outaged, whereas the influence graph states record only the last outaged line.

5.2.3 Method of translating DAG into an influence graph

The weight of an edge in DAG is the probability that a line fails given all the outaged lines that have outaged. The weight of an edge in the influence graph is the probability that a line fails given the line that outaged in the current generation.[3]

Lines are numbered 1,2,3,... and the generations are indexed 0,1,2,... Write 12 for the event lines $\{1, 2\}$ outaged in generation 1. Write $12 \to 123$ for $\{1, 2\}$ in generation 1 transitioning to $\{1, 2, 3\}$ in generation 2. This type of transition is in DAG.

Write $1_1 \to 2_2$ for the outage of line 1 in the transition to generation 1 transitioning to the outage of line 2 in generation 2. This type of transition is in the influence graph from generation 1 to generation 2.

Given a state xy with $x, y \neq 2$, the probability of it having 2 in the next transition is $P[xy \to xy2]$. Given a state $x1$ the probability of it having 1 in the preceding transition is $\frac{P[x]P[x\to 1]}{P[1]P[1\to x]+P[x]P[x\to 1]}$. These two transitions are independent by assumption. So given a specific state $x1$ with $x \neq 2$, the probability of it having 1 in the preceding transition and 2 in the next transition is

$$P[1_1 \to 2_2 | \text{via } x1] = \frac{P[x]P[x \to 1]}{P[1]P[1 \to x] + P[x]P[x \to 1]} P[x1 \to x12] \tag{5.7}$$

and since the passages through the various $x1$ are disjoint:

$$P[1_1 \to 2_2] = \sum_{x \neq 1} \frac{P[x]P[x \to 1]}{P[1]P[1 \to x] + P[x]P[x \to 1]} P[x1 \to x12] \tag{5.8}$$

More generally, we want to calculate the probability of line i in the preceding transition and line j in the following transition when we are in generation k. First suppose this happens via state $\underline{x}i = x_1x_2...x_ki$. Given the state $\underline{x}i = x_1x_2...x_ki$ with $j \notin \underline{x}$, the probability of

[3] As DAG only considers single-line outages, the states in the influence graph are all single-line outages.

$\underline{x}i = x_1x_2...x_ki$ having i in the preceding transition is

$$\frac{P[\underline{x}]P[\underline{x} \to \underline{x}i]}{\sum_{r=1}^{k+1} P[(\underline{x}i)^{(-r)}]P[(\underline{x}i)^{(-r)} \to \underline{x}i]} \tag{5.9}$$

where $(y_1, y_2, ..., y_{k+1})^{(-r)} = (y_1, y_2, ..., y_{r-1}, y_{r+1}, y_{r+2}, ..., y_{k+1},)$; that is, $^{(-r)}$ removes the rth element. The probability of $\underline{x}i = x_1x_2...x_ki$ with $j \notin \underline{x}$ having i in the preceding transition and j in the following transition is

$$\frac{P[\underline{x}]P[\underline{x} \to \underline{x}i]}{\sum_{r=1}^{k+1} P[(\underline{x}i)^{(-r)}]P[(\underline{x}i)^{(-r)} \to \underline{x}i]} P[\underline{x}i \to \underline{x}ij] \tag{5.10}$$

Then the probability of generation k having i in the preceding transition and j in the following transition is

$$P[i_{k-1} \to j_k] = \sum_{\underline{x}; i, j \notin \underline{x}} \frac{P[\underline{x}]P[\underline{x} \to \underline{x}i]}{\sum_{r=1}^{k+1} P[(\underline{x}i)^{(-r)}]P[(\underline{x}i)^{(-r)} \to \underline{x}i]} P[\underline{x}i \to \underline{x}ij] \tag{5.11}$$

First, the absolute probabilities $P[\underline{x}]$ can be computed recursively using

$$P[\underline{x}] = \sum_{r \in \underline{x}} P[\underline{x}^{(-r)}]P[\underline{x}^{(-r)} \to \underline{x}] \tag{5.12}$$

The transition probabilities for the first generation in the influence graph are the same as the failure probabilities for the first generation in DAG. So we go through the DAG and apply (5.11) to compute the transition probabilities for the influence graph.

DAG models the failure time for a line as an exponential distribution. The failure probability in DAG is the cumulative density function (CDF). We need to assign the simulation time interval to calculate the failure probability from the exponential distribution of the failure time. The longer the simulation time interval, the larger the failure probability. As the simulation time interval is assigned freely, the corresponding transition probability in the influence graph varies. Thus, the row sum in the transition matrix in the influence graph is not always one. Therefore, we assume that (1) each line outage possibly propagates to any other line outage with a small probability; (2) a cascade either propagates further or stops at a generation. Thus we adjust the transition matrix of the influence graph by (1) adding a small number to all the transition probabilities; and (2) adjusting the stopping probabilities to make the total of each row be one.

5.2.4 Case study

We use the method in Section 5.2.3 to translate DAG into an influence graph and verify the influence graph by comparing the cascade size distribution computed from the influence graph with the distribution of cascade size generated by DAG. The IEEE 118-bus system is used as a test system.

5.2.4.1 An overview of 118-bus system

The IEEE 118-bus system represents a portion of the power system in Midwestern of U.S. as of December 1962. This test model contains 186 branches (177 lines and 9 transformers), 19 generators, 35 synchronous condensers. Figure 5.11 shows the single line diagram of IEEE 118-bus system.

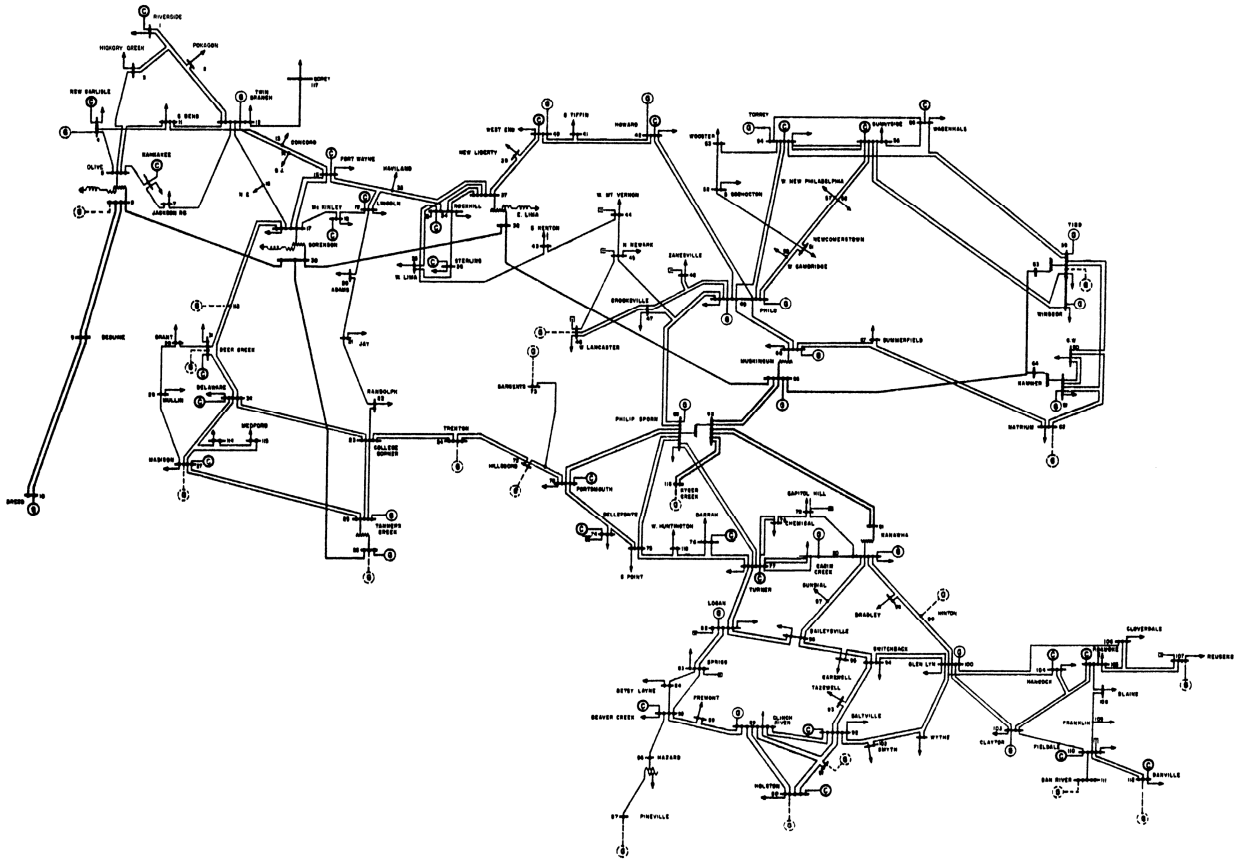


Figure 5.11: Single line diagram of the IEEE 118-bus system

5.2.4.2 DAG and cascade data

A full DAG of the IEEE 118-bus system has 2^{186} vertices theoretically, which is practically impossible to be formed. Therefore, fifty cascading sequences are simulated to prune the full DAG. Cascading sequences are divided into cascades according to the timing of outages. The outage time difference in a cascade is less than 1 hour. In other words, if the time difference between two outages are greater than 1 hour, these two outages belong to different cascades. Based on this principle, 50 sequences are divided into 109 cascades. They have a distribution of cascade size in terms of number of line outages shown in Table 5.2.

Table 5.2: Cascade sizes of cascade data

size	3	4	5	6	7	8	9	10	11	12	13
count	95	4	2	1	1	3	1	0	0	1	0
size	14	15	16	17	18	19	20	21	22	23	24
count	0	0	0	0	0	0	0	0	0	1	0

The initial outages are two-line outages. They are sampled according to the distance distribution between lines in multiple-line outages. Specifically, the first line is drawn randomly from all lines; then, a distance d is sampled according to the distance distribution; finally, a second line is drawn randomly from all lines that are distance d from the first line.

This DAG has 20,395 vertices and 40,759 edges. 152 out of 186 lines have outaged at least once. Let the top layer of DAG be layer 0, the second layer be layer 1, etc. Initial outages are 2-line outages, so that layer 0 are vertices with two line outages. If there are m_k vertices in layer k and m_{k+1} vertices in layer $k+1$, then theoretically there are $m_k m_{k+1}$ edges between layer k and layer $k+1$.

5.2.4.3 Translate DAG into an influence graph

The influence graph is denoted by $IG(\pi_0, P_0, P_1, P_2, ...)$, where π_0 is the state distribution for initial outages, P_0 is the matrix for transition from generation 0 to generation 1, and P_i is the transition matrix from generation k to $k+1$. Figure 5.13(b) illustrate the influence graph

structure. P_0 has a special structure because initial outages are two-line outages and dependent outages are single-line outages. P_0 has dimension $s_0 \times n_{line}+1$ (there are n_{line} lines and one stop state). $P_k(k>0)$ is derived based on P_+, which is the average of the transition matrices $\hat{P}_k$ derived directly from layer k to layer $k+1$ by method in Section 5.2.3. The reason that $\hat{P}_k$s are lumped together is that they are sparse. The rest of this section describes the detail of forming the influence graph.

Layer 0 in DAG includes size-2 vertices, and layer 1 is a set of size-3 vertices. We assume vertices in layer 0 are uniformly distributed. That is, the probability of any vertex in layer 0 is $1/s_0$, where $s_0 = 49$ is the number of vertices in layer 0. We form a transition matrix P_0 from layer 0 to layer 1, which is a s_0 by $n_{line}+1$ matrix. 1 is the number of the stop state. $P_0[i,j](j>1)$ is the transition probability that line $j+1$ ($j=1$ is the stop state) is outaged given 2 lines outaged that are represented by i-th vertex in layer 0.

We form a transition matrix $\hat{P}_k$ from layer k to layer $k+1$ ($k>0$). $\hat{P}_k$ is an n by n matrix. $n = n_{line}+1$, in which 1 corresponds to the stop state. $P_k[1,1]=1$ because if a cascade stops, it stays in stop state. $P_k[i,1](i>1)$ is the transition probability that a cascade stops given line $i-1$ outaged in previous generation. $P_k[i,j](i,j>1)$ is the transition probability that line $j-1$ outaged given line $i-1$ outaged in previous generation. $\hat{P}_k$ is a sparse matrix. If a row in $\hat{P}_k$ are all zeros, it means we do not observe the corresponding state transitioning to other states. At least 110 states out of 187 do not have interactions with other states from layer 0 to layer 12. That is why we lump $\hat{P}_k$ together and form P_+ . That is

$$P_+ = \frac{1}{12}\sum_{k=1}^{12}\hat{P}_k \tag{5.13}$$

Besides layer 0 to layer 12, there are layer 19 to layer 25. We do not consider layer 19 to layer 25 because there is only one cascade that has 23 line outages.

Introducing P_+ can mitigate this problem to some extent, but P_+ still has 108 rows that are all zeros. Therefore, it is necessary to adjust the stopping probability (first column of P_+). This adjustment is done by assuming stopping counts follow binomial distribution with Beta prior. For the details, refer to Section VI-A of [39].

We cannot use P_+ because it implies a constant propagation rate. However, the propagation rate is increasing as the cascade propagates. The average propagation ρ_k for generation k is estimated from the data using

$$\begin{aligned}\hat{\rho}_k &= \frac{\text{Number of cascades with } > k+1 \text{ generations}}{\text{Number of cascades with } > k \text{ generations}} \\ &= \frac{\pi_{k+1}(\mathbf{1} - \boldsymbol{e}_0)}{\pi_k(\mathbf{1} - \boldsymbol{e}_0)}\end{aligned} \tag{5.14}$$

where π_k is the state distribution for generation k, $\mathbf{1}$ is a column vector whose all elements are 1, and e_0 is a column vector with the first element 1 and the rest element 0. An important feature of the cascading data is that average propagation ρ_k increases with generation k as shown in Table 5.3. There are only several cascades with 5 or more generations, which makes the

Table 5.3: Propagations of generations $k = 1$ to 5

k	1	2	3	4	5
$\hat{\rho}_k$	0.11	0.67	0.75	0.83	0.80

estimation of $\hat{\rho}$ less unreliable for 5 or more generations. So we use the $\hat{\rho}_5$ as an estimation for $k > 5$, and we write $\hat{\rho}_{5+}$.

P_+ is adjusted according to $\hat{\rho}$ by a matrix A_k that has structure as in (5.15). Specifically, $P_1 = P_+A_1$, $P_2 = P_+A_2$, $P_3 = P_+A_3$, $P_4 = P_+A_4$, $P_{5+} = P_+A_5$. Values of A_k are determined by solving (5.16). π_k is the state distribution for generation k, which is $\pi_k = \pi_{k-1}P_{k-1}$. We obtain π_k and P_k in an alternating fashion. That is, we obtain them in order of $\pi_1, P_1, \pi_2, P_2, ...$

$$A_k = \begin{pmatrix} 1 & 0 & ... & 0 \\ a_k & 1-a_k & ... & 0 \\ \vdots & & \ddots & \\ a_k & 0 & ... & 1-a_k \end{pmatrix} \tag{5.15}$$

$$\hat{\rho}_k = \frac{\pi_k P_+ A_k(\mathbf{1} - e_0)}{\pi_k(\mathbf{1} - e_0)} = (1 - a_k)\frac{1 - \pi_k P_+ \boldsymbol{e}_0}{1 - \pi_k \boldsymbol{e}_0} \tag{5.16}$$

5.2.4.4 Comparison of DAG and the influence graph

The DAG describes the probability that a line outages given all previous outaged lines, while the influence graph describes the probability that a line outages given outaged lines in previous generation by assuming the Markov property. Figure 5.13 shows an illustration of the two graph models. DAG has 20395 states in all layers, while the influence graph only has 236 = 49 + 187 states. As more cascades are explored, DAG may include more states, while the influence graph does not add new states but estimation of transition probabilities is more reliable.

The influence graph is used to compute the cascade size distribution in terms of generations. Let $S(k)$ be the probability that a cascade has more than k generations, which is the survival function of number of generations. Then, $S(k) = 1 - \pi_k[0]$ because $\pi_k[0]$ is the probability that a cascade propagates to the next generation. More generally,

$$\begin{aligned} S(k) &= 1 - \pi_k[0] = \boldsymbol{\pi}_k(\mathbf{1} - \boldsymbol{e}_0) \\ &= \boldsymbol{\pi}_0 \boldsymbol{P}_0 \boldsymbol{P}_1 ... \boldsymbol{P}_{k-2} \boldsymbol{P}_{k-1} (\mathbf{1} - \boldsymbol{e}_0), \end{aligned} \tag{5.17}$$

For this specific influence graph, the number of line outages equals to the number of generations plus one because only initial outages are two-line outages and all the subsequent generations are one-line outages.

The DAG is associated with a data set of cascades. To verify that the influence graph produces the same statistics of cascades, we compare the cascade size distribution estimated from cascade data and computed based on the influence graph. Figure 5.12 shows the survival function of the cascade size. $S(1) = S(2) = 1$, where S is the cascade size because all cascades have at least 2 line outages. It also shows that the cascade size distribution computed from the influence graph matches the empirical cascade size distribution estimated from cascade data. However, this result may have a high variance. That is because cascade data as shown in Table 5.2 are limited, and the transition matrix of the influence graph is sparse.

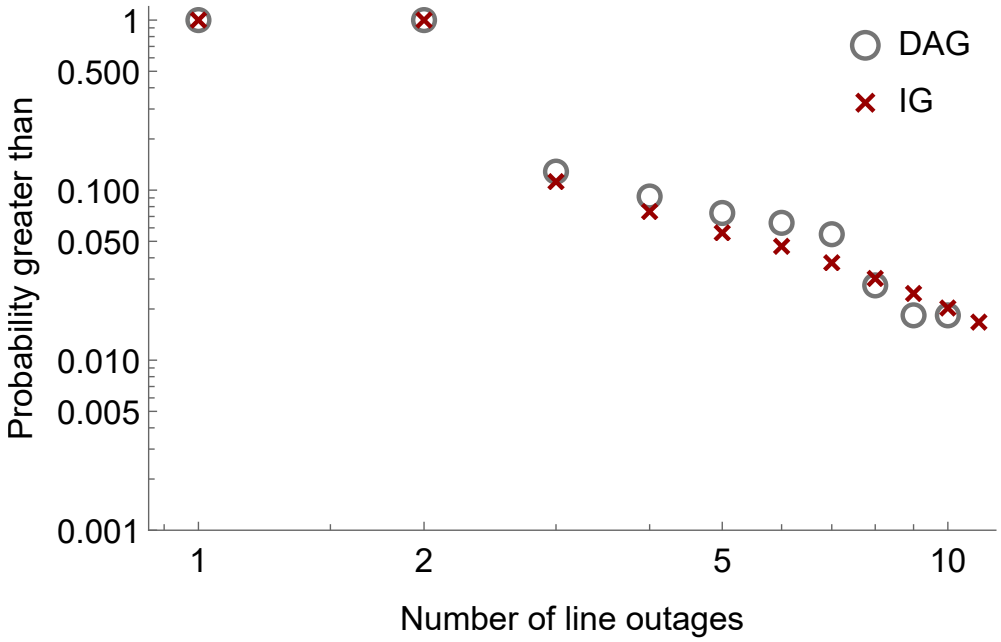


Figure 5.12: The distribution of cascade size calculated using the influence graph (red cross) and estimated from cascade data associated with DAG (gray circle).

5.2.5 Conclusion

The influence graph assumes line outages only depend on the preceding line outage. To test this assumption, this chapter compares the influence graph with the DAG, which is also a Markov chain but considers all failed line outages. We make the test on the IEEE 118-bus system and the influence graph produces the same statistics as cascade data associated with DAG.

An advantage of the influence graph is that the size of the influence graph is fixed, while the size of DAG is increasing as more cascades are simulated. The number of states of the influence graph is the number of transmission lines in the system. The DAG would have more vertices and edges as more cascades are simulated, as failure rate of a line is conditional on all previously outaged lines. The DAG has much more detailed and numerous Markov states, but the cost is that the size of DAG is huge.

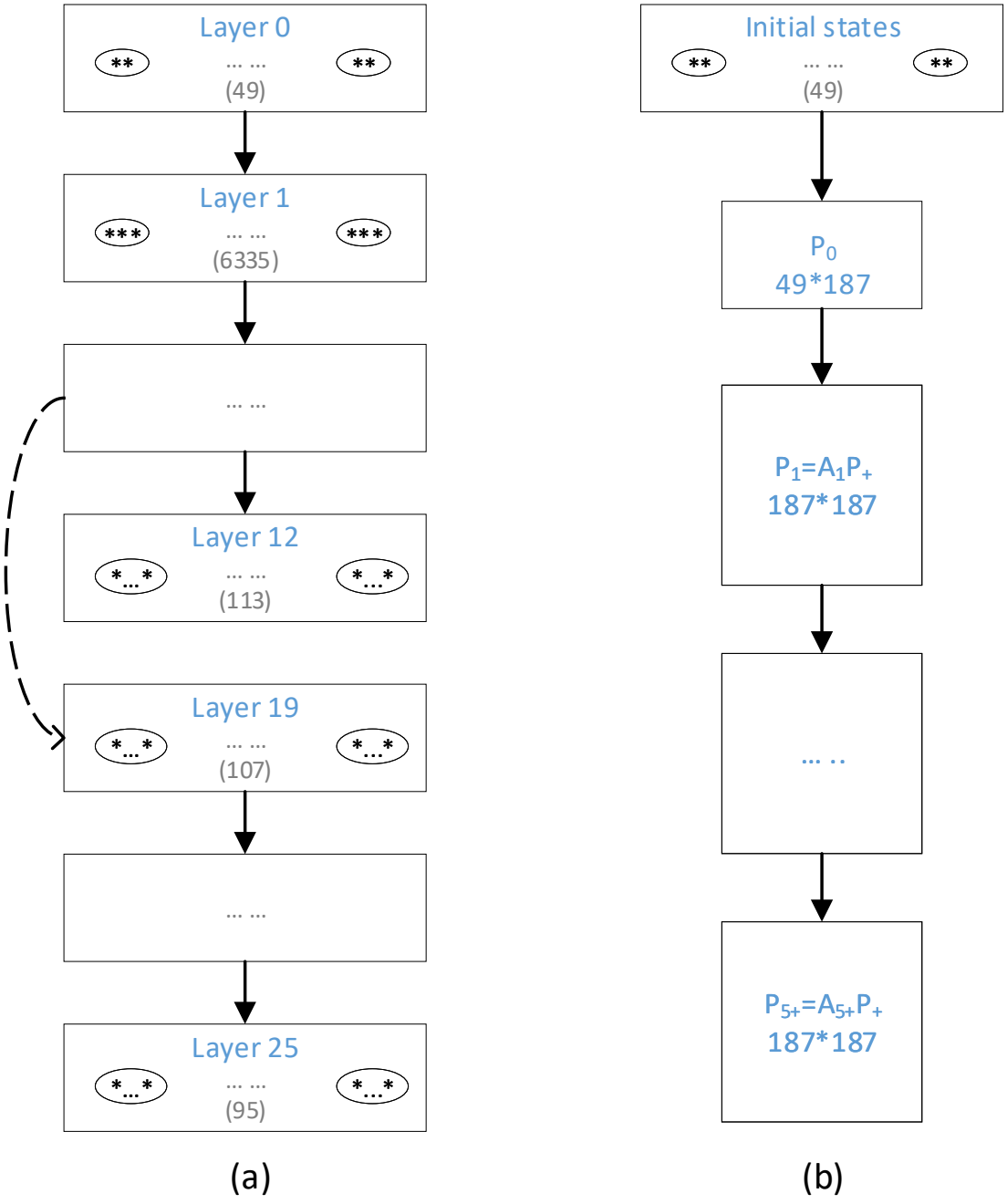


Figure 5.13: An illustration of DAG without stop state (a) and influence graph (b). In (a) , the number in parentheses is the number of vertices in this layer, a star represents a line outage, a solid line is a group of edges, and a dashed line is an edge from layer 9 to layer 19. In (b), 49 * 187 is the dimension of P_0.

CHAPTER 6. BAYESIAN ESTIMATES OF TRANSMISSION LINE OUTAGE RATES

Transmission line outage rates are foundation for many reliability calculations. However, line outages are infrequent, so outage data is limited. This chapter proposes a Bayesian hierarchical model that leverages line dependencies to better estimate outage rates of individual transmission lines. The Bayesian hierarchical model produces more accurate estimates of individual line outage rates and the uncertainty of these estimates.

This chapter's work is done in collaboration with L. Wehenkel, University of Liège, Belgium, J. R. Cruise, Riverlane Research, UK, C. J. Dent and A. Wilson, University of Edinburgh, Scotland. L. Wehenkel and J. R. Cruise particularly helped with the formulation of the problem. The material in this chapter is published in [91].

6.1 Introduction

Transmission lines are partially similar in several ways, such as their length, rating, geographical location, and their proximity. We leverage these partial similarities with a Bayesian hierarchical model to improve the estimation of line outage rates from historical data.

The conventional method of estimating annual line outage rates divides the number of outages by the number of years of data. However, these estimates have a high variance when the data are insufficient. Indeed, many lines either do not fail or only fail once in a year.

One pragmatic approach to mitigate the problem of limited outage counts is to group or pool similar lines together to get an estimate for the outage rate of that group. The lines can be grouped by areas [48, 51, 54], or by line voltage ratings. Lines in the same area experience similar weather conditions, and lines of the same rating have similar construction. However, the

similarity between lines in these groups is only partial, variations of outage rates within the groups can be neglected, and it is unwieldy to group lines according to multiple characteristics.

Transmission line outage rates are often supposed to be proportional to line length, and they are often quoted as rates per unit length [49, 50]. However, a line's outage rate is not strictly proportional to the line length because of substation and other effects, making the dependence on line length only a partial dependence. Indeed, our historical line outage data show only a limited dependence on line length.

There is a middle ground between pooling lines in groups assuming perfect line dependencies within the group, and completely neglecting dependencies between lines by computing individual line outage rates in isolation. To exploit the partial dependencies of transmission line outage rates, this chapter proposes a Bayesian hierarchical model to estimate outage rates of individual transmission lines. In particular, our method can leverage the multiple partial dependencies in line length, rating, network proximity, and geographical area to give better outage rates of individual lines. This is done by explicitly modeling the dependence of outage rates on line length and rating and by using covariance kernels to model the dependencies between lines in close proximity. Our method can, therefore, learn about the outage rates of individual lines from lines close-by and with similar lengths and ratings. This means that where there is little data associated with a line (because the outage rate is small), our method can still estimate an outage rate for that line and its uncertainty. Also, by borrowing information from other lines, we can expect smaller uncertainties associated with estimates of outage rates, without assuming that all lines within a group have the same outage rate (as would be the case if we pooled the data).

6.2 Exploring historical outage data and modeling line dependencies

Utilities routinely collect detailed outage data. For example, NERC's Transmission Availability Data System (TADS) collects outage data from North American utilities. Here, to illustrate our method, we use a publicly available historical line outage data [7].

Table 6.1: Annual Outage Counts, Line attributes, and Bayesian estimates of outage rates after 1st, 7th and 14th years for 4 lines

Line	Outage counts in different years														Line attributes			Annual outage rate		
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Voltage(kV)	Length(mile)	District	1st	after 7th	after 14th
29	0	0	0	0	0	0	0	0	3	2	0	0	0	0	230	8.3	P	0.32	0.17	0.37
11	0	0	1	0	0	1	0	0	1	0	0	0	0	1	500	22.65	N	0.36	0.33	0.34
2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	230	7.62	A	0.73	0.48	0.28
8	1	2	4	2	1	2	2	2	2	1	3	8	6	2	500	148.86	E	0.93	1.85	2.56

6.2.1 Historical outage data

The historical line outage data we use consists of transmission line outages recorded by a North American utility [7] for 14 years since 1999. The data record forced and scheduled line outages, including the sending and receiving bus names of outaged lines, outage start and end times and dates, line attributes such as lengths, voltage ratings, districts in which a line is, and outage causes. Some lines cross several districts. There are 549 lines outaged in the data with rated voltages of 69, 115, 230, 287, 345, and 500 kV.

We neglect the scheduled outages and only consider the forced line outages. We also exclude two 1000 kV HVDC lines, and momentary outages (outage duration does not exceed one minute). There are lines that failed once or twice in most of years but suddenly failed, for example, ten times in one year. One common reason that a line could fail several times in a day is outages and reclosures for the same cause. So if a line fails several times in a day, we only count it once. Table 6.1 shows an example of the outage data.

6.2.2 Data exploration

We initially explore the line outage data using the conventional method of estimating annual line outage rates by dividing the number of outages by the number of years. We first pool all the line data together (i.e. treat as one homogeneous data set) to calculate the overall mean and standard deviation of outage rates, which are 0.6 and 0.7 outages per year, respectively. Next, we

examine the individual conventional line outage rates. The mean variance-to-mean ratio of outage counts for each line is 1.2, which indicates that the outage counts show some overdispersion[1].

The power system network can be deduced directly from the outage data using the method in [1], and we show the conventional outage rates on the network in Figure 6.1 to visualize the spatial correlation. Close lines tend to have close colors, which indicates line dependencies from network proximity.

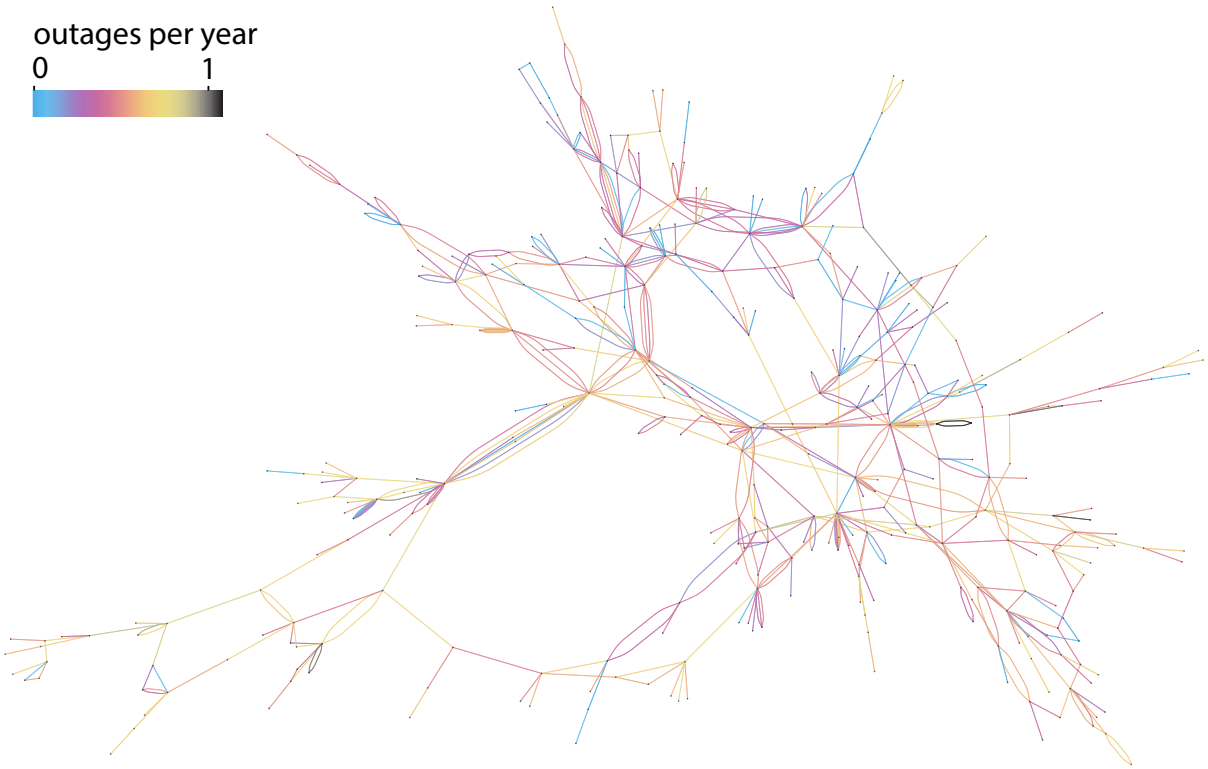


Figure 6.1: The number of average annual forced outages over 14 years on network indicated by different colors (network layout is not geographic).

[1]Overdispersion means that the variance is larger than the mean. The Poisson distribution commonly used for count data does not apply when there is overdispersion because the Poisson mean and variance are equal.

6.2.3 Scaling line lengths and voltage ratings

The line lengths and voltage ratings are transformed and scaled so that their magnitudes and variations are scale-free and comparable. We do this in ways suggested by Gelman [92] for generic priors.

The line lengths in the vector $\boldsymbol{L}$ are first transformed by the natural logarithm to make the range of values less extreme, and then divided by the scale so that their variations are order of magnitude one:

$$\boldsymbol{x}_L = \frac{\ln \boldsymbol{L}}{\text{scale}(\ln \boldsymbol{L})} \tag{6.1}$$

Here the scale of the sample in a vector $\boldsymbol{z}$ is estimated by the Mean Absolute Deviation, which is $\text{scale}(\boldsymbol{z}) = \text{median}(\boldsymbol{z} - \text{median}(\boldsymbol{z}))$. Note that we use bold variables for vectors, and functions such as ln are applied element-wise so that $\ln \boldsymbol{L} = [\ln L_1, ..., \ln L_N]'$.

Similarly, the line voltage ratings $\boldsymbol{V}$ are first scaled by $\text{SD}(\boldsymbol{V})$, the standard deviation of $\boldsymbol{V}$, and then divided by the scale:

$$\boldsymbol{x}_V = \frac{\boldsymbol{V}/\text{SD}(\boldsymbol{V})}{\text{scale}(\boldsymbol{V}/\text{SD}(\boldsymbol{V}))} \tag{6.2}$$

It is usually considered that the line length and voltage rating have a positive correlation. Indeed, the BPA data show this correlation, but it is a weak correlation: the Pearson correlation coefficient is 0.34 (0.12 for transformed lengths and voltage ratings).

6.2.4 Line proximity

The proximity of lines is quantified by the weighted sum of two kernels, which reflect two aspects of proximity. The first kernel is based on districts. Lines in the same district are more likely to experience the same weather conditions. Another kernel is based on network distance in terms of line length, which, to some extent, reflects both geographic proximity and the physical and engineering interactions in the power grid. We fit a linear regression model with correlated lines (described below) to support the form of the Bayesian hierarchical model and give guidance on setting priors.

6.2.4.1 Districts

There are twelve districts (Figure 6.2). The mean number of automatic outages per line per year in each district are shown in Table 6.2. Lines with two or more districts are excluded from this computation. The difference in rates makes it plausible that district information can inform the rates.

Table 6.2: The mean number of automatic outages per line per year in each district

COV	EUG	KAL	LGV	OLY	RED	SAL	SNO	SPK	TDA	TRI	WEN
3.84	8.33	18.81	7.05	9.08	12.96	10.11	5.55	7.65	12.89	6.82	9.22

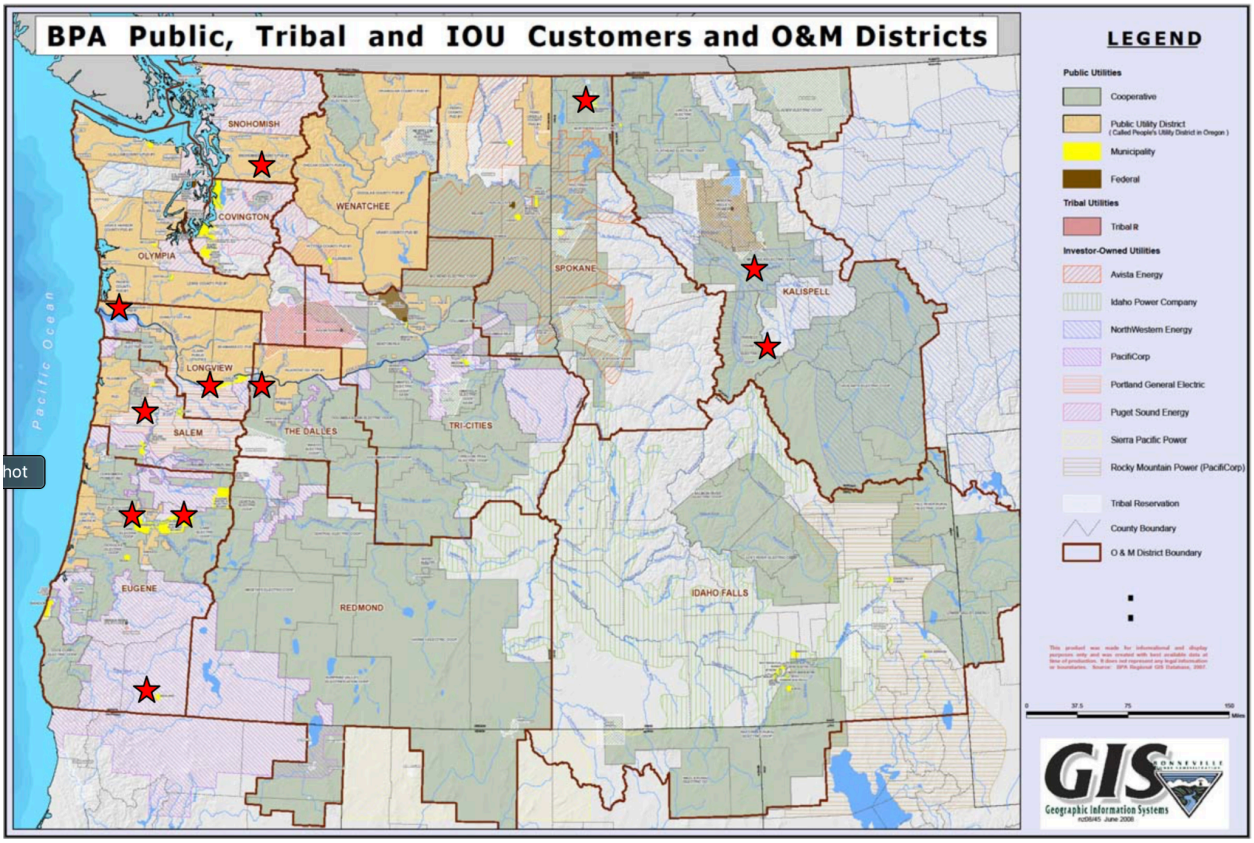


Figure 6.2: BPA districts

The districts for each line are represented by a feature vector $\phi_{dis} \in \{0,1\}^{12}$ whose coordinates correspond to the districts and are set to 1 for the district crossed by that line, and to 0 otherwise. The scalar product in this feature space thus counts the number of common districts crossed by two lines.

We define the district kernel as:

$$\mathbf{\Sigma}_1 = \exp\left(-||\phi_{dis}(i) - \phi_{dis}(j)||_2^2 - \mathbb{I}_{i \neq j}\right) \tag{6.3}$$

where $||\cdot||_2$ stands for the two-norm[2], and $\mathbb{I}_{i \neq j}$ is the indicator function. The reason why $\mathbb{I}_{i \neq j}$ is included is that a line is most similar to itself. The kernel $\mathbf{\Sigma}_1$ has the form of a correlation matrix since it is positive definite.

6.2.4.2 Network distance

The network distance between lines L_i and L_j along the network lines is defined as

$$d_{ij} = d(L_i, L_j) = \text{minimum length in miles of a network path joining midpoint of } L_i \text{ to midpoint of } L_j.$$

For example, the distance of a line to itself is zero and the distance of a line to a neighboring line with at least one bus in common is half of the total length of the two lines.

Then we use the exponential kernel $\mathbf{\Sigma}_2$ which is

$$\mathbf{\Sigma}_2 = \exp[-2d(L_i, L_j)] \tag{6.4}$$

As $d(L_i, L_i) = 0$, the diagonal elements of $\mathbf{\Sigma}_2$ are one.

6.2.4.3 Combining the two kernels

The network proximity matrix $\mathbf{\Sigma}$ is the weighted sum of above two kernels:

$$\mathbf{\Sigma} = w\mathbf{\Sigma}_1 + (1-w)\mathbf{\Sigma}_2, \tag{6.5}$$

[2]Since the vectors only have entries 0 or ± 1, the one-norm is the same as the two-norm in this context.

where $0 < w < 1$. For example, if the two kernels are equally important, then $w = 0.5$.

We find the weight by fitting a linear regression model for the logarithm of average outage counts with $\boldsymbol{\beta}_0$ following a multivariate normal distribution to model the correlation:

$$\ln \frac{\boldsymbol{N}}{t} = \boldsymbol{\beta}_0 + \beta_L \boldsymbol{x}_L + \beta_V \boldsymbol{x}_V, \tag{6.6}$$

$$\boldsymbol{\beta}_0 \sim \mathcal{N}(m\mathbf{1}, \sigma^2 \boldsymbol{\Sigma}), \tag{6.7}$$

where $\boldsymbol{N}$ is a column vector whose entry N_i is the total number of counts in t years for line i, $\mathbf{1}$ is a column vector of ones, m, β_L, β_V are scalars, and

$$\sigma^2 \boldsymbol{\Sigma} = \sigma^2(w\boldsymbol{\Sigma}_1 + (1-w)\boldsymbol{\Sigma}_2) = \sigma_1^2 \boldsymbol{\Sigma}_1 + \sigma_2^2 \boldsymbol{\Sigma}_2. \tag{6.8}$$

For computation convenience, we decouple the dependencies between different lines in (6.8) by a coordinate transformation to diagonalize the covariance matrix $\sigma^2\boldsymbol{\Sigma}$. This transforms the multivariate normal random vector $\boldsymbol{\beta}_0$ in (6.7) into independent univariate normal random variables in the vector $\boldsymbol{\beta}_0'$. This decoupling facilitates the maximum likelihood calculation below. In particular, by simultaneous diagonalization [93, p.286], we find a matrix $\boldsymbol{Q}$ such that $\boldsymbol{Q}^T\boldsymbol{\Sigma}_1\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{Q}^T\boldsymbol{\Sigma}_2\boldsymbol{Q} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix. Define $\boldsymbol{\beta}_0' = \boldsymbol{Q}^T\boldsymbol{\beta}_0$, then

$$\begin{aligned} \boldsymbol{\beta}_0' &\sim \mathcal{N}(m\boldsymbol{Q}^T\mathbf{1}, \boldsymbol{Q}^T(\sigma_1^2\boldsymbol{\Sigma}_1 + \sigma_2^2\boldsymbol{\Sigma}_2)\boldsymbol{Q}) \\ &\sim \mathcal{N}(m\boldsymbol{Q}^T\mathbf{1}, \sigma_1^2\boldsymbol{I} + \sigma_2^2\boldsymbol{\Lambda}). \end{aligned} \tag{6.9}$$

We use Maximum Likelihood Estimation to estimate the parameters $\sigma_1^2, \sigma_2^2, m, \beta_L, \beta_V$ from the utility data. The log likelihood logL is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Q}^T(\ln \frac{\boldsymbol{N}}{t} - \beta_L \boldsymbol{x}_L - \beta_V \boldsymbol{x}_V) \\ \text{logL} &= \sum_i \ln f(y_i | m(\boldsymbol{Q}^T\mathbf{1})_i, \sigma_1^2 + \sigma_2^2\Lambda_i) \end{aligned} \tag{6.10}$$

where $\boldsymbol{y}$ is a column vector with ith entry y_i, $f(\cdot|\mu, \sigma^2)$ is the PDF of a normal distribution with mean μ and variance σ^2, $(\boldsymbol{Q}^T\mathbf{1})_i$ is the ith entry of $\boldsymbol{Q}^T\mathbf{1}$, and Λ_i stands for the ith diagonal entry of $\boldsymbol{\Lambda}$.

The maximum of logL in (6.10) is attained when $\sigma_1^2 = 0.45, \sigma_2^2 = 0.42$, $m = -1.5$ and $(\beta_L, \beta_V) = (0.13, 0.12)$. By normalizing σ_1^2 and σ_2^2, we have $w = \sigma_1^2/(\sigma_1^2 + \sigma_2^2) = 0.52$. The positive values of β_L and β_V indicate that longer lines or higher voltage lines tend to have higher outage rates, which is reasonable. These values shall give guidance on setting priors in Section 6.3.

We check the model assumptions by using the residual plot and QQ-plot as shown in Figure 6.3. $\boldsymbol{\beta}_0'$ has no correlation, so we focus on the transformed linear model, and Pearson residuals are used here as $\boldsymbol{\beta}_0'$ has heterogeneous variance. The Pearson residual is estimated by $\epsilon_i' = \epsilon_i / \sqrt{\sigma_1^2 + \sigma_2^2 \Lambda_i}$, where the raw residuals are $\boldsymbol{\epsilon} = \boldsymbol{Q}^T(\ln \boldsymbol{N}/t - \beta_L \boldsymbol{x}_L - \beta_V \boldsymbol{x}_V - m\mathbf{1})$. There is no noticeable trend in the residual plot, and the QQ-plot shows that the Pearson residual follows the normal distribution.

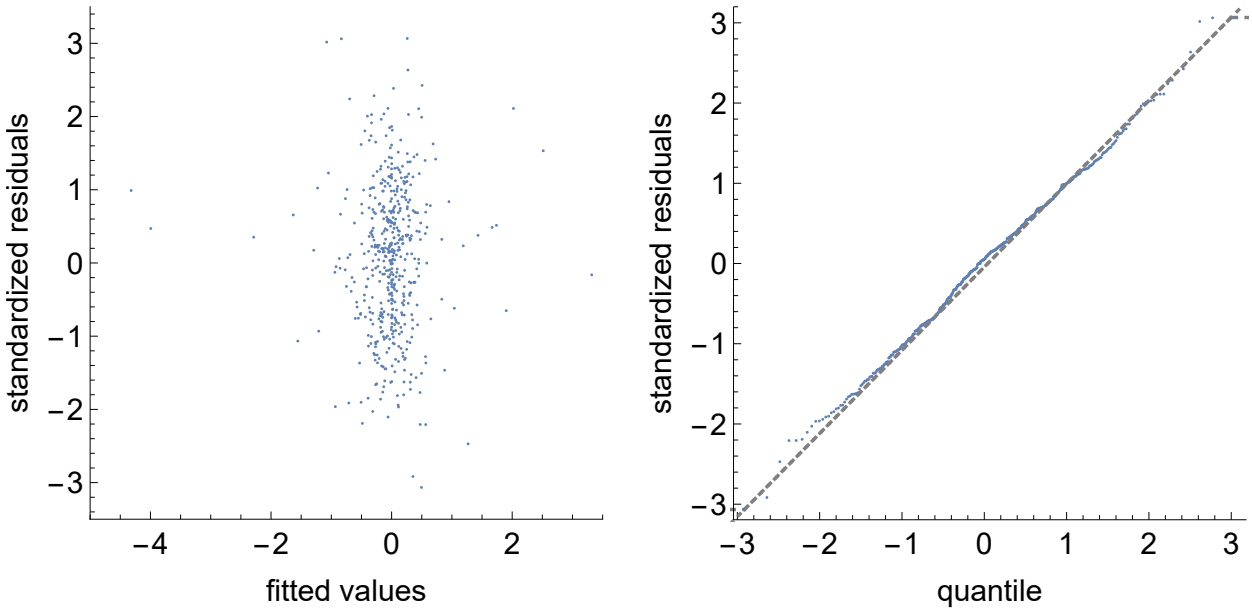


Figure 6.3: Residual plot (left) and QQ-plot (right) for Pearson residuals.

6.3 The Bayesian hierarchical model with line dependencies

We propose a Bayesian hierarchical model of outage counts incorporating line dependencies.

We assume that outage counts follow a Poisson distribution:

$$N_i \sim \text{Poisson}(\lambda_i t_i), \quad i = 1, ..., n \tag{6.11}$$

where N_i is the outage count for line i over t_i years, λ_i is the annual outage rate, and n is the number of lines.

We assume that the outage rates λ_i follow a Gamma distribution:

$$\lambda_i \sim \text{Gamma}(\alpha, \alpha/\mu_i), \quad i = 1, ..., n \tag{6.12}$$

The Gamma distribution is chosen for two reasons. First, it is a conjugate prior for the Poisson distribution. Second, the Gamma distribution mean is μ_i and its variance is μ_i^2/α; the variance of the Gamma distribution increases quadratically as the mean increases, which allows for the overdispersion observed in Section 6.2.2.

The mean outage rate μ_i is modeled via a linear regression model with correlated lines. The linear regression model assumes the predicted variable is normally distributed, but μ_i is positive and may have a large range of values, so μ_i is transformed by a log function [94, Sec. 3.6]:

$$\ln \boldsymbol{\mu} = \boldsymbol{\beta}_0 + \beta_L \boldsymbol{x}_L + \beta_V \boldsymbol{x}_V \tag{6.13}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\beta}_0$ are column vectors.

$\boldsymbol{\beta}_0$ follows a multivariate normal distribution:

$$\boldsymbol{\beta}_0 \sim \mathcal{N}(m\mathbf{1}, \sigma^2(w\boldsymbol{\Sigma}_1 + (1-w)\boldsymbol{\Sigma}_2)) \tag{6.14}$$

the line proximity dependency is captured by the covariance matrix of this multivariate normal distribution, σ^2 is a scalar which controls the magnitude of the covariance and w controls the weights of the two kernels. The parameters α, β_V, β_V, m, σ^2 and w will be estimated using prior distributions in combination with the data as described below.

The prior distributions are:

$$\begin{aligned} \alpha &\sim \text{Half Normal}(0.7, 8^2) & \beta_L &\sim \text{Normal}(0.13, 5^2) \\ m &\sim \text{Normal}(-1.5, 5^2) & \beta_V &\sim \text{Normal}(0.12, 5^2) \\ \sigma^2 &\sim \text{Half Normal}(0, 0.5^2) & w &\sim \text{Beta}(1, 1) \end{aligned} \tag{6.15}$$

These priors are set to ensure that the parameters have a reasonable range and/or mean [3] when compared to our knowledge about the system and the model tested in Section 6.2.2. As there is not much information about the standard deviations about these priors, we make these priors weakly informative. The detail is as follows.

The prior for α is a half-normal distribution with $\alpha > 0$. As discussed in Section 6.2.2, the mean annual outage rate is 0.6, and the standard deviation is 0.7. This suggests the expected value of μ is 0.6, so the expected value of α would be $0.6^2/0.7^2 = 0.7$ (as $\mu_i^2/\alpha = \text{Var}\lambda_i$). The standard deviation of α is $\frac{(0.6+2\times0.7)^2}{0.7^2} - 0.6 \approx 8$ (the numerator is the maximum of μ in a typical range estimated by two times the standard deviation, $\frac{(0.6+2\times0.7)^2}{0.7^2}$ is the maximum of α).

The priors for m, β_L, β_V are normal distributions. The linear regression model in Section 6.2.4 suggests expected values for these parameters. $\boldsymbol{x}_L$ and $\boldsymbol{x}_V$ have range $[-10, 10]$ after scaling using method described in Section 6.2.3, and we observe that the range of $\ln \boldsymbol{N}/t$ is $[-10, 10]$ conservatively. Therefore, we set the standard deviations of m, β_L, β_V to 5 so that 95% of the values lie in $[-10, 10]$ and they vary mostly in the same magnitude, which produces weakly informative priors.

σ^2 functions as a variance. The inverse-gamma prior is usually preferred since it is a conditional conjugate distribution. Gelman [95], however, does not recommend the inverse-gamma prior as the estimation of σ^2 would be sensitive to the parameters of inverse-gamma distribution when σ^2 is near zero. Thus, we let σ^2 have a half-normal prior. Section 6.2.4 shows that σ_1^2, σ_2^2 are about 0.5, so we set the standard deviation of σ^2 to 0.5 to make at least 95% of the values of σ^2 to lie in $[0, 1]$.

We give w a uniform prior as we know that w lies in $[0, 1]$ and the expectation of w is $0.52 \approx 0.5$ from Section 6.2.4.

We now summarize the Bayesian hierarchical model. The Bayesian hierarchical model is specified by (6.11,6.12,6.13,6.14) together with the prior distributions of the parameters (6.15). Note that the partial dependencies between lines are expressed in (6.13,6.14).

[3] By saying that a range or mean is reasonable, we mean that the distribution of the prior has mean or range that is consistent with our prior knowledge, and it does not incorporate any further information.

The model parameters, including the outage rates $\boldsymbol{\lambda}$, are

$$\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\beta}_0, \alpha, \beta_L, \beta_V, m, w) \tag{6.16}$$

The objective is to estimate the posterior distribution of the parameters $p(\boldsymbol{\theta}|\boldsymbol{N})$ that is informed by the line outage counts $\boldsymbol{N}$. By Bayes' theorem, the posterior distribution is

$$p(\boldsymbol{\theta}|\boldsymbol{N}) = \frac{p(\boldsymbol{N}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{N})} \tag{6.17}$$

Because normalization can be applied later, it is sufficient to calculate the unnormalized numerator of (6.17). We can exploit the dependencies in the hierarchical model (12,13,14) to get

$$p(\boldsymbol{N}|\boldsymbol{\theta}) = p(\boldsymbol{N}|\boldsymbol{\lambda}) = \prod_i p(N_i|\lambda_i) \tag{6.18}$$

$$\begin{aligned} p(\boldsymbol{\theta}) = \prod_i p(\lambda_i|\alpha, \mu_i) p(\boldsymbol{\mu}|\boldsymbol{\beta}_0, \beta_L, \beta_V) p(\boldsymbol{\beta}_0|m, w) \\ \times p(\alpha)p(\beta_L)p(\beta_V)p(m)p(w) \end{aligned} \tag{6.19}$$

so that

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{N}) &\propto p(\boldsymbol{N}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \prod_i p(N_i|\lambda_i) \prod_i p(\lambda_i|\alpha, \mu_i) p(\boldsymbol{\mu}|\boldsymbol{\beta}_0, \beta_L, \beta_V) \\ &\quad \times p(\boldsymbol{\beta}_0|m, w)p(\alpha)p(\beta_L)p(\beta_V)p(m)p(w) \end{aligned} \tag{6.20}$$

6.4 Bayesian Processing of real data

The Bayesian hierarchical model described in the previous section is applied to the historical outage data.

6.4.1 Sampling posterior distributions using Stan

The posterior distributions (6.20) of the parameters (6.16) can be evaluated numerically by repeated sampling from the distribution with a Monte Carlo Markov Chain (MCMC) algorithm. MCMC is a class of algorithms for sampling from a probability distribution. We use the software

Stan, which implements MCMC as Hamiltonian Monte Carlo (HMC) [96] with the algorithm adaptively tuned by the No-U-Turn Sampler (NUTS) [97]. Appendix C reproduces the algorithm of HMC with some explanatory comments and gives a detailed guide to the introductory and advanced literature on HMC.

We sample 2000 times, and the first 1000 samples are burn-in. Appendix D discusses technical details of model diagnostics and algorithm convergence. In this section, we focus on the result of the sampling.

6.4.2 Results of Bayesian estimates

We use the posterior mean as the point estimate of a line outage rate because the posterior mean minimizes the Bayes risk in terms of squared error loss. Figure 6.4 shows the point estimates of line outage rates and their 95% credible intervals[4]. The mean outage rate of all lines is 0.74 outages per year, and 82% of lines have rates less than 1 outage per year. There are two lines with very high outage rates. By inspecting the cause codes of these outages, one line outaged mainly because of foreign trouble (which is an external cause outside the power system, such as vehicles striking towers), while the other outaged mainly because of a remedial action scheme.

The values of β_L and β_V reveal the relationship between line lengths, voltage ratings, and outage rates. Figure 6.5 shows the posterior distributions of β_L and β_V and their correlation. The means of β_L and β_V are both 0.1. So the logarithm of the outage rate has a weakly positive correlation with transformed line length and transformed voltage rating. β_L and β_V have a very weak correlation, which is reasonable as $\boldsymbol{x}_L$ and $\boldsymbol{x}_V$ have a very weak correlation.

We use weakly informative priors in the Bayesian model. If we had access to previous studies in the region, or outage rates for other similar regions then these could be used to refine the priors. In this case we would expect the uncertainty in the outage rate estimates to be reduced.

We also test the sensitivity of the Bayesian model to the priors using 14-year data using two different sets of priors. The first case uses somewhat stronger informative priors. We reduce the

[4]The credible interval is described by the multiplicative factor κ within which the outage rate λ_i can vary from the point estimate $\hat{\lambda}_i$ with 95% probability; that is, $P[\hat{\lambda}_i/\kappa \leq \lambda_i \leq \hat{\lambda}_i\kappa] = 95\%$.

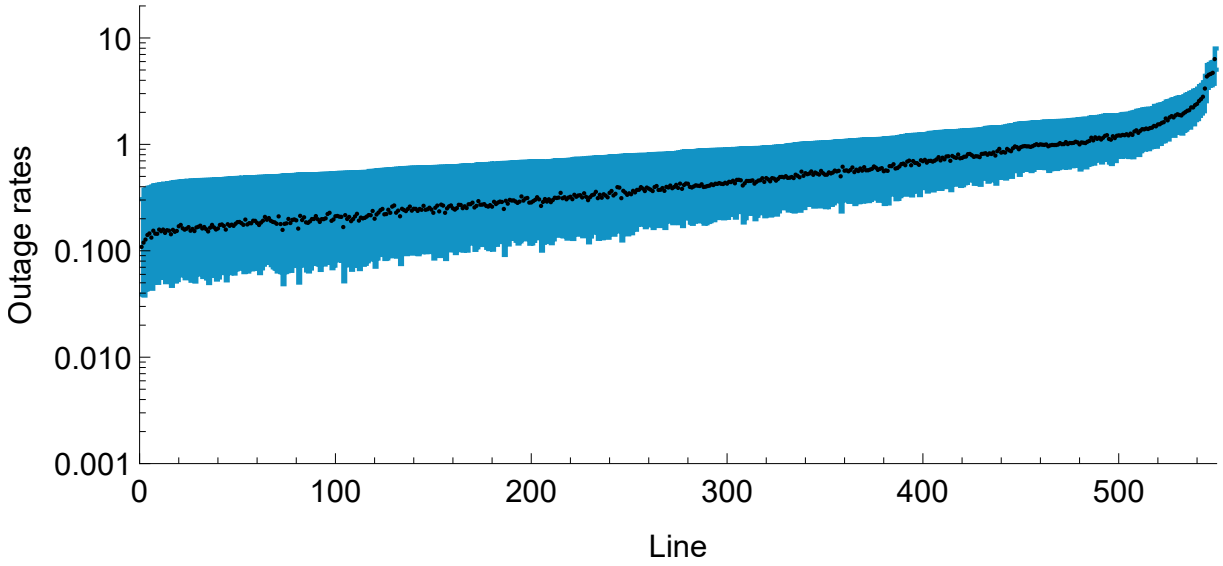


Figure 6.4: Point estimates (black dots) and 95% credible intervals (blue bars) of annual outage rates. Lines are ordered by point estimates.

standard deviation of the prior distributions of m, β_L, β_V from 5 to 1 and redo the calculations. In the second case, we randomly set parameters of priors by sampling from uniform distributions; then, we run the MCMC to estimate the posterior distributions. We compare the posterior mean and standard deviation of outage rates $\boldsymbol{\lambda}$ calculated using different priors, and find there is not much difference.

6.4.3 Comparing the standard deviations of Bayesian and conventional estimates

The Bayesian method produces the distribution of the outage rate, and it is straightforward to compute the standard deviation of this distribution. The conventional method estimates the outage rate with the sample mean. The standard deviation of the sample mean can be estimated as $s/\sqrt{n}$, where s is the sample standard deviation, and n is the sample size.

Figure 6.6 shows the ratio of the standard deviations of the Bayesian and conventional estimators. It shows that the standard deviation of the Bayesian estimator is typically smaller than the conventional estimator, especially when the data is limited to one year. The median ratio of standard deviations is 0.66 for one year of data, while the median ratio is 0.93 for 14 years

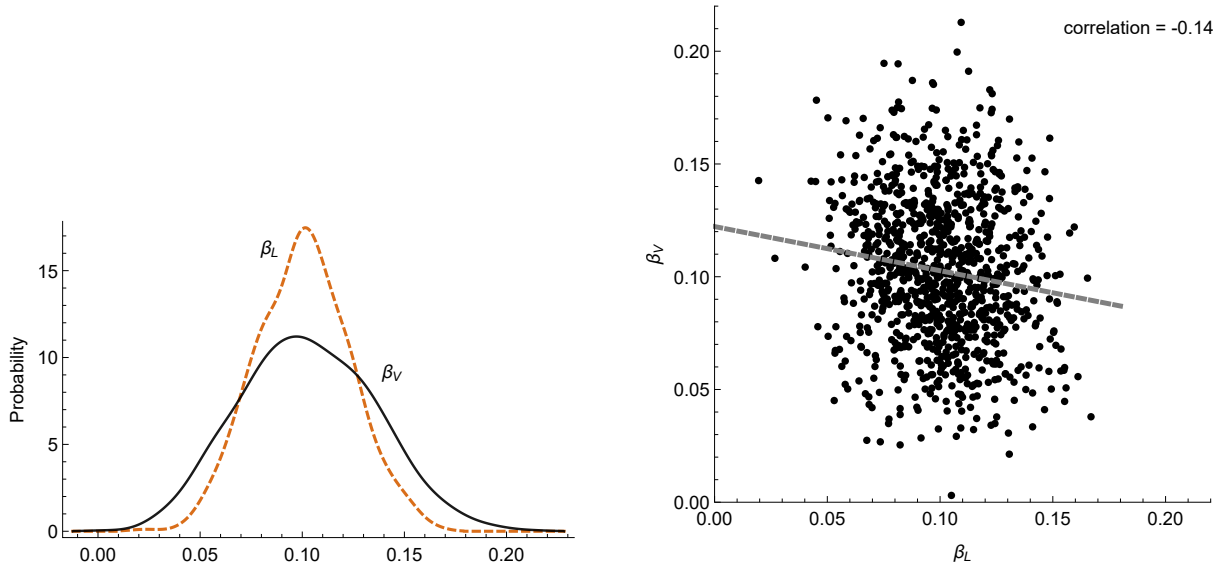


Figure 6.5: Distributions of β_L and β_V (top) and their scatter plot and correlation (bottom).

of data. Thus the Bayesian estimator typically achieves a lower standard deviation than the conventional estimator for limited data. Another way to present this finding is that given the same acceptable precision, the Bayesian method requires fewer data. Since the standard deviation is proportional to the square root of sample size, the Bayes estimator using one year of data achieves the same standard deviation as the conventional estimator using 2.30 years of data ($1/(0.66^2) = 2.30$). Similarly, the Bayesian estimator using 14 years of data achieves the same standard deviation as the conventional estimator using 16.2 years of data ($14/(0.93^2) = 16.2$).

6.4.4 Performance on rarely outaged lines

One advantage of the Bayesian method is that it provides a principled way of making line outage rates with no observed outages. The conventional estimate of outage rate is zero if a line has no outage in a year. However, it is more reasonable that the underlying outage rate of this line is a small value.

Table 6.1 calculates 4 line outage rates with the data available after the 1st year, after the 7th year, and after the 14th year. In Table 6.1, line 29 has no outage except in the 9th and 10th year. The Bayesian estimate of the outage rate of line 29 for the 1st year is 0.32, which is informed by

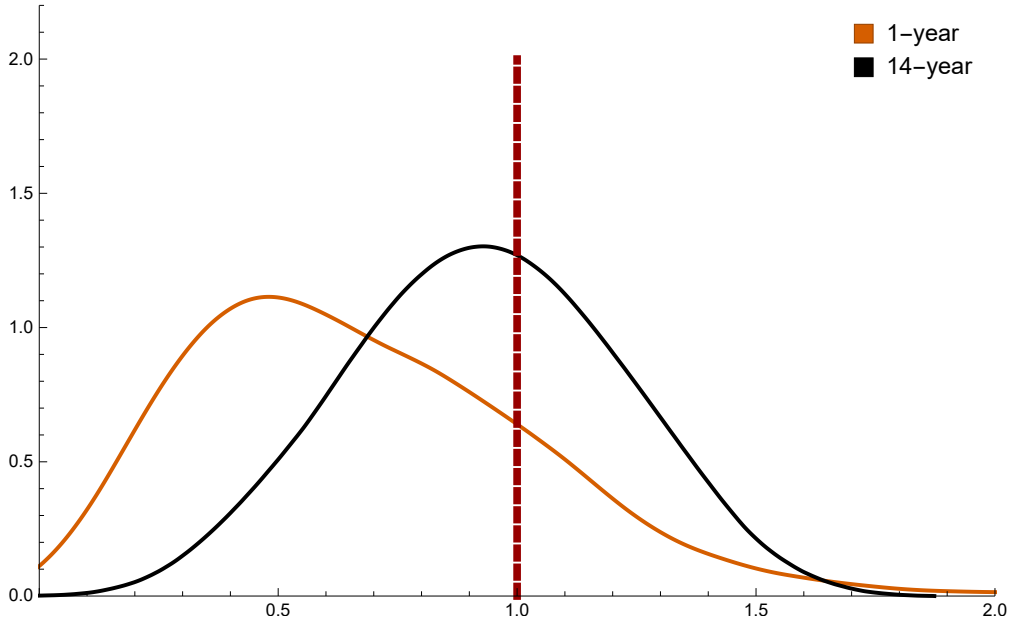


Figure 6.6: Distributions of ratios of standard deviations of Bayesian estimator and conventional estimator using 1-year and 14-year data respectively. The ratio is SD(Bayesian)/SD(conventional).

correlations with other lines. By the 7th year, more years with no outages have been observed, so that the estimated outage rate decreases to 0.17. Line 29 outages several times in the 9th and 10th years, so its estimated rate over 14 years increases. There are also many zeros for lines 11 and 2, but the two outage rates vary differently as the distribution of zeros has different patterns. Most counts for line 11 are zeros, and single outages appear every several years. So we believe that the outage rate is roughly constant and small, which is captured by the Bayesian estimator. At the beginning, line 2 had several outages, and then it stops having outages. So this line has a decreasing outage rate. Line 8 is an example of a line with a high and increasing outage rate.

6.4.5 Validation of the Bayesian hierarchical model

Section 6.2.4 fits a linear regression model to the data, and Figure 6.3 shows that the assumptions for this regression model hold. This validates that the form of the Bayesian hierarchical model (particularly for (6.13), (6.14)) is reasonable.

As we do not know the true outage rates using real data, we generate synthetic data to further validate the Bayesian model in Section 6.5. That is, assuming that the real outage data follow the model detailed in 6.2.4, we test that the Bayesian model accurately estimates the outage rates. As we have checked in Figure 6.3 that the model in 6.2.4 is a good fit to the real outage data, this is a reasonable method for validating the model when we do not have the true outage rates.

6.5 Test Bayesian estimates on synthetic data

We build a generative model for synthetic datasets of arbitrary size, so the data are not limited in size, and the ground truth values are known. Then we test the Bayesian hierarchical model and the conventional estimates on the synthetic data. It turns out that the Bayesian hierarchical model predicts the outage rates well, and the Bayesian estimates compare favorably with the conventional method.

We also construct and test with synthetic data sets a Bayesian hierarchical model without correlations between the lines to evaluate the effect of line dependencies, which shows that modeling the dependencies reduces the variation of estimates.

6.5.1 The generative model for the synthetic data

In Section 6.2.4, we fit a linear regression model with correlated lines. Based on this model, we generate outage counts according to the following model:

$$N_i \sim \text{Poisson}(\lambda_i G) \tag{6.21}$$

$$G \sim \text{Gamma}(a, a) \tag{6.22}$$

$$\ln \boldsymbol{\lambda} \sim \mathcal{N}(m\mathbf{1} + \beta_L \boldsymbol{x}_L + \beta_V \boldsymbol{x}_V, \boldsymbol{\Sigma}) \tag{6.23}$$

The parameters in (6.21–6.23) are assigned values according to the linear regression model with correlated lines. That is, $m = -1.5$, $\beta_L = 0.13, \beta_V = 0.12$, and $\boldsymbol{\Sigma} = 0.52\boldsymbol{\Sigma}_1 + 0.48\boldsymbol{\Sigma}_2$, which models the line dependencies.

Once we draw a sample from (6.23), the failure rate is known and fixed. So the variation of outage counts comes from the Poisson and Gamma distributions. In particular, using $EG = 1$, we derive from (6.21), (6.22) that the mean of N_i is the same as only using a Poisson distribution and that a controls the overdispersion:

$$\mathrm{E}N_i = \mathrm{E}[\mathrm{E}[N_i|G]] = \lambda_i \tag{6.24}$$

$$\mathrm{Var}N_i = \mathrm{E}[\mathrm{Var}[N_i|G]] + \mathrm{Var}[\mathrm{E}[N_i|G]] = \lambda_i + \lambda_i^2/a \tag{6.25}$$

The value of a is chosen so that the variance of the model matches the empirical variance calculated from the data. In particular, we find the quadratic that best fits the relationship between the empirical variance and mean to be $\sigma^2 = 0.14 + 0.54\mu + 0.53\mu^2$ (where σ^2 is the variance, μ is the mean). Since the coefficients of μ and μ^2 are close, we choose $a = 1$.

We generate three datasets with different sizes so that we have the equivalents of 1-year, 5-year, and 100-year data:

1) draw a sample of $\ln \boldsymbol{\lambda}$ from the multivariate normal distribution (6.23); 2) draw a sample of G from the Gamma distribution (6.22); 3) draw samples of N_i from the Poisson distribution (6.21) n times ($n \in \{1, 5, 100\}$). Thus, we obtain n annual outage counts for each line, and we know the true values of the outage rates $\boldsymbol{\lambda}$.

6.5.2 Comparing to the conventional estimates

The conventional estimates of outage rates are average outage counts per year. The conventional estimates and their standard deviations are obtained using Monte Carlo simulation: draw $B = 1000$ samples according to model (6.21), calculate the average count of each sample, and then calculate the standard deviation of the estimates.

We apply the Bayesian hierarchical model to synthetic datasets using MCMC with the same configuration as in Section 6.3, and use the mean of the posterior distribution as a point estimate.

6.5.2.1 Errors of point estimates

Figure 6.7 shows the distribution of errors of the Bayesian estimates and the conventional estimates (the estimation errors of the Bayesian method and conventional method have the same distribution for 100-year data, so that the plot is not shown). In general, the less the data, the wider the histogram. The error of the conventional estimates has two modes, and the probability of error near zero is lower for 1-year data. As the data size increases, the two modes merge into one. Moreover, for 1-year data, the standard deviation of the error is 0.6 for Bayesian estimates and 0.9 for conventional estimates; for 5-year data, the standard deviation is 0.3 for Bayesian estimates and 0.4 for conventional estimates. Therefore, the Bayesian estimates have a high chance of obtaining more accurate point estimates, especially when data is limited.

On the other hand, there is not much difference in the bias. Specifically, the bias is -0.007 for Bayesian estimates and -0.004 for conventional estimates using 1-year data, and the bias is 0.003 for both Bayesian estimates and conventional estimates using 5-year data.

6.5.2.2 Standard deviation

Figure 6.8 shows the distribution of the ratio of the standard deviation of the Bayesian estimator to that of the conventional estimator. The Bayesian estimator has a lower standard deviation when the data set is smaller. Specifically, the median of the ratio is 0.74 for 1-year data, 0.90 for 5-year data, and 0.99 for 100-year data.

6.5.2.3 Interval estimates

Figure 6.9 shows 95% credible intervals of the Bayes estimator using 1-year, 5-year, and 100-year data respectively. As the size of the dataset increases, we gain more information, and the width of the credible intervals decreases. Figure 6.9 also shows the true values of the outage rates as black dots. As expected with a 95% credible interval, approximately 5% of the true values lie outside the credible interval. The Bayesian point estimates (not indicated in Figure 6.9) lie in the center of the credible intervals and tend to be larger than the true values for low outage rates and

smaller than the true values for high outage rates. This can be explained as the shrinkage towards the mean expected with Bayesian methods; see [44, Sec. 1.5].

6.5.3 Comparing to the Bayesian hierarchical model with independent lines

Previous work does not compute individual line outage rates while considering spatial dependencies between lines. We test the effect of the spatial dependence by removing it. The Bayesian hierarchical model with independent lines is:

$$N_i \sim \text{Poisson}(\lambda_i t_i) \tag{6.26}$$

$$\lambda_i \sim \text{Gamma}(\alpha, \alpha/\mu_i) \tag{6.27}$$

$$\ln \mu_i = \beta_0 + \beta_L x_{Li} + \beta_V x_{Vi} \tag{6.28}$$

The prior distributions are:

$$\begin{aligned} \alpha &\sim \text{Half Normal}(0.7, 8^2) & \beta_L &\sim \text{Normal}(0.13, 5^2) \\ \beta_0 &\sim \text{Normal}(0, 1) & \beta_V &\sim \text{Normal}(0.12, 5^2) \end{aligned} \tag{6.29}$$

We apply this restricted model to synthetic datasets using MCMC with the same configurations as in Section IV. The standard deviation when considering line dependencies is smaller than that without considering line dependencies. The medians of standard deviation ratios of this model to the conventional estimator for 100-year, 5-year, and 1-year data are 0.99, 0.93, and 0.89, which are greater than standard deviation ratios of the Bayesian model with line dependencies to the conventional estimator.

6.6 Conclusion and discussion

We use a Bayesian hierarchical model to improve the estimation of annual outage rates for individual transmission lines. This Bayesian method incorporates several types of dependencies between lines and is applied to real outage data and tested with synthetic data. Particularly for the shorter observation periods with the lower outage counts, the Bayesian estimates perform

better than the conventional estimates that simply divide the number of outages by the observation time: estimates of the individual line outage rates are more accurate, and the uncertainty of the estimates is reduced. Moreover, the comparison with a Bayesian model assuming spatially independent lines shows modeling line spatial dependencies reduces the standard deviation of estimates.

Our Bayesian hierarchical model offers an improvement over the conventional estimates for two reasons. Firstly, the Bayesian method can appropriately capture our prior knowledge of the parameter uncertainties with prior distributions. Secondly, because the model is hierarchical and models the dependence between lines, information about multiple partial commonalities can be appropriately shared across similar lines. These reasons imply that estimates can be improved for lines with no (or a small number of) outages.

Geographically close and neighboring lines experience similar weather conditions, may have a similar design, and share some physical and engineering interactions through the network. We model these line dependencies as a covariance matrix in the Bayesian hierarchical model. The covariance matrix is the weighted sum of two kernels that represent geographic district commonalities and network line proximity, respectively. The Bayesian model learns the weights of the two kernels from the outage data. Our modeling of these dependencies can be realized from a single utility outage dataset that is routinely collected, since the line district is recorded in the dataset, and the network can be readily deduced from the dataset [1]. Using only one dataset is advantageous since coordinating and combining different datasets is often arduous. However, it is conceivable that further advantage could be gained by including other factors such as average wind speed or altitude.

Previous work has often assumed that transmission line outage rates are proportional to line length [49] or grouped together lines of the same area [48, 51, 54]. We model these dependencies by linear factors in the outage rate, and the Bayesian model learns the weights for these factors. The results for our data are that individual line outage rates are only partially correlated with the

line length or the voltage rating. Therefore, it is more reasonable to consider the outage rate for a whole line instead of the rate per mile.

The Bayesian method estimates the distribution of individual line outage rates. This is an advantage compared to methods that return point estimates, as a complete picture of the uncertainty around estimates is needed to make robust decisions about risk and maintenance. For example, if a line has a high point estimate outage rate that is very uncertain, it may be beneficial to wait to gather more information. If desired, any point or interval estimates can be easily obtained from the distribution, depending on the desired application of the outage rates. The quantification of the uncertainty of estimates is useful when the outage rates are used in other models and simulations. For example, a Monte Carlo simulation of transmission reliability can easily be modified to sample from the outage rate distribution to better capture the uncertainty in the estimated reliability.

We focus on overall line outage rates without considering different outage causes in work. However, the proposed Bayesian method can naturally be extended to investigate line outage rates for specific causes.

When data is limited, which is generally true for power system outage data, Bayesian estimates have smaller uncertainty than conventional estimates. Equivalently, with a specific acceptable standard deviation, the proposed Bayesian method needs less data than the conventional method. Thus, utilities can monitor individual line outage rates with fewer years of recording outages. There is a potential to more quickly identify lines with increasing outage rates and aging problems so that maintenance can be scheduled. For example, if utilities need two years of data using the conventional method to estimate line outage rates with a given uncertainty, they typically only need one year of data using the proposed Bayesian method to obtain an outage rate estimate that meets the same uncertainty requirement.

The general advantages of the hierarchical Bayesian method discussed above suggest benefits for various applications of line outage rates. We apply the hierarchical Bayesian method to start to explore and quantify these benefits in [75], which shows improved performance in detecting

deterioration in line outage rates, quantifying the effect of storms, and a system reliability calculation.

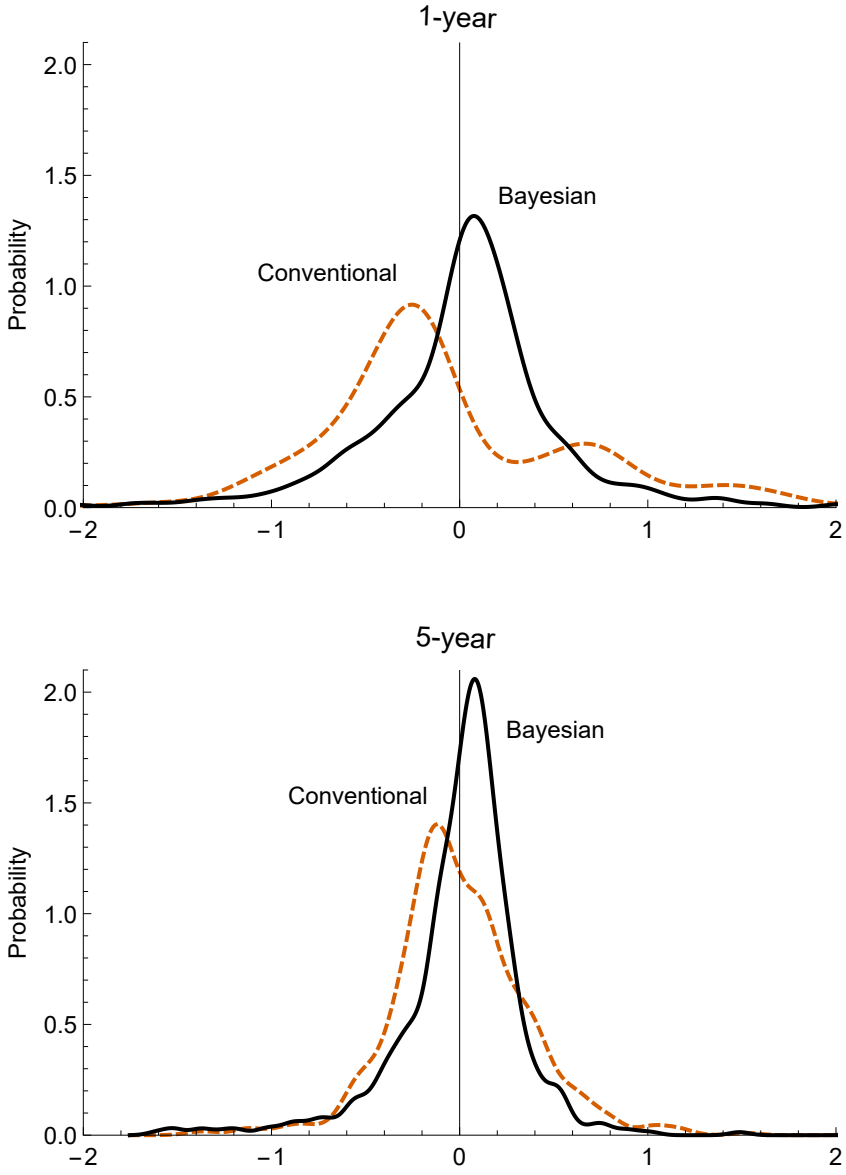


Figure 6.7: Distributions of point estimation errors of Bayes estimates (posterior mean) and conventional estimates using 1-year and 5-year data.

Figure 6.8: Distributions of ratios of standard deviations of Bayes estimator and conventional estimator. The ratio is SD(Bayes)/SD(conventional).

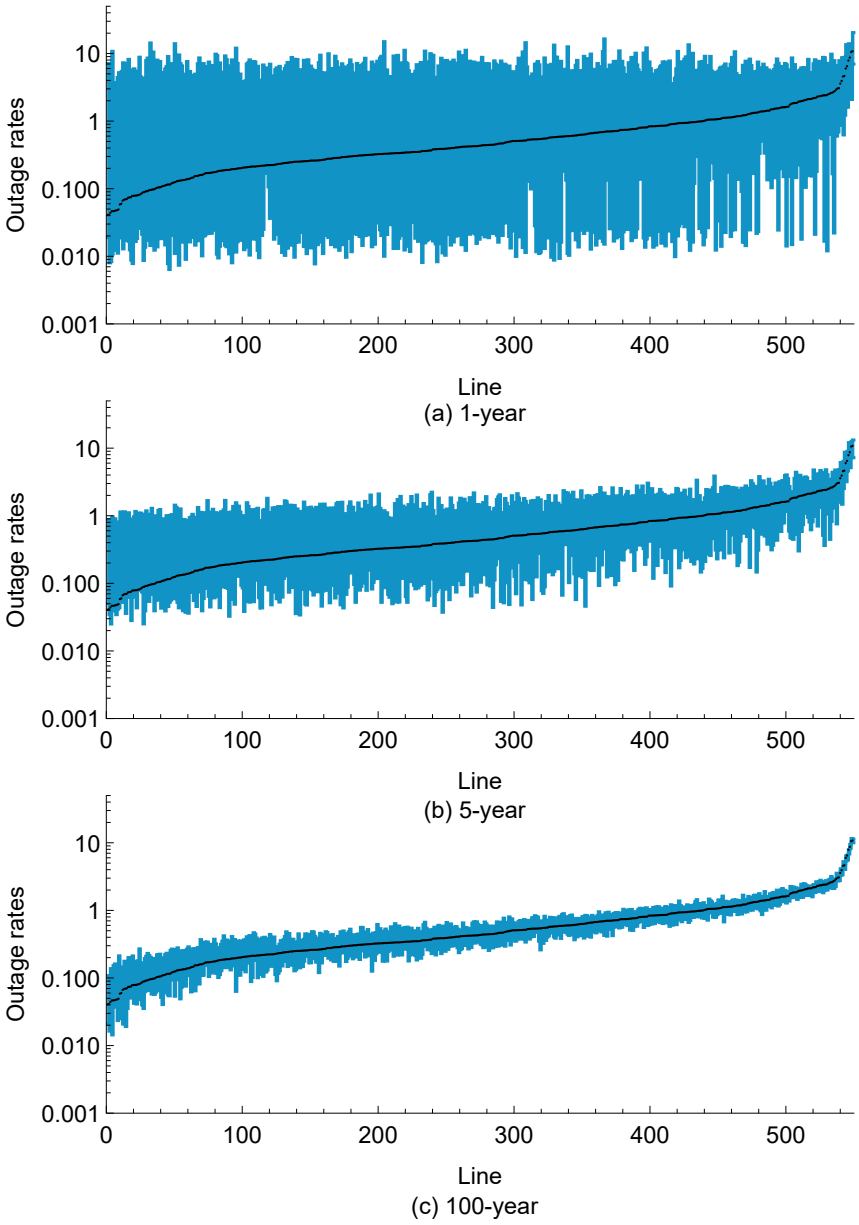


Figure 6.9: 95% credible intervals of Bayesian estimates using 1-year, 5-year and 100-year data. Lines are ordered by outage rates (black dots).

CHAPTER 7. APPLYING BAYESIAN ESTIMATES OF INDIVIDUAL TRANSMISSION LINE OUTAGE RATES

It is a straightforward application of estimating individual transmission line outage rates to identify critical lines in initial outages. As transmission line outage rates are fundamental to power system reliability calculations, it is worth exploring other benefits of better estimates of line outage rates. This chapter explores what can be achieved with this new Bayesian hierarchical model using real utility data. In particular, we assess the capability to detect increases in line outage rates over time, quantify the influence of bad weather on outage rates, and discuss the effect of outage rate uncertainty on a simple availability calculation.

This chapter is developed with assistance from L. Wehenkel, University of Liège, Belgium, J. R. Cruise, Riverlane Research, UK, C. J. Dent and A. Wilson, University of Edinburgh, Scotland. The material in this chapter is published in [98].

7.1 Introduction

Chapter 6 shows that the Bayesian hierarchical model can leverage multiple partial similarities to get better estimates of individual line outage rates from utility data. Typical results are that, for the lines with less frequent outages and using one year of data, the annual outage rates estimated with Bayesian methods have less than half the standard deviation of the conventional estimates. Another way to state this typical result is that the Bayesian hierarchical model for one year of data gives the same accuracy as the conventional estimator for two years of data. Thus, the Bayesian hierarchical model mitigates to some extent the problem of estimating individual line outage rates. This chapter explores how much advantage can be gained from applying our method to give these improved line annual outage rates from utility data. In the following sections, we consider three problems:

7.1.0.1 Detecting lines with reduced reliability

We determine with statistical validity which lines have deteriorated reliability over time to better discriminate which lines should be considered for further analysis and maintenance or upgrade.

7.1.0.2 Storm and no storm data

We often want to partition the data set to get more specific information, and we illustrate the capability of our proposed method in this regard by comparing line outage rates during storms with line outage rates when there is no storm.

7.1.0.3 Effects on reliability calculations

The Bayesian hierarchical model not only gives better estimates of individual line outage rates but also gives the uncertainty of these estimates. We discuss a simple example of an availability calculation to illustrate the impact of these advantages on a system reliability calculation.

7.2 Detecting lines with increased outage rates

It is desirable to examine historical transmission line outages and judge whether the outage rate has increased and the reliability of the line has deteriorated. If there is a high chance that the outage rate has increased significantly, then the condition of this line should be evaluated and decisions about its maintenance, operational limits, or upgrade could be considered[1]. This section applies Bayesian estimates to this problem.

We divide the 14 years of utility data into the first 7 years and the last 7 years. Applying the Bayesian method for each line k, we obtain an outage rate probability distribution $\lambda_k^{(1)}$ for the first 7 years and an outage rate probability distribution $\lambda_k^{(2)}$ for the last 7 years. An example is shown in Figure 7.1.

[1]Similarly, we note that detecting significantly decreased outage rates could be used to verify previous reliability investments.

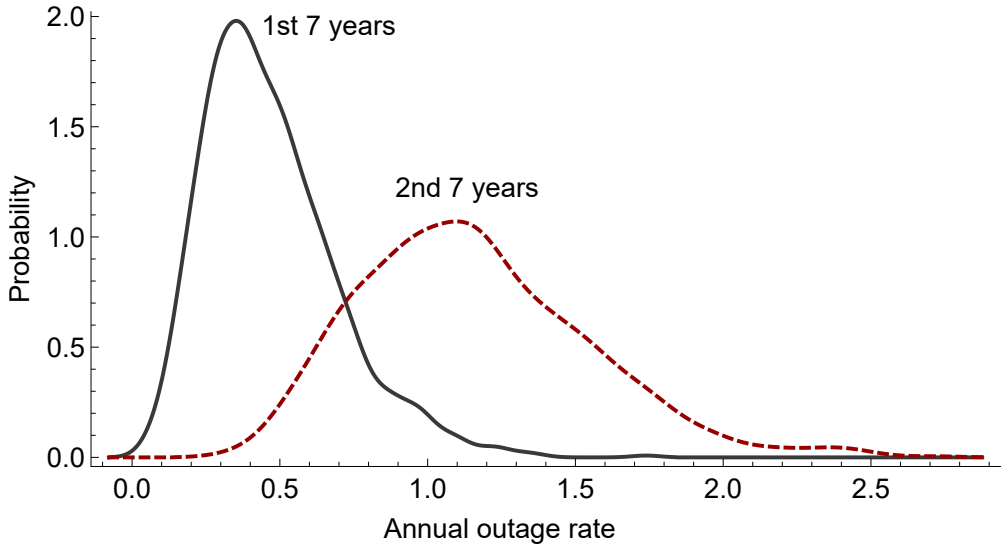


Figure 7.1: The distribution of outage rates for one line in the first 7 years and in the second 7 years.

We are interested in the probability

$$p_k = \mathrm{P}[\lambda_k^{(2)} > \kappa\lambda_k^{(1)}] \tag{7.1}$$

of the line k outage rate increasing by more than some factor κ. We will show results for $\kappa = 1$, $\kappa = 1.5$, and $\kappa = 2$. If $\lambda_k^{(2)} > \kappa\lambda_k^{(1)}$, a larger value of κ indicates a more significant increase in the outage rate.

We evaluate the probability (7.1) empirically by sampling 10 000 times from the probability distributions $\lambda_k^{(1)}$ and $\lambda_k^{(2)}$, which are assumed to be independent. That is,

$$p_k \cong \frac{1}{10\,000}[\text{number of samples with } \lambda_k^{(2)} > \kappa\lambda_k^{(1)}] \tag{7.2}$$

Figure 7.2 shows the probability p_k for each line k. We choose a significance level 0.05; that is, if $p_k > 0.95$, $\lambda_k^{(2)}$ is significantly greater than $\kappa\lambda_k^{(1)}$, and we conclude that the outage rate for line k increases significantly in the last 7 years. According to this rule, we identify 31 lines with increased outage rates for $\kappa = 1$, 8 lines for $\kappa = 1.5$, and 1 line for $\kappa = 2$.

After identifying those lines with significant increases in outage rates, it is worthwhile checking the outage records to find out more about the possible specific causes of the increases, such as the

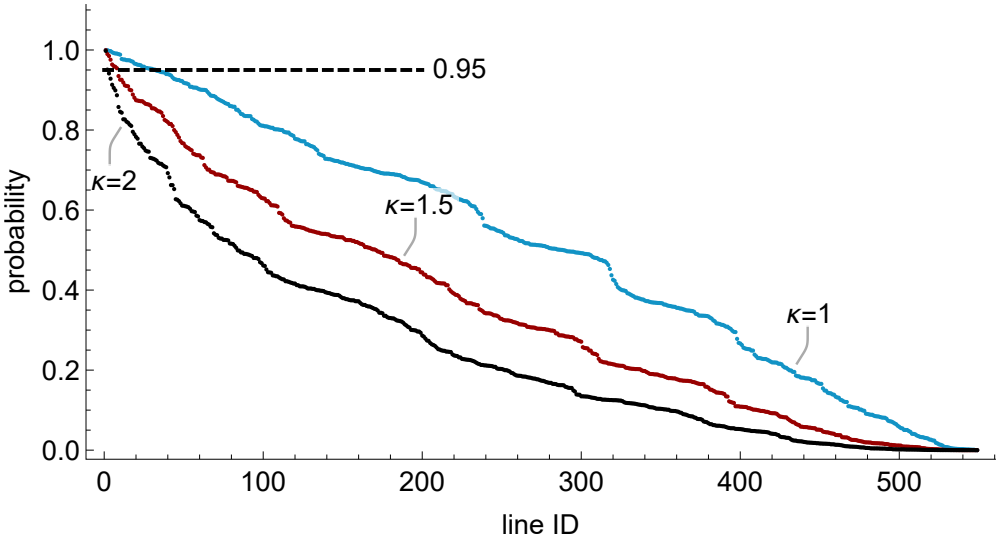


Figure 7.2: The probability p_k that the outage rate increases by at least a factor of κ in the second half time period from the first time period for each line k. Lines are ordered by the $\kappa = 1$ probabilities.

cause codes for each outage. Utilities have much richer information about the outage and the grid conditions and can investigate much further. To illustrate this, we check the top three lines with the highest probability of outage rate increases. Table 7.1 shows the observed counts for these three lines, and there is an obvious increase in counts during the last 7 years. Outage causes for line 151 are mainly recorded as foreign utility and foreign trouble. Given the lack of outages in years 1 through 6, it should also be checked whether the line was newly installed in year 7. Line 138 has various cause codes during the last 7 years, such as tree blown, line material failure, foreign trouble, and vegetable management. Line upgrade or tree-trimming may help lower the outage rate. Most of the causes for line 539 are wind related, so weather variations have a big influence on this line, and the line spacers, damping, and icing could also be reviewed.

We compare the results with the conventional method, which estimates mean annual outage rates of individual lines by simply evaluating the average outage counts in a year. The standard deviation of the annual counts is also estimated. Then we fit a Gamma distribution for each line in each 7-year period using the method of moments. (Here we prefer the method of moments to maximum likelihood estimation because there are only 7 data points, and several of them are

Table 7.1: Observed outage counts of the top three lines with highest probability of increases in annual outage rates

Line	Outage counts in different years													
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14
151	0	0	0	0	0	0	0	1	2	9	0	1	2	0
138	0	0	0	1	2	1	2	2	5	5	2	0	3	6
539	0	0	0	1	0	0	0	2	2	2	1	1	1	1

zeros; maximum likelihood estimation would exclude the zero observations, which reduces the information further and the optimization to find the maximum likelihood may fail.)

Note that the conventional method cannot deal with the lines with all zero counts in a 7-year period, while the Bayesian hierarchical model can solve this case. So we compare the two methods for lines with at least one nonzero count.

For each line in each 7-year period, we sample from the fitted Gamma distribution and use the same sampling method described at the beginning of this section to estimate the probability p_k. We call this procedure the "basic method" (the mean estimation is conventional, but we are not sure to what extent industry computes uncertainty of the conventional mean estimate). This basic method identifies 2 lines with increased outage rates for $\kappa = 1$, 1 lines for $\kappa = 1.5$, and no lines for $\kappa = 2$. Thus the increased uncertainty for the basic method detects significantly fewer lines with statistically verified increased outage rates.

The two lines identified by the basic method are line 539 and line 32. Line 539 is also identified in the above Bayesian method. The basic method does not identify line 151 because this line has no outage in the first 7 years. Line 138 is not identified by the basic method but is identified by the Bayesian method. The posterior distribution of the outage rate for line 138 has mean 0.81 and standard deviation 0.29 in the first 7 years, and mean 2.93 and standard deviation 0.62 in the second 7 years. Whereas in the basic method for line 138, the Gamma distribution has mean 0.86 and standard deviation 0.90 in the first 7 years, and mean 3.29 and standard deviation 2.14 in the second 7 years. The standard deviation of the posterior distribution is obviously lower than the standard deviation of the Gamma distribution. This low standard deviation makes the

distributions in the two 7-year periods sufficiently different, while the two Gamma distributions overlap due to their larger standard deviation.

Figure 7.3 compares the means and standard deviations of the posterior distribution in the Bayesian method and the Gamma distribution in the basic method. Although the two methods have close means, the posterior distribution has a smaller standard deviation. This observation confirms the result in our journal paper [91] that hierarchical Bayesian estimates of outage rates have a lower standard deviation than the conventional estimates. The lower uncertainty of the Bayesian estimates explains why the Bayesian method more effectively detects lines with significant outage rate increases.

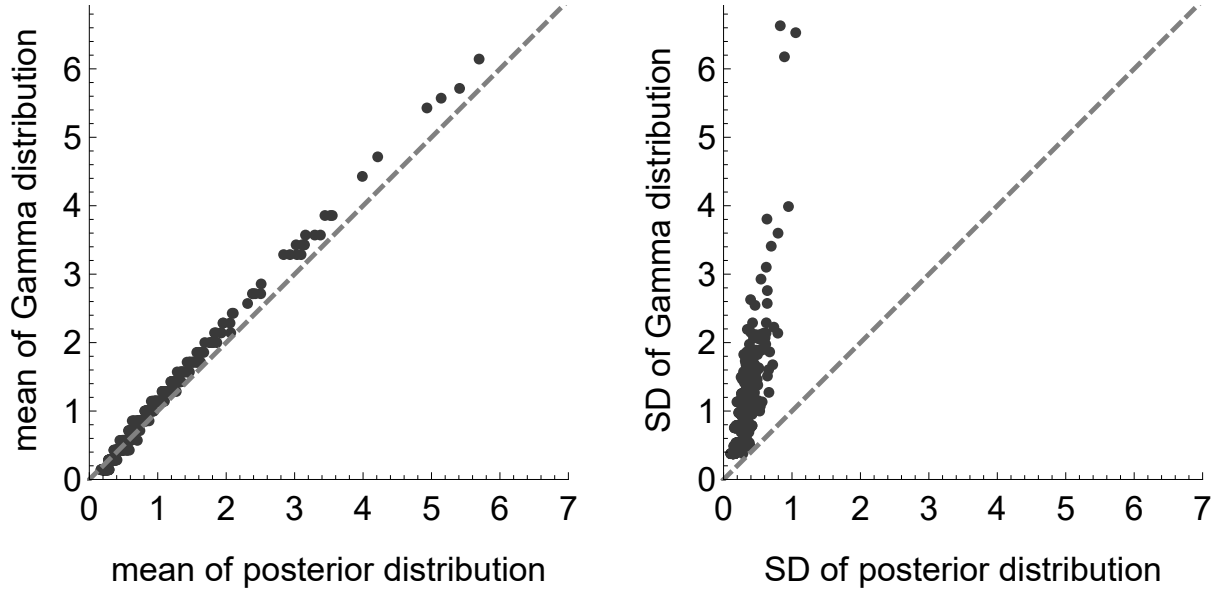


Figure 7.3: Comparing means (left panel) and standard deviations (right panel) of the posterior distribution and the Gamma distribution in two methods.

7.3 Effect of storms on outage rates

Since the proposed Bayesian hierarchical method mitigates the limited data problem in estimating individual outage rates, we can study further by investigating a subset of the outage data. For example, we can evaluate the effect of weather on outage rates.

We define a line outage as a storm outage if it occurs during a storm, otherwise it is called a non-storm outage. Then the annual storm outage rate is the number of storm outages divided by the total storm time in a year; similarly, the annual non-storm outage rate is the number of non-storm outages divided by the total non-storm time in a year. (Note that the Bayesian model does not directly produce the storm/non-storm outage rate. It outputs the average storm/non-storm outages over a year without considering the storm/non-storm time. So we need to divide the average storm/non-storm outages over a year by the storm/non-storm probability, which is the storm/non-storm time divided by the total time.)

The weather data is from the USA National Oceanic and Atmospheric Administration (NOAA) which includes storm events and other significant weather phenomena [99]. Using the method described in [100], we classify outages as storm outages and non-storm outages.

Figure 7.4 compares the storm and non-storm outage rates estimated using the Bayesian hierarchical model. 93% of lines have storm outage rates greater than non-storm outage rates (using the posterior mean as point estimation). The average storm outage rate is 4.5 per year which is nine times greater than the average non-storm outage rate 0.5 per year. This result confirms the finding in [100] and provides more information due to the lower uncertainty of the Bayesian estimates.

The BPA data records the cause code reported for each outage. Table 7.3 tallies the frequency of each dispatcher cause code in the recorded outage data. We now summarize how storms affect the cause codes. The proportion of cause codes "tree blown", "wind", "weather", "ice", "power system condition", "SCADA", and "galloping conductors" for storm outages are at least one order of magnitude greater than that for non-storm outages; while the proportion of causes "equipment/miscellaneous", "RAS initiated", "human element", "foreign utility", "fire", "smoke", and "maintenance" for non-storm outages are at least one order of magnitude greater than that for storm outages. Fewer human errors are reported during storms, as "human element" is a cause for 0.08% of storm outages, compared to 1% for non-storm outages, and the proportions of "dispatcher" are about the same. Storms increase the outages caused by "SCADA"

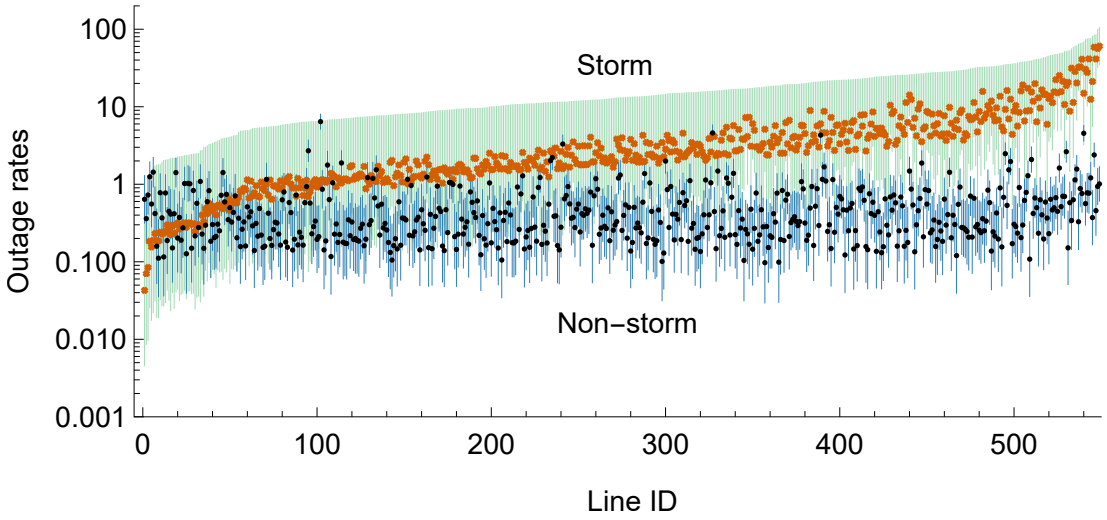


Figure 7.4: 95% credible intervals of outage rates and posterior means (sorted according to the upper bound of storm outage rates). Orange crosses are storm outage rates, and black dots are non-storm outage rates.

and "galloping conductors", but do not increase other equipment causes such as "improper relaying", "terminal equipment failure", and "arc while switching".

7.4 Effect of outage rate variation on a simple unavailability calculation

This section shows how variation and uncertainty in outage rates affect an elementary transmission system reliability calculation. One of the simplest idealized availability calculations has 3 lines that minimally satisfy the N–1 criterion; that is, the system is available if all lines or 2 out of 3 lines are operating, and unavailable otherwise. The 3 lines are independent with exponential failure rates λ_1, λ_2, λ_3 and exponential repair rate μ. State 1 is no lines out, states 2,3,4 are one line out, states 5,6,7 are two lines out, and state 8 is three lines out. The Markovian

transition rate matrix Q is

$$\begin{bmatrix}
-\lambda_1-\lambda_2-\lambda_3 & \lambda_1 & \lambda_2 & \lambda_3 & 0 & 0 & 0 & 0 \\
\mu & -\lambda_2-\lambda_3-\mu & 0 & 0 & \lambda_2 & 0 & \lambda_3 & 0 \\
\mu & 0 & -\lambda_1-\lambda_3-\mu & 0 & \lambda_1 & \lambda_3 & 0 & 0 \\
\mu & 0 & 0 & -\lambda_1-\lambda_2-\mu & 0 & \lambda_2 & \lambda_1 & 0 \\
0 & \mu & \mu & 0 & -\lambda_3-2\mu & 0 & 0 & \lambda_3 \\
0 & 0 & \mu & \mu & 0 & -\lambda_1-2\mu & 0 & \lambda_1 \\
0 & \mu & 0 & \mu & 0 & 0 & -\lambda_2-2\mu & \lambda_2 \\
0 & 0 & 0 & 0 & \mu & \mu & \mu & -3\mu
\end{bmatrix}$$

The steady state probability distribution of states is given by the row vector π, where $\pi Q = 0$ and the entries of π add to one. The probability of unavailability is the sum of the last 4 entries of π. It is convenient to express the unavailability as the expected number of minutes of unavailability in a year by multiplying the probability of unavailability by 525 600, the number of minutes in a year.

In our raw line data, the mean outage rate is $\overline{\lambda} = 0.6$ per year, and the standard deviation is 0.7. The mean restoration time is 907 minutes [101], which corresponds to the restoration rate of $\mu = 579$ per year that we use throughout this section.

We consider the effect of using an average outage rate for all three lines when their outage rates differ. A parameter α is used to control the variation of the line outage rates while keeping the mean outage rate constant. The outage rates in Table 7.2 satisfy $\lambda_2 = \alpha\lambda_1$, $\lambda_3 = \lambda_1/\alpha$, and $\text{Mean}\{\lambda_1, \lambda_2, \lambda_3\} = 0.6$ for several values of α. $\alpha = 5$ gives a plausible variation of outage rates (one standard deviation from the mean is 0.6 ± 0.7) and approximately half the unavailability. That is, if the outage rates do vary according to $\alpha = 5$, then using an average outage rate for all three lines approximately doubles the unavailability.

We consider the effect of uncertainty in the estimated line outage rates on the unavailability. We model the uncertainty in estimates of $\lambda_1, \lambda_2, \lambda_3$ by three independent Gamma distributions, each with mean 0.6 and standard deviation σ. The resulting probability distributions in the

Table 7.2: System unavailability for several annual outage rates

α	λ_1	λ_2	λ_3	unavailability
1	0.6	0.6	0.6	1.7 min
2	0.51	1.03	0.26	1.4 min
5	0.29	1.45	0.06	0.8 min

unavailability are shown in Figure 7.5 for $\sigma = 0.7$ and $\sigma = 0.17$. $\sigma = 0.7$ is the average of the standard deviations of individual transmission lines used in the basic estimation in section 7.2. $\sigma = 0.17$ is the average of the standard deviations of individual transmission lines in the Bayesian estimation.

If we neglect the uncertainty in the estimated outage rates, the deterministic calculation with $\lambda_1 = \lambda_2 = \lambda_3 = 0.6$ gives an unavailability of 1.69 minutes. If we use the average uncertainty $\sigma = 0.17$ that is typical of the Bayesian estimates, the 95% probability interval for the unavailability is $\{1.25, 3.01\}$. If we use the average uncertainty $\sigma = 0.7$ that is typical of that used in the basic method of section 7.2, the 95% probability interval for the unavailability is $\{0.24, 8.63\}$. For this example, a typical uncertainty in the line outage rates appreciably affects the unavailability. The smaller uncertainty provided by the Bayesian estimates is clearly advantageous compared to the uncertainty provided by the basic method in section 7.2.

We note that either the Bayesian or basic method considered above of estimating the standard deviation of individual line outage rates is better than conventionally estimating the outage rates of all the lines and computing the mean and standard deviation of this combined data. This procedure gives a standard deviation of 1.14, which is larger because it includes not only the uncertainty of individual line estimates but also the variation in individual line outage rates from their combined mean. In the unavailability calculation, the larger standard deviation $\sigma = 1.14$ gives unacceptably large variation in the calculated unavailability, with a 95% probability interval $\{0.02, 14.27\}$.

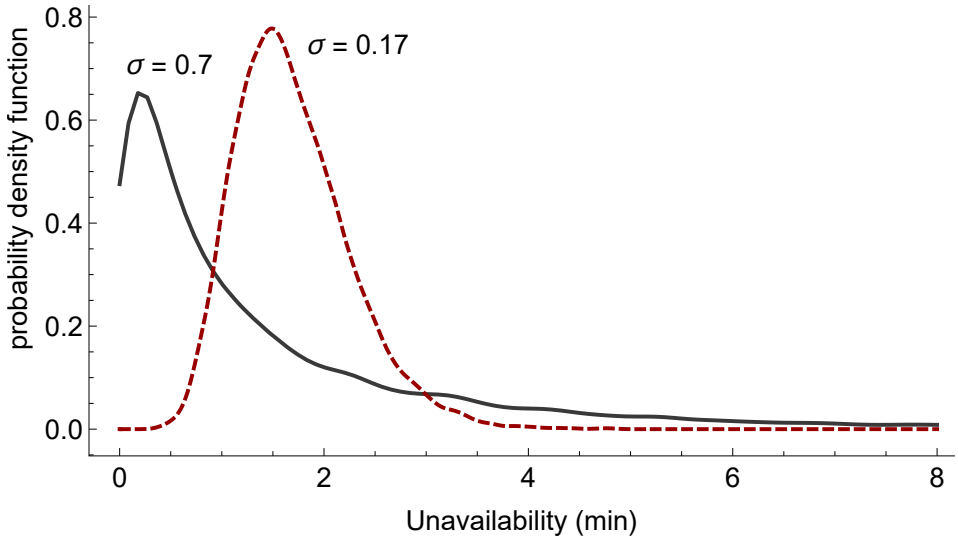


Figure 7.5: Probability distributions of calculated unavailability for several values of standard deviation σ for the estimated line outage rate. The distribution of unavailability has standard deviation 2.6 for $\sigma = 0.7$ and standard deviation 0.56 for $\sigma = 0.17$. The mean unavailability is 1.69 minutes per year.

7.5 Conclusion

The Bayesian hierarchical model can process standard transmission line outage data routinely collected by utilities to give improved estimates of individual line outage rates [91]. When the outage counts are low, the Bayesian hierarchical model estimates have lower variance than the conventional calculation of outage rate that simply divides counts of outages by the observation time elapsed. The Bayesian model does this by combining line data with data from other lines with partial similarities in rating, length, and proximity. This work uses real utility data to explore several ways in which the improved performance of the Bayesian outage rates for individual lines can be exploited.

It is useful to be able to detect deterioration in line outage rates so that corrective action can be taken. We use the Bayesian outage rates to calculate the probability that an individual line outage rate has increased in a second 7-year period compared to a first 7-year period. Since the Bayesian outage rates have lower uncertainty, they can better detect significant outage rates increases in more lines than a basic conventional method. The significant increase in outage rate

can be used to select lines that are likely to have deteriorated reliability in a principled way, so that these lines can be further investigated to inform upgrade, maintenance, modification, or derating decisions.

It is useful to split historical data sets for separate analyses to investigate the effect of factors such as storms. We illustrate the performance of the Bayesian method in distinguishing storm and no-storm outage rates. For our utility data, the average storm line outage rate is 4.5 per year, which is nine times the average non-storm line outage rate of 0.5 per year.

Bayesian methods calculate probability distributions of line outage rates, so that the mean gives a point estimate of the outage rate and the standard deviation indicates the uncertainty of the point estimate. It is desirable to account for the uncertainty of outage line rate estimates in transmission system reliability calculations, and the Bayesian uncertainties are smaller than the conventional uncertainties. To start to discuss and quantify the effect of this on system reliability calculations, we contrast Bayesian hierarchical models and conventional methods for an elementary availability computation for a 3-line system. For this computation, using individual line outage rates as opposed to average outage rates for pooled data can halve the unavailability. Moreover, the reduced uncertainty of the Bayesian outage rates compared to conventional uncertainties gives significantly smaller probability intervals for the unavailability.

Overall, our results indicate that the reduced uncertainty in individual line outage rates enabled by the Bayesian hierarchical model can be useful. We also expect that routinely quantifying the uncertainty in individual line outage rates will help to better justify decisions based on reliability calculations that depend on these outage rates.

Table 7.3: Dispatcher cause code frequency in BPA outage data during 1999 to 2012

CAUSE CODE	ALL	>20 min
Foreign Trouble	3163	1285
Unknown	2433	181
Lightning	2296	124
Terminal Equipment Failure	298	123
Forced (Configuration)	242	87
Wind	238	80
Tree blown	226	198
RAS Initiated	202	32
Weather	194	55
Equipment/Miscellaneous	187	51
Line Material Failure	175	119
Fire	129	37
Improper Relaying	115	35
Foreign Utility	114	33
Human Element	114	15
Ice	76	39
Maintenance	73	16
Smoke	72	18
Malicious	66	61
Substation Operations	58	5
Tree	57	52
Tree cut	49	39
Construction	47	6
Staged Test	41	2
Imp Install/Design/Applica	40	8
Arc while switching	34	11
Bird droppings	30	7
Sympathetic	26	7
Dispatcher	20	3
Machinery, Construction	15	8
Machinery, Logging	15	9
Bird or Animal	15	10
TT Noise	14	2
Earthquake	11	9
Vehicle	9	8
Foreign Object	6	3
Aircraft	6	3
Voltage	5	3
Machinery, Farming	5	3
Galloping Conductors	5	3
Power System Condition	5	1
Earth slide	4	3
Contamination	4	2
Tree growth	3	1
Industrial	2	2
Agricultural	1	1
SCADA	1	0
Frequency	1	1

CHAPTER 8. N-k CONTINGENCY SELECTION USING NETWORK MOTIFS AND SPATIAL STATISTICS OBSERVED IN OUTAGE DATA

8.1 Motivation

Selecting contingencies or initial outages according to their probability is significant to accurately evaluate the power system resilience and cascading risk. If N-k contingencies are chosen to initiate cascading outages, such as using the Random Chemistry algorithm in [84], we would overestimate the cascading risk or underestimate the system resilience. Indeed, if we identify contingencies that cause severe impact and then take corresponding preventive/ corrective actions, the system would be more resilient. However, with limited resources, it is more effective to identify high-risk contingencies instead of only high-impact contingencies. The risk measurement usually has two dimensions: impact and frequency. Simulations using power system models with various levels of details can estimate the impact of certain contingencies, however, it is difficult to model and simulate all the rare contingencies so here we pursue the estimation of contingency frequency through real outage data. Therefore, this chapter studies the contingency probability exploiting network theory.

Another more important motivation is that the contingency analysis, especially for multiple contingencies, has been a challenge for many years. Contingency analysis is one of the three main functions of power system security that includes system monitoring, contingency analysis, and security-constrained optimal power flow. Contingency analysis aims to detect system problems in advance and take necessary preventive and corrective actions so that the system can withstand the impact of an contingency and operate normally without violations in line thermal limits, bus voltages, and dynamic stability. It requires operators to test all possible credible contingencies. This is a challenging problem because theoretically there are $\binom{N}{k}$ contingencies for k-component contingencies in a N-component system, which is a huge number even for $k = 2$ in interconnected

power systems. Contingency analysis must be analyzed quickly so that the results are useful for operators. An observation is that not all contingencies would cause violations in power systems. Therefore, we can select contingencies that could cause problems, and only study these limited credible contingencies in detail. This is contingency selection, which produces a list of credible contingencies. Many researches have studied the multiple contingency selection based on model-based methods, graph theory, statistical sampling, and optimization methods. This study, however, selects multiple contingencies based on the observed probability of their graphical patterns that occurred historically in the power network.

8.2 Multiple contingencies occur frequently in contingency motifs

We represent multiple contingencies as subgraphs of the power network. Some patterns are frequently recurrent, and we adapt and apply a concept, network motifs, to represent those patterns. The presence of motifs reflects the basic structure of power systems. Thus, they give a general and practical guidance on contingency selection.

Before we give the formal definition of contingency motifs, it is necessary to describe the statistics of random patterns of the power network and statistics of patterns appeared in outage data.

8.2.1 Subgraphs of the power network

The power transmission grid is comprised of substations and transmission lines. It is represented by a graph/network as shown in Figure 8.9, which is a utility in the Northwest of the US. Substations correspond to nodes and transmission lines correspond to edges/lines. The power grid sometimes has multiple transmission lines between two substations, and they are represented by one line in the power network in this study.

A k-edge subgraph $s_{k,i}$ is an edge-induced subgraph, which is a subset of edges of a graph together with vertices that are their endpoints. Figure 8.1 illustrates a 2-edge subgraph $\{1-3,\ 1-6\}$.

Two subgraphs are isomorphic when there exists a mapping between their vertices such that two vertices are adjacent in one subgraph implies that the two corresponding vertices in the other subgraph are also adjacent. We say two subgraphs are the same when their nodes and edges are exactly the same. For example, in Figure 8.1, subgraph $\{1-3,\ 1-6\}$ and $\{1-5,\ 4-5\}$ are isomorphic, but they are different subgraphs.

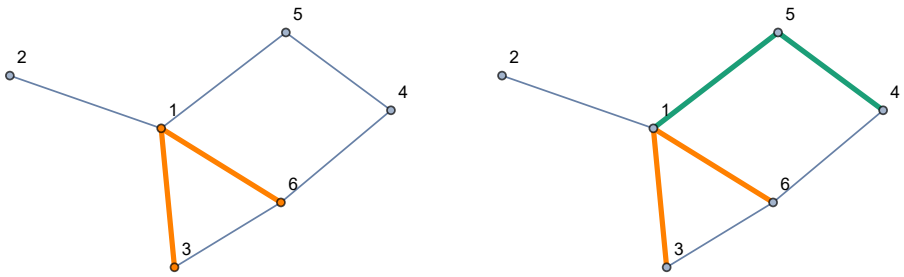


Figure 8.1: Example of different subgraphs. (A) 2-edge subgraph. (B) Orange subgraph and green subgraph are isomorphic subgraphs.

A pattern is a set of isomorphic subgraphs in the network. $S_{k,i}$ denotes a pattern that is a set of $s_{k,i}$, where k is the number of edges and i is the ID of that pattern. The exception is $S_{4,*}$, which is the set of 4-edge subgraphs that are not the members of $S_{4,i}$ for $i = 1, 2, 3, 4$. Figure 8.2 shows some patterns of the power network and the number of all distinct subgraphs in each pattern, which is also the size $|S_{k,i}|$ of the pattern.

8.2.2 Statistics of multiple contingencies

When a $N-k$ contingency occurs, we can imagine the k outaged lines in the power network are highlighted and we have a subgraph. These subgraphs are called contingency subgraphs. Thus, each multiple contingency corresponds to a subgraph $s_{k,i}$. As multiple contingencies are always grouped as $N-k$ contingencies, the corresponding contingency subgraphs are also grouped according to the number of edges k. Figure 8.3 shows the statistics of the contingency subgraphs observed in outage data.

If no other information is available, it is natural to assume all contingency subgraphs occur uniformly in all k-edge subgraphs. We call this the uniform assumption. Let $p^{\text{uni}}_{s_{k,i}}$ be the

	$S_{k,i}$	$\lvert S_{k,i} \rvert$	$p^{uni}_{k,i}$	n_k	$n_{k,i}$	$\hat{p}^{obs}_{k,i}$
	$S_{2,1}$	2116	0.015209	392	317	0.808673
	$S_{2,2}$	137 012	0.984791	392	75	0.191327
	$S_{3,1}$	4653	0.000191	127	74	0.582677
	$S_{3,2}$	1 083 833	0.044431	127	31	0.244094
	$S_{3,3}$	7519	0.000308	127	18	0.141732
	$S_{3,4}$	62	$3. \times 10^{-6}$	127	3	0.023622
	$S_{3,5}$	23 297 709	0.955068	127	1	0.007874
	$S_{4,1}$	9799	$3. \times 10^{-6}$	23	9	0.391304
	$S_{4,2}$	2 354 215	0.000735	23	5	0.217391
	$S_{4,3}$	48 581	0.000015	23	5	0.217391
	$S_{4,4}$	26 028	$8. \times 10^{-6}$	23	2	0.086957
others	$S_{4,*}$	3 199 244 477	0.999238	23	2	0.086957

Figure 8.2: Probabilities of patterns in outage data and in random subgraphs. $S_{4,*}$ is the set of 4-edge subgraphs that are not the members of $S_{4,i}$ for $i = 1, 2, 3, 4$.

probability of $s_{k,i}$ under the uniform distribution, where "uni" indicates the uniform assumption. Then,

$$p^{\text{uni}}_{s_{k,i}} = \frac{1}{\binom{n}{k}} \tag{8.1}$$

where n is the number of lines in the power network, which is also equal to N as in $N - k$ contingencies. And the probability of a pattern $p^{\text{uni}}_{k,i}$ given k under the uniform assumption is

$$p^{\text{uni}}_{k,i} = \frac{|S_{k,i}|}{\binom{n}{k}} \tag{8.2}$$

	$S_{k,i}$	frequency	frequency (%)		$S_{k,i}$	frequency	frequency (%)
	$S_{2,1}$	317	58.5		$S_{2,2}$	75	13.8
	$S_{3,1}$	74	13.7		$S_{3,2}$	31	5.7
	$S_{3,3}$	18	3.3		$S_{3,4}$	3	0.6
	$S_{3,5}$	1	0.2				
	$S_{4,1}$	9	1.7		$S_{4,2}$	5	0.9
	$S_{4,3}$	5	0.9		$S_{4,4}$	2	0.4
others	$S_{4,*}$	2	0.4				

Figure 8.3: Statistics of contingency subgraphs in outage data.

Figure 8.2 shows the probabilities of patterns under the uniform assumption.

However, the observations from historical outage data differ greatly from the uniform assumption. For example, $p_{3,1}^{\text{uni}}$ is greater than $p_{3,2}^{\text{uni}}$, while the frequency of $S_{3,1}$ is comparable with that of $S_{3,2}$ in outage data. This observation implies that some patterns recur much more frequently than under the uniform assumption. For those significantly recurrent patterns in contingency subgraphs, we can define them as motifs, which is discussed in detail in the next section. Before that, we estimate the probabilities of different patterns based on outage data.

The probability of $S_{k,i}$ is estimated from the outage data by

$$\widehat{p}_{k,i}^{\text{obs}} = \frac{n_{k,i}}{n_k} \tag{8.3}$$

where "obs" indicates the probability of a pattern is estimated from outage data, $n_{k,i}$ is the number of contingency subgraphs $s_{k,i}$ appearing in the outage data, and n_k is the number of k-edge contingency subgraphs s_k in the outage data. Note $\sum_i n_{k,i} = n_k$.

Assuming contingency subgraphs $s_{k,i}$ are uniformly distributed in $S_{k,i}$, then the probability of a contingency subgraph $s_{k,i}$ based on the outage data is

$$\widehat{p}^{\text{obs}}_{s_{k,i}} = \frac{\widehat{p}^{\text{obs}}_{k,i}}{|S_{k,i}|} \tag{8.4}$$

At the same time, the standard error of the estimate is

$$\sigma\left(\widehat{p}^{\text{obs}}_{s_{k,i}}\right) = \frac{1}{|S_{k,i}|}\sigma\left(\widehat{p}^{\text{obs}}_{k,i}\right) = \frac{1}{|S_{k,i}|}\sqrt{\frac{\widehat{p}^{\text{obs}}_{k,i}\left(1-\widehat{p}^{\text{obs}}_{k,i}\right)}{n_k}} \tag{8.5}$$

In summary, we define a probability $p_{k,i}$ that a pattern $S_{k,i}$ appears in contingency subgraphs. Two methods are proposed to estimate $p_{k,i}$: the uniform assumption as (8.2) and the other is based on observed outage data as (8.3). Figure 8.2 shows $p^{\text{uni}}_{k,i}$ and $\widehat{p}^{\text{obs}}_{k,i}$ for different patterns. Moreover, in the set $S_{k,i}$, all contingency subgraphs $s_{k,i}$ are assumed uniformly distributed so that their probability $p_{s_{k,i}}$ is given by either (8.1) or (8.4).

8.2.3 Definition of contingency motifs

The conventional definition of the motif introduced by Milo considers connected subgraphs with a specific number of nodes. For example, possible size-3 motifs are the subgraphs $\{1-3,\ 1-6,\ 3-6\}$ and $\{1-2,\ 1-5\}$ in Figure 8.1. The detection algorithm computes the relative frequency of each pattern and compares it with the frequency of the pattern in random graphs that has the same global property (eg. degree distribution) as the original network [59, 102]. However, contingency subgraphs are naturally grouped based on the number of edges, not the number of nodes, and they could be disconnected subgraphs. Moreover, the power network is fixed and known. Therefore, the conventional detection of the motif cannot be directly applied, and needs to be adapted to our application to power network contingencies.

Instead of comparing the frequency of a pattern in the network to that in random graphs, we compare the frequency of the contingency pattern in outage data to that in subgraphs sampled randomly from the original network. Therefore, we define a k-edge contingency motif in a power network as a k-edge pattern whose probability of occurrence is significantly greater than that

when all k-edge subgraphs in the network have the same probability of occurrence. For example, to determine 3-edge motifs, we estimate the probability of $S_{3,i}$ for all i from outage data, and then compute the probability of $S_{3,i}$ in the network under uniform assumption. If the probability of a pattern in outage data is significantly greater than that under uniform assumption, we say this patter is a contingency motif. That is, we define $S_{k,i}$ as a contingency motif when

$$p_{k,i} > a p_{k,i}^{\text{uni}} \tag{8.6}$$

where $a \geq 1$ is large enough so that it is a conservative comparison. We let $a = 10$ in this study.

8.2.4 Detecting contingency motifs

To detect a contingency motif, we need to compare the probability of the contingency pattern observed in outage data and the probability of that pattern in random subgraphs under the uniform assumption. This problem can be formulated as a hypothesis test:

$$H_0 : p_{k,i} \leq 10 p_{k,i}^{\text{uni}} \quad \text{versus} \quad H_1 : p_{k,i} > 10 p_{k,i}^{\text{uni}} \tag{8.7}$$

8.2.4.1 Statistical model of multiple contingencies

Let X be the number of $s_{k,i}$ in a total of n_k $N-k$ contingencies. X follows a binomial distribution:

$$P_{S_{k,i}}(X = x) = \binom{n_k}{x} p_{k,i}^{x} \left(1 - p_{k,i}\right)^{n_k - x} \tag{8.8}$$

8.2.4.2 Frequentist hypothesis test

Under H_0, the likelihood of obtaining $n_{k,i}$ or more $s_{k,i}$ is

$$L\left(p_{k,i} | n_{k,i}\right) = \sum_{j=n_{k,i}}^{n_k} \binom{n_k}{j} (10 p_{k,i}^{\text{uni}})^{j} (1 - 10 p_{k,i}^{\text{uni}})^{(n_k - j)} \tag{8.9}$$

When the likelihood is less than significance level 0.01, we reject H_0, which means that the probability that H_0 is true but we reject it is less than 0.01.

8.2.4.3 Bayesian hypothesis test

We compare the posterior probability $P(H_0|n_{k,i})$ with $P(H_1|n_{k,i})$. If $P(H_0|n_{k,i}) \geq P(H_1|n_{k,i})$, we accept H_0; otherwise, we reject H_0 and accept H_1.

$$\begin{aligned} P(H_0|n_{k,i}) &= \frac{P(n_{k,i}|H_0)P(H_0)}{P(n_{k,i})} \\ &= \frac{P(n_{k,i}|H_0)P(H_0)}{P(n_{k,i}|H_0)P(H_0) + P(n_{k,i}|H_1)P(H_1)} \\ &= \frac{1}{1 + \frac{P(n_{k,i}|H_1)P(H_1)}{P(n_{k,i}|H_0)P(H_0)}} \end{aligned} \tag{8.10}$$

where $P(H_0)$ is the prior probability and

$$P(n_{k,i}|H_0) = \int_0^{10p_{k,i}^{\text{uni}}} P(n_{k,i}|p_{k,i}) f(p_{k,i}|H_0) dp_{k,i} \tag{8.11}$$

is the marginal likelihood under H_0. $f(p_{k,i}|H_0)$ is the prior for parameter $p_{k,i}$ when H_0 is true, which is assumed to be a uniform distribution.

The Bayes Factor for H_1 relative to H_0 is defined by

$$BF(H_1 : H_0) = \frac{P(n_{k,i}|H_1)}{P(n_{k,i}|H_0)} \tag{8.12}$$

Assume $P(H_0) = P(H_1) = 0.5$, and $p_{k,i}|H_0$ and $p_{k,i}|H_1$ both follow uniform distributions. Then, the posterior probability of H_0 turns out to be

$$P(H_0|n_{k,i}) = \frac{1}{1 + BF} = \frac{1}{1 + \frac{1 - F(10p_{k,i}^{\text{uni}})}{F(10p_{k,i}^{\text{uni}})}} \tag{8.13}$$

where $F(x)$ is the cumulative density function of a beta distribution with parameters $n_{k,i} + 1$ and $n_k - n_{k,i} + 1$. Formula (8.14) shows the derivation, where $B(a, b)$ is the beta function with parameter a, b, $B(x; a, b)$ is the incomplete beta function, and $I_x(a, b)$ is the regularized incomplete beta function. $I_x(a, b)$ is also the cumulative distribution function of the beta distribution $F(x; a, b)$.

$$\begin{aligned} & BF(H_1 : H_0) \\ =& \frac{P(n_{k,i}|H_1)}{P(n_{k,i}|H_0)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\int_{10p_{k,i}^{\text{uni}}}^{1} P(n_{k,i}|p_{k,i}) f(p_{k,i}|H_1) dp_{k,i}}{\int_{0}^{10p_{k,i}^{\text{uni}}} P(n_{k,i}|p_{k,i}) f(p_{k,i}|H_0) dp_{k,i}} \\
&= \frac{\int_{10p_{k,i}^{\text{uni}}}^{1} \binom{n_k}{n_{k,i}} p_{k,i}^{n_{k,i}} (1-p_{k,i})^{n_k - n_{k,i}} \cdot 1 dp_{k,i}}{\int_{0}^{10p_{k,i}^{\text{uni}}} \binom{n_k}{n_{k,i}} p_{k,i}^{n_{k,i}} (1-p_{k,i})^{n_k - n_{k,i}} \cdot 1 dp_{k,i}} \\
&= \frac{\int_{10p_{k,i}^{\text{uni}}}^{1} p_{k,i}^{n_{k,i}} (1-p_{k,i})^{n_k - n_{k,i}} dp_{k,i}}{\int_{0}^{10p_{k,i}^{\text{uni}}} p_{k,i}^{n_{k,i}} (1-p_{k,i})^{n_k - n_{k,i}} dp_{k,i}} \\
&= \frac{B(n_{k,i}+1, n_k - n_{k,i}+1) - B(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)}{B(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)} \\
&= \frac{1 - \frac{B(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)}{B(n_{k,i}+1, n_k - n_{k,i}+1)}}{\frac{B(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)}{B(n_{k,i}+1, n_k - n_{k,i}+1)}} \\
&= \frac{1 - I_{10p_{k,i}^{\text{uni}}}(n_{k,i}+1, n_k - n_{k,i}+1)}{I_{10p_{k,i}^{\text{uni}}}(n_{k,i}+1, n_k - n_{k,i}+1)} \\
&= \frac{1 - F(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)}{F(10p_{k,i}^{\text{uni}};\ n_{k,i}+1, n_k - n_{k,i}+1)}
\end{aligned}
\tag{8.14}
$$

8.2.4.4 Results and discussion

The likelihood and posterior of the null hypothesis are shown in Figure 8.4. The results of the frequentist and Bayesian hypothesis test are the same. It turns out that $S_{2,1}$ is a 2-edge contingency motif, $S_{3,1}$, $S_{3,3}$ and $S_{3,4}$ are 3-edge contingency motifs, and $S_{4,1}$, $S_{4,2}$, $S_{4,3}$ and $S_{4,4}$ are 4-edge contingency motifs. p_{H_0} is probability of a pattern not being a motif. The lower the probability, the more unlikely that a pattern is not a motif. It shows that stars ($S_{2,1}$, $S_{3,1}$, and $S_{4,1}$) are the most significant motifs.

In the graph representation of the power system, only lines are considered as elements of multiple contingencies, and node outages (mainly generators and transformer outages) are not considered. However, line outages usually accompany node outages, and they have equivalent effect on the power flow model. As for physical elements in power systems, multiple contingencies involve primary devices (generators, lines, transformers, compensators, circuit breakers, bus-bar sections) and secondary devices (protections and telecommunication equipment). Outages of these devices would result in multiple contingencies of transmission lines. NERC standard [103]

	$S_{k,i}$	p-value	p_{H_0}	motif
	$S_{2,1}$	0.	0.	True
	$S_{2,2}$	1	1.	False
	$S_{3,1}$	0.	0.	True
	$S_{3,2}$	1	1.	False
	$S_{3,3}$	0.	$3.\times10^{-26}$	True
	$S_{3,4}$	0.	$4.\times10^{-12}$	True
	$S_{3,5}$	1	1.	False
	$S_{4,1}$	0.	$1.\times10^{-39}$	True
	$S_{4,2}$	0.	$2.\times10^{-8}$	True
	$S_{4,3}$	0.	$2.\times10^{-18}$	True
	$S_{4,4}$	0.	$1.\times10^{-9}$	True
others	$S_{4,*}$	1	1.	False

Figure 8.4: Contingency motifs.

suggests seven categories of contingencies related to various devices, and they can be further grouped into three types: $N-1$, $N-1-1$, $N-2$ and $N-k$. For example, category P3 is single-phase short circuit to ground of a bus-bar section. If the bus-bar section connects k lines, then a $N-k$ contingency occurs.

Multiple contingencies can be divided into dependent contingencies and independent contingencies. Dependent multiple contingencies are closely related to bus configurations and protective relay design . It needs a lot of effort to build a detailed power system model including the relays [104]. Scheduled maintenance, planned and forced outages change the topology of the

power network, and hidden failures in protective relay system are inevitable. There are also common environment multiple contingencies that are caused by extreme weather or other external factors. Network motifs are used successfully in biology as a technique to identify functionally important local structures in gene transcription regulation networks; we show that contingency motifs in power systems can assist engineer by indicating multiple contingencies with high probabilities. The existence of the motif is a result of complex physical dependencies in power systems. Given motifs in a power network, engineers with field knowledge can better identify vulnerable sites in the network.

Two-edge stars could be two transmission lines connected on the same bus-bar section faulted simultaneously by coincidence, primary protection fails and zone 2 protection is activated, a bus-bar section connecting two transmission lines faults, a circuit breaker or a tie break stuck in a Breaker and Half substation, hidden failures of relay systems, etc. In the first cause, the two line outages are independent because one line outage does not necessarily cause the other line outage; while for the rest of the causes, the two line outages are dependent via physical or engineered structure. On the other hand, $S_{2,1}$ exist in common mode contingencies. A common mode contingency is a multiple contingency caused by a single event where outages are not consequences of each other [105]. For example, a single lightning stroke can cause two line outages on a common tower. Thus, $S_{2,1}$ as a motif usually reflects some inherent dependence of two lines.

The causes for three-edge and four-edge stars could include faults of transmission lines connected on the same bus-bar section, faults of bus-bar sections, transformers outages, and breaker stuck, etc. $S_{3,3}$ is composed of three lines in a row. A possible cause is that these three lines are in a protection control group. $S_{3,4}$ is a triangle, which is a special local structure in the power network that is limited in number.

The exactly inherent physical dependence for a specific motif is not clear without detailed knowledge of a system, but their existence implies a direction to study multiple contingency mechanisms in detail, such as protection groups, remedial action schemes, or common mode contingencies.

8.3 Diameter distribution of multiple contingencies

Multiple contingencies not only form connected subgraphs but also disconnected subgraphs. This section inspects the spatial extent of connected and disconnected subgraphs of multiple contingencies.

Multiple contingencies have different frequencies when they have different extent in the network. For example, in Figure 8.1 the subgraph with lines $1-3$ and $4-5$ has larger extent than the subgraph with lines $1-3$ and $1-6$. To quantify the extent of a subgraph, we define a diameter for the subgraph. The diameter of a subgraph is the longest distance between any two lines in the subgraph, in which the distance is the minimum number of nodes of a network path joining two lines.

The distance of any two lines l_i, l_j is defined as [1, 106]:

$$d(l_i, l_j) = \text{minimum number of nodes of a network path joining } l_i \text{ to } l_j \tag{8.15}$$

and the diameter of a k-edge subgraph $s_{k,i} = (l_l, ..., l_m)$ with some pattern $S_{k,i}$ is:

$$d(s_{k,i} = (l_l, ..., l_m)) = \max_{i,j} d(l_i, l_j) \tag{8.16}$$

For example, the distance between line 1-2, 1-6 in Figure 8.1 is $d(1-2, 1-6) = 1$, $d(1-2, 4-6) = 2$, and the diameter of 3-edge subgraph $d(s_{3,3} = (1-2, 1-6, 4-6)) = 2$.

Figure 8.5 shows the histogram of the diameters of contingency subgraphs formed from outage data. It also shows the diameter distribution when we randomly select k-edge subgraphs under the uniform assumption. The diameter distribution of $N-k$ is not the same as that of randomly selected k-edge subgraphs. Therefore, the diameter distribution of outages is not a result of the power network topology.

We fit the diameters of contingency subgraphs into a Zipf distribution. Figure 8.5 shows that the PDF of the fitted Zipf distribution matches the empirical diameter distribution. Also, the Kolmogorov–Smirnov goodness-of-fit test has a p-value 0.99.

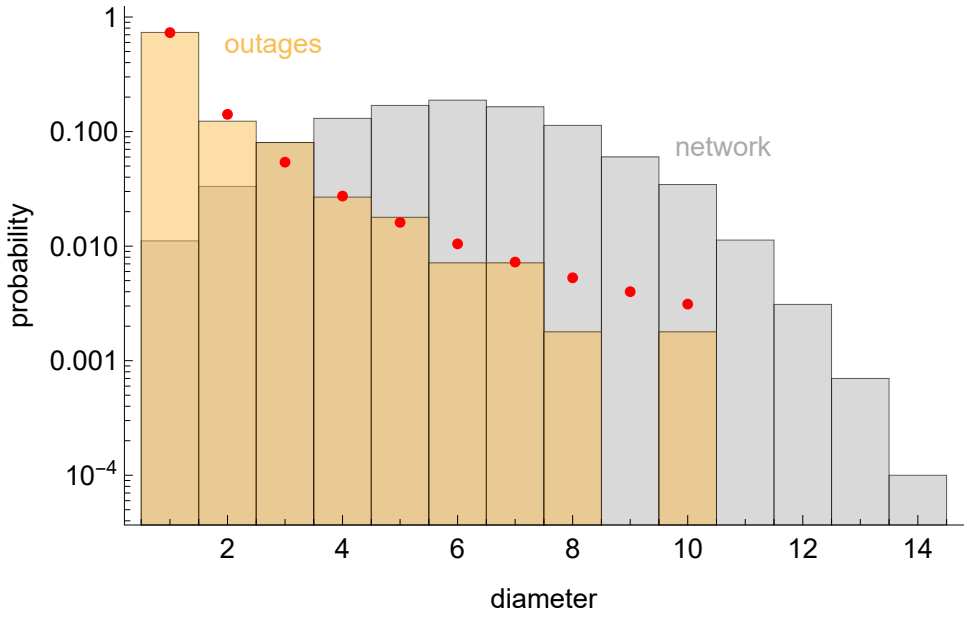


Figure 8.5: Histogram of diameters of subgraphs form by multiple contingencies (yellow bars), the fitted Zipf distribution (red dots), and histogram of diameters of random k-edge subgraphs drawn from the power network (gray bars).

The PDF of the fitted Zipf distribution is

$$P(d(s_{k,i} = (l_l, ..., l_m)) = x) = \frac{x^{-\rho}}{H(r, \rho)}, \qquad x = 1, ..., r \tag{8.17}$$

where $H(z) = \sum_{x=1}^{r} x^{-\rho}$ is the generalized harmonic number. r is the range of x because the diameter of the power network is limited. The maximum likelihood estimations [107] of parameters are $\widehat{r} = 10$ and $\widehat{\rho} = 2.37$.

8.4 Estimating probabilities of multiple contingencies

Since multiple contingencies occur much frequently in contingency motifs, and multiple contingencies with large diameters do occur. Thus, we can partition the whole contingency space according to different patterns and their diameters. The idea of the partition is illustrated in Figure 8.6. The ellipse represents the space of multiple contingency subgraphs, which include $N-2$, $N-3$, and $N-4$. According to different patterns, $N-k$ are further divided into groups $S_{k,i}$. Furthermore, disconnected $S_{k,i}$ are divided into subgroups according to their diameters.

Each cell represents a $S_{k,i}$ with a specific diameter d, and multiple contingencies $s_{k,i}$ in each cell are assumed to be uniformly distributed.

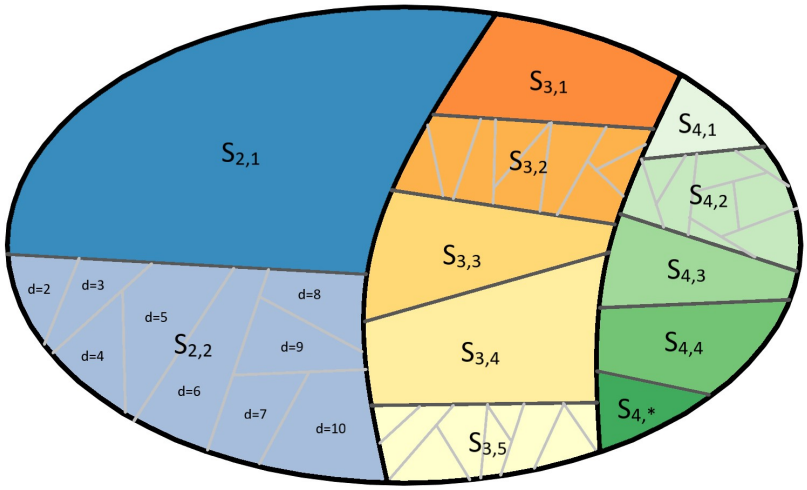


Figure 8.6: Partition of multiple contingency subgraphs. Each cell represents a pattern $S_{k,i}$ with a specific diameter d, and multiple contingencies $s_{k,i}$ in each cell are uniformly distributed.

We build a probabilistic model for estimating the probability of the multiple contingency $s_{k,i}$ with k lines $l_l, ..., l_m$ based on statistics of outage data. That is,

$$\begin{aligned}
&P\left(s_{k,i} = (l_l, ..., l_m)\right) \\
=&P\left(k, S_{k,i}, d, s_{k,i}\right) \\
=&P\left(k\right) P\left(S_{k,i}, d, s_{k,i}|k\right) \\
=&P\left(k\right) P\left(S_{k,i}|k\right) P\left(s_{k,i}, d|k, S_{k,i}\right) \\
=&P\left(k\right) P\left(S_{k,i}|k\right) P\left(d|k, S_{k,i}\right) P\left(s_{k,i}|k, S_{k,i}, d\right) \\
=&P\left(k\right) P\left(S_{k,i}|k\right) P\left(d|S_{k,i}\right) P\left(s_{k,i}|S_{k,i}, d\right) \qquad (8.18)
\end{aligned}$$

where $P\left(k\right)$ is the probability of the number of line outages, $P\left(S_{k,i}|k\right)$ is the probability of pattern $S_{k,i}$ given k line outages, $P\left(d|S_{k,i}\right)$ is the probability that pattern $S_{k,i}$ has diameter d, and $P\left(s_{k,i}|d, S_{k,i}\right)$ is the probability of a specific multiple contingency given its pattern and diameter.

8.4.1 Probability of the number of line outages

It is natural to estimate the probability $P(k)$ by

$$P(k) = \frac{n_k}{\sum_{k=2}^{4} n_k} \tag{8.19}$$

The distribution of k is shown in Table 8.1.

Table 8.1: Distribution of the number of line outages in a contingency.

k	2	3	4
$P(k)$	0.72	0.23	0.04

8.4.2 Probability of a pattern given the number of line outages

Then, we determine the probability $P(S_{k,i}|k)$, which is also $p_{k,i}$. The probability of patterns are estimated during the detection process. That is $P(S_{k,i}|k) = p_{k,i}^{\text{obs}}$ as shown in Table 8.2.

Table 8.2: Distribution of patterns $P(S_{k,i}|k)$

$S_{k,i}$	$S_{2,1}$	$S_{2,2}$	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	$S_{3,5}$	$S_{4,1}$	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$	$S_{4,*}$
$P(S_{k,i}\|k)$	0.809	0.191	0.583	0.244	0.141	0.024	0.008	0.391	0.217	0.217	0.087	0.087

8.4.3 Probability of the diameter of a contingency given its pattern

Section 8.3 actually gives the marginal distribution of the diameter. Here, we estimate the conditional distribution of the diameter given a pattern $S_{k,i}$ for the probabilistic model.

For a connected contingency subgraph $s_{k,i}$, the diameter provides no new information or negligible information. Connected $s_{k,i}$ for $k = 2, 3, 4$ except $s_{4,4}$ have a constant diameter. Diameters of $s_{2,1}$, $s_{3,1}$, $s_{3,4}$, and $s_{4,1}$ are all 1; diameters of $s_{3,3}$ and $s_{4,3}$ are both 2. Although the diameter of $s_{4,4}$ is 2 or 3, 98% of $s_{4,4}$ have diameter 3 in the BPA network. Therefore, the diameter distributions of different connected contingency patterns (for $k = 2, 3, 4$) can be modeled as degenerate distributions with probability 1 at a constant diameter. Since $S_{k,*}$ has a small probability but contains many different patterns, it is convenient to assume subgraphs in $S_{k,*}$ are uniformly distributed.

For a disconnected contingency subgraph $s_{k,i}$, the diameter is always greater than 1. For example, the minimum diameter of $s_{2,2}$ is 2. Given a disconnected pattern $S_{k,i}$, the diameters observed in outage data are distributed according to some distribution. Let $P\left(d\left(s_{k,i}=(l_l,...,l_m)\right)|S_{k,i}\right)$ (or $P\left(d|S_{k,i}\right)$ for simplicity) denote the conditional diameter distribution given a disconnected pattern $S_{k,i}$. It can be estimated from the outage data. However, due to limited outage data, we assume diameter distributions $P(d|S_{k,i})$ for disconnected $S_{k,i}$ are the same and we estimate it by lumping all disconnected contingency subgraphs together. That is, $P\left(d|S_{k,i}\right) = P(d|$ all disconnected $S_{k,i})$, and it is represented by the empirical distribution estimated from the outage data.

In summary, the diameter distribution conditional on pattern $S_{k,i}$ is computed by

$$P\left(d|S_{k,i}\right) = \begin{cases} 1 & S_{k,i} \in C \\ P(d|C^c) & S_{k,i} \in C^c \end{cases} \tag{8.20}$$

where C is the union of connected patterns and $S_{4,*}$, $C = S_{2,1} \cup S_{3,1} \cup S_{3,3} \cup S_{3,4} \cup S_{4,1} \cup S_{4,3} \cup S_{4,4} \cup S_{4,*}$; and C^c is the complement of C, the union of disconnected patterns, $C^c = S_{2,2} \cup S_{3,2} \cup S_{3,5} \cup S_{4,2}$.

8.4.4 Probability of a contingency given its pattern and diameter

Finally, assume that subgraphs of a pattern $S_{k,i}$ with diameter d are uniformly distributed. That is, $P\left(s_{k,i}|S_{k,i},d\right)$ is a discrete uniform distribution as shown in (8.21). $|S^d_{k,i}|$ denotes the number of subgraphs in $S_{k,i}$ with diameter d. It is approximated by sampling a large number of $s_{k,i}$ and computing their diameters, which is shown in Table 8.3.

$$P\left(s_{k,i}|S_{k,i},d)\right) = \frac{1}{|S^d_{k,i}|} \tag{8.21}$$

Thus, given a specific multiple contingency $s_{k,i} = (l_1,...,l_m)$, we can estimate its probability through formula (8.18) by substituting values in Table 8.1 and 8.2, and computing probabilities by (8.20) and (8.21).

Table 8.3: Number of distinct subgraphs with different diameters in $S_{k,i}$

d	$\lvert S^d_{2,2}\rvert$	$\lvert S^d_{3,2}\rvert$	$\lvert S^d_{3,5}\rvert$	$\lvert S^d_{4,2}\rvert$
2	6592	47035	10129	110725
3	14330	133813	333798	232684
4	22607	203860	1442182	414143
5	27069	231926	3395950	522119
6	25360	201685	5041582	475201
7	18777	133339	5069585	318717
8	11653	73988	3781098	160323
9	6283	36730	2327380	75294
10	2897	15178	1213277	34563

8.5 Contingency selection scheme

Contingency selection aims to select a list of credible contingencies sorted by their probability and/or impact for detailed analysis. There is no well-established scheme for multiple contingency selection. The practice is to select $N-k$ according to expert knowledge or operator's experience. We aim to augment the engineering judgement with an objective method driven by observed data.

By analyzing outage data, we find that multiple contingencies occur more frequently in contingency motifs and that close lines tend to fail simultaneously more frequently than far away lines. Based on these, we propose a probabilistic model to estimate the probability of a multiple contingency. However, it is not practical to compute probabilities for all multiple contingencies because of the enormous number of possible contingencies. Therefore, we propose a systematic scheme to select multiple contingencies by sampling according to the probabilities of multiple contingencies. Formula (8.18) implies the systematic scheme of sampling a multiple contingency. This scheme is comprised of four steps: (1) sample k according to $P(k)$; (2) sample $S_{k,i}$ according to $P(S_{k,i}|k)$; (3) sample diameter d according to $P(d|S_{k,i})$; (4) sample a $s_{k,i}$ uniformly from all subgraphs in pattern $S_{k,i}$ with diameter d. The first three steps are straightforward as the random variables are represented by integers. However, the fourth step is tricky because there is not an effective way to find all subgraphs $s_{k,i}$ with pattern $S_{k,i}$ and diameter d and randomly draw one from these subgraphs. Instead, we sample a $s_{k,i}$ by drawing lines sequentially. For

$N-2$, we first draw a line randomly; then we find all lines that are distance d from the first line; finally, we randomly draw a second line to form a $N-2$ together with the first line. For $N-3$, we draw the first line randomly and draw the second line that is distance d from the first line, as we do for $N-2$; then, we randomly draw the third line from lines that have a distance not greater than d from either the first or the second line and forms the desired pattern together with the previous two lines. For connected $N-4$, the distance is fixed. We first sample a 3-edge subgraph, which is a subsubgraph of the desired pattern, and then sample the last line randomly from lines that can form the desired pattern. For $s_{4,2}$, we first draw a 3-edge star $s_{3,1}$ and then draw a line that is maximum distance d from any of the three lines in the star. For $s_{4,*}$, we randomly sample 4 lines; if they do not form $s_{4,*}$, we sample again until they form a $s_{4,*}$.

8.6 Case study

We use 19 years of historical outage data recorded by a utility and publicly available at [7] to test the contingency selection scheme. The network topology is deduced from the data using [1]. The first 14 years of data is used to estimate the probabilistic model and then sample a contingency list; then, we evaluate the percentage of contingencies in the last 5 years that are covered by the contingency list.

To evaluate the performance of the proposed contingency selection scheme, we compare the contingency list to a same size list produced by a random scheme based on the uniform assumption. The random scheme randomly selects multiple contingencies with equal probability. Specifically, it samples the number of line outages k according to $P(k)$ and then randomly samples a $N-k$ contingency by draw k lines randomly from all lines.

Using the systematic scheme, we draw 10,000 contingency samples that form a contingency list. Contingencies with high probabilities generally appear in the top of the list, but we can compute their probabilities and sort contingencies in descending order of the probability. Table 8.4 shows the top three contingencies and their probabilities. They have the same probability

because they all belong to $S_{2,1}$. We also show the probabilities of all contingencies in the list in Figure 8.7.

Table 8.4: The three most probable contingencies.

ID	$S_{k,i}$	outaged lines	probability
1	$S_{2,1}$	$\{348-365, 348-385\}$ [1]	0.0003
2	$S_{2,1}$	$\{342-378, 350-378\}$	0.0003
3	$S_{2,1}$	$\{340-353, 340-354\}$	0.0003

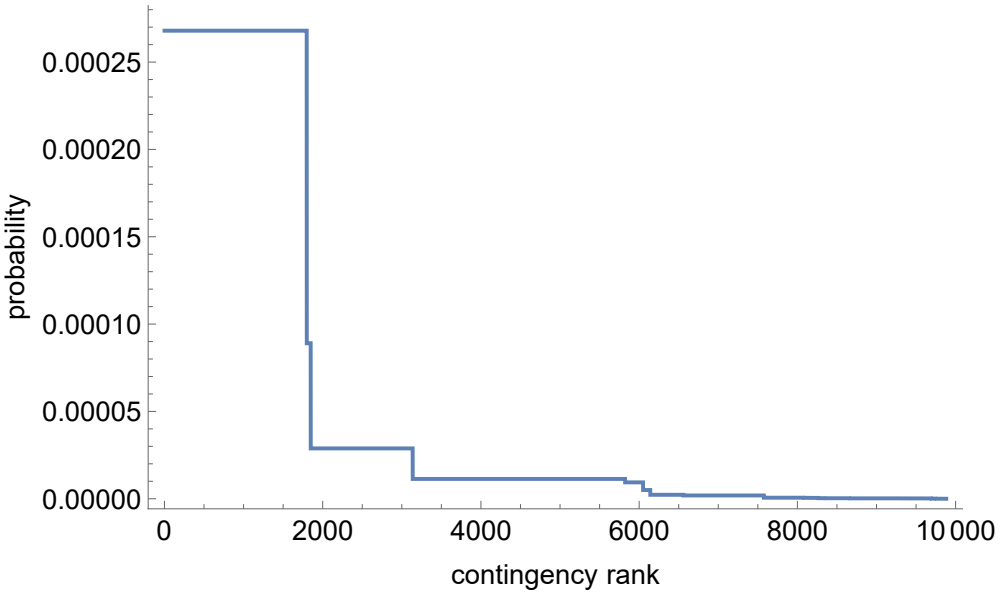


Figure 8.7: Probabilities of contingencies in descending order.

Let $M(r)$ be the percentage of actual contingencies in test data that are predicted by the contingency list that contains r contingencies. Figure 8.8 shows how $M(r)$ increases as r increases. It is obvious that the systematic scheme is much more efficient than the random scheme. Since the systematic scheme is a sampling method, we draw ten groups of samples with size r to estimate the mean and standard deviation of $M(r)$. For the systematic scheme, the average $M(10000)$ is 82%, and the standard deviation is 2%; for the random scheme, the average of $M(10000)$ is only 10%.

[1] $\{348-365, 348-385\}$ stands for a $N-2$. $348-365$ is a line with bus 348 and bus 365.

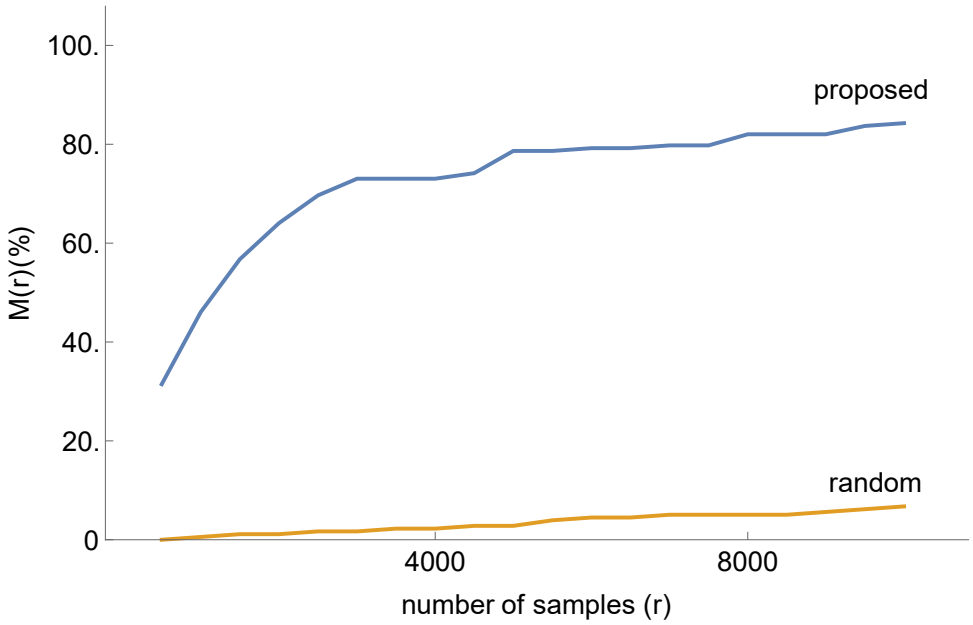


Figure 8.8: Percentage $M(r)$ of contingencies in test data that is predicted in the contingency list with r samples for the proposed systematic scheme (blue) and the random scheme (orange). The blue line does not start at 0 because we compute $M(r)$ for $r = 500, 1000, 1500, ...$

There is a correspondence between Figure 8.7 and Figure 8.8. Even though we draw samples randomly, the contingencies with high probabilities are more likely to be drawn at the early stage. Therefore, the curve in Figure 8.8 has a high derivative at the beginning and then the curve becomes nearly flat. The first 3000 contingencies covers about 75% of contingencies in the test data. On the other hand, the first 3000 random contingencies covers only 4%.

8.7 Conclusion and discussion

This study analyzes the spatial patterns of multiple contingencies using historical outage data. Some patterns occur significantly more frequently than others in outage data, and we define them as contingency motifs. The existence of these motifs in the power network reveals the physical dependencies of the power network. Contingency motifs indicate the basic functional groups as a result of bus configurations and protection systems in the power grid. They point out an objective way of selecting contingencies for detailed analysis based on historical outages.

Some patterns are disconnected subgraphs. The network diameters of contingency subgraphs follow a Zipf distribution. The Zipf distribution has a heavy tail, which implies multiple contingencies that contains far away components do occur.

Based on the above two findings, this study formulates a probabilistic model to estimate probabilities of multiple contingencies. The probabilistic model implies a systematic scheme to sample contingencies to form a contingency list with the most likely contingencies appearing first. The study tests the effectiveness of the systematic scheme by training the probabilistic model on the first 14-year outage data and testing it on the last 5-year outage data. It turns out that the systematic scheme is much more efficient than the random scheme: 10,000 sampled contingencies produced by the systematic scheme contain 82% of the multiple outages in test data, while 10,000 sampled contingencies produced by the random scheme only contain 10%. As the systematic scheme is a sampling method, more likely contingencies are sampled first. For example, the first 3,000 samples cover 75% of contingencies, in contrast to 6% for the random scheme.

We can sample multiple contingencies according to the systematic scheme and estimate the probabilities of these contingencies. This approach based on observed contingencies is complementary to methods of determining multiple contingencies by engineering knowledge of specific mechanisms. One advantage of contingency motifs is that their probability is estimated from routinely collected outage data. It is often difficult to get data on the probability of specific engineering mechanisms for multiple contingencies. Then, for industry, detailed analysis can be conducted to evaluate the severity/impact of different contingencies. As the risk is the product of the probability and severity, this would lead to a risk-based contingency analysis. The systematic scheme is a sampling method that can be flexibly used in planning or on-line analysis. Moreover, instead of selecting initial outages randomly, this systematic scheme gives a practical method of selecting initial outages for cascading study and resilience evaluation.

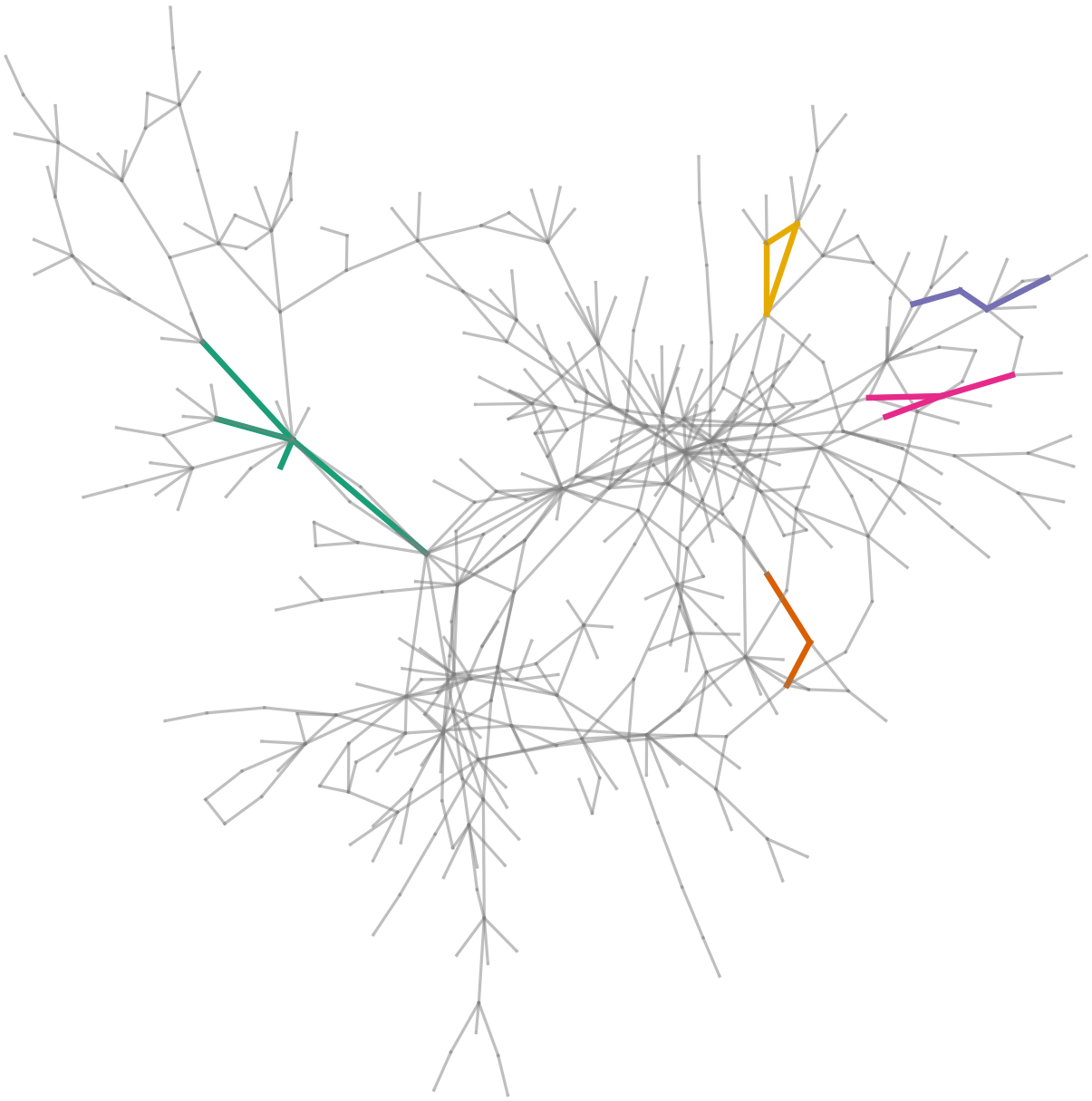


Figure 8.9: Power network of a utility derived from outage data [1]. Highlighted subgraphs with different colors are five multiple contingencies.

CHAPTER 9. CONCLUSION AND CONTRIBUTION

9.1 Conclusions

Outage data are routinely collected by utilities. Data-driven methods are needed to extract useful information from real outage data.

To study the cascading phenomena, line outages are grouped into cascades and generations within a cascade. Then, a Markovian influence graph is formed from the cascade data. The Markovian influence graph describes the probabilistic interactions between generations. It is a rigorous Markov chain and reproduces the distribution of cascade sizes in the cascade data. The quasi-stationary distribution of the Markov chain gives the probabilities of lines involved in long cascades. This distribution indicates critical lines in the propagation of cascading outages.

The Markovian influence graph driven by historical data can be used to simulate cascades, assuming some initial outages. Sampling of the Markovian influence graph is easily improved to allow sample size uniformly distributed across all sizes of cascades. This produces more samples than straightforward sampling from a Markov chain. Therefore, we can better estimate the risk of large cascades that are rare but significant. The cascade samples are comprised of specific line outages. The probability distribution of the load shed is estimated by a model-based simulation OPA. We estimate the conditional distribution of load shed given the number of line outages, and then the load shed of cascades is the weighted sum of the conditional distribution of load shed. There are strengths and limitations of the Markovian influence graph based on historical data versus model-based simulation. The two simulation approaches are complementary.

The Markovian influence graph is validated by two tests. The first test shows that the influence graph captures the mitigation effect produced by different upgrading measures. The mitigation is simply modeled by adjusting the transition probabilities to critical lines. However, the mitigation effect is dependent on the modeling of cascades and the measurement of cascade

sizes. The second test shows that the assumption that line outages only depend on the preceding line outage is reasonable. The result of the Markovian influence graph and another Markov model, in which the current line outages depend on all previous line outages, is consistent.

One essential and fundamental reliability is calculating the individual transmission line outage rates. As the line outages are infrequent, the outage data is limited. To better estimate the individual line outage rates, the study proposes a Bayesian hierarchical model that leverages line dependencies. This Bayesian hierarchical model produces estimates that have a lower standard deviation than simply dividing the number of outages by the time period. It estimates the distribution of individual line outage rates, which is an advantage compared to methods that only produce point estimates. The Bayesian estimates of individual line outage rates benefit the reliability calculations. This thesis demonstrates this by three applications: detecting lines with reduced reliability, estimates outage rates for specific causes, and testing the effect on the system unavailability calculation.

Contingency selection is one of the key functions of power system operation and planning. Analysis of the historical outage data shows that multiple contingencies occur frequently in contingency motifs of the power network, and that the diameter of contingency subgraphs follows a Zipf distribution. Based on these two findings, a probabilistic model of multiple contingencies and the corresponding contingency selection scheme are proposed. The systematic sampling scheme is more efficient than random sampling contingencies.

9.2 Contributions

The Markovian influence graph

- uses real data observed and routinely collected by utilities.
- obtains a clearly defined influence graph that solves the problem of multiple simultaneous outages by using additional states with multiple outages. This generalized influence graph rigorously defines a Markov chain.
- mitigates the problems of limited cascading data with several new methods; in particular, it combines Bayesian methods of estimation with elaborate methods of distinguishing and combining different events. This better estimates the transition matrices of the influence graph while matching the increasing cascade propagation and retaining possibilities of analysis.
- computes the probabilities of small, medium and large cascades, and these match the historical data statistics.
- makes innovative use of the bootstrap to estimate the standard deviation of the probabilities of small, medium and large cascades. This allows checking that the estimated probabilities of small, medium and large cascades are accurate enough to be useful.
- identifies critical lines most involved in large cascades directly from the Markov chain as the quasi-stationary distribution contains the probability of lines involved in large cascades.
- is validated on simulations for mitigation modeling and KMC model for the assumption that outages only depend on preceding outages.
- simulates cascade samples that encompass rare large cascades assuming some initial damages by extreme events.

The Bayesian hierarchical model

- estimates annual outage rates for individual transmission lines more accurately by leveraging partial similarities between lines, including proximity, length, and rated voltage, especially when the annual outage counts are low or the data is limited.
- has performance better than the conventional method of simply dividing the number of outages by the number of years observed, especially when the data is limited. The estimates have a lower standard deviation for given data, or the same standard deviation for less data. For one-year data, the standard deviation halves comparing to the conventional estimates.
- instead of pooling lines with one characteristic in common, gives a way to combine multiple partial similarities between lines.
- provides not only point estimates of line outage rates, but also the uncertainty.
- shows that line length and rated voltage correlate with line outage rate, but the correlation is not strong.
- works using a single standard line outage dataset routinely collected by transmission utilities worldwide.
- benefits for reliability evaluation. Applications are detecting lines with deteriorated reliability, estimating rates for specific causes, and computing more accurate system availability.

The analysis of spatial characteristics of initial outages

- finds that multiple contingencies occur much more frequently in contingency motifs of the power network.
- finds that the network diameter of multiple contingencies follows a Zipf distribution.
- helps to construct a probabilistic model to estimate the probability of multiple contingencies.

- produces a systematic scheme for multiple contingency selection, which is much more efficient than random sampling. Specifically, 10,000 samples cover 82% contingencies, in contrast to 10% using random scheme; furthermore, the first 3,000 samples cover 75% contingencies, in contrast to 6% using random scheme.

9.3 Future work

- The Markovian influence graph driven by historical data and model-based simulation are complementary. We can combine the two approaches through the influence graph by forming the Markovian influence graph from historical data and simulation data for the same system, then taking the weighted sum of the two transition matrices. Combining different data sources into the same influence graph would be particularly useful when extending the influence graph to interactions between the power grid and other critical infrastructures.

- Study a standard and easy-to-use cascading outages simulation environment considering uncertain renewable energy. Chapter 5 uses different cascade simulations to test the influence graph. We find that the modeling methods of cascading outages have different mitigation results of blackouts. This makes it difficult for researchers to test and compare different mitigation measures. Therefore, there is a need to make a standard and easy-to-use cascading simulation environment. Moreover, as more renewables are integrated into the power system, the uncertainty of high-penetration renewables should be considered in the cascading simulation.

- Study operational actions to mitigate cascading outages. Mitigation strategies are needed for both planning and operation. This work studied the mitigation in planning by upgrading critical components. [108] shows that cascading outages have a slow phase before blackouts, which justifies that mitigation actions can be taken during the propagation of cascading outages. Therefore, more research can be conducted on blackout mitigation strategies for operation.

9.4 Publications

Journal

1. K. Zhou, I. Dobson, Z. Wang, A. Roitershtein, A. P. Ghosh, "A Markovian influence graph formed from utility line outage data to mitigate large cascades," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3224-3235, Jul. 2020.

   This paper corresponds to the material in Chapter 3.

2. K. Zhou, J. R. Cruise, C. J. Dent, I. Dobson, L. Wehenkel, Z. Wang, A. Wilson, "Bayesian estimates of transmission line outage rates that consider line dependencies," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1095-1106, Mar. 2021.

   This paper corresponds to the material in Chapter 6.

Conference

1. K. Zhou, I. Dobson, P. D. H. Hines, Z. Wang, "Can an influence graph driven by outage data determine transmission line upgrades that mitigate cascading blackouts?," in *IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Boise, ID, USA, Jun. 2018.

   This paper is not included in this thesis as Chapter 3 describes a better solution to the problem.

2. K. Zhou, I. Dobson, Z. Wang, "Can the Markovian influence graph simulate cascading resilience from historical outage data?," in *IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Liege, Belgium, Aug. 2020.

   This paper corresponds to the material in Chapter 4.

3. K. Zhou, J. R. Cruise, C. J. Dent, I. Dobson, L. Wehenkel, Z. Wang, A. Wilson, "Applying Bayesian estimates of individual transmission line outage rates," in *IEEE International*

Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Liege, Belgium, Aug. 2020.

This paper corresponds to the material in Chapter 7.

BIBLIOGRAPHY

[1] I. Dobson *et al.*, "Obtaining statistics of cascading line outages spreading in an electric transmission network from standard utility data," *IEEE Trans. Power Syst.*, vol. 31, pp. 4831–4841, Nov. 2016.

[2] R. Baldick, B. Chowdhury, I. Dobson, *et al.*, "Initial review of methods for cascading failure analysis in electric power transmission systems," in *IEEE PES General Meeting*, (Pittsburgh, PA, USA), July 2008.

[3] P. Hines, J. Apt, and S. Talukdar, "Large blackouts in North America: Historical trends and policy implications," *Energy Policy*, vol. 37, pp. 5249–5259, Dec. 2009.

[4] B. A. Carreras, D. E. Newman, and I. Dobson, "North American blackout time series statistics and implications for blackout risk," *IEEE Trans. Power Syst.*, vol. 31, pp. 4406–4414, Nov. 2016.

[5] H. Haes Alhelou, M. E. Hamedani-Golshan, T. C. Njenda, and P. Siano, "A survey on power system blackout and cascading events: Research motivations and challenges," *Energies*, vol. 12, no. 4, p. 682, 2019.

[6] M. Papic, S. Ekisheva, and E. Cotilla-Sanchez, "A risk-based approach to assess the operational resilience of transmission grids," *Applied Sciences*, vol. 10, no. 14, p. 4761, 2020.

[7] "Bonneville Power Administration Transmission Services Operations & reliability." https://transmission.bpa.gov/Business/Operations/Outages. Accessed: 2019-01-21.

[8] Y. Bapin, S. Ekisheva, M. Papic, and V. Zarikas, "Outage data analysis of the overhead transmission lines in kazakhstan power system," in *2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pp. 1–6, 2020.

[9] J. Bialek *et al.*, "Benchmarking and validation of cascading failure analysis tools," *IEEE Trans. Power Syst.*, vol. 31, pp. 4887–4900, Nov. 2016.

[10] E. Ciapessoni *et al.*, "Benchmarking quasi-steady state cascading outage analysis methodologies," in *Prob. Methods Applied to Power Syst.*, (Boise, ID, USA), June 2018.

[11] Y. Yang, T. Nishikawa, and A. E. Motter, "Small vulnerable sets determine large network cascades in power grids," *Science*, vol. 358, no. 6365, 2017.

[12] C. Asavathiratham, S. Roy, B. Lesieutre, and G. Verghese, "The influence model," *IEEE Control Syst. Mag.*, vol. 21, pp. 52–64, Dec. 2001.

[13] M. Rahnamay-Naeini, "Designing cascade-resilient interdependent networks by optimum allocation of interdependencies," in *Int. Conf. Computing Networking and Communications*, (Kauai, HI, USA), Feb. 2016.

[14] P. Hines, I. Dobson, E. Cotilla-Sanchez, *et al.*, "'Dual graph' and 'random chemistry' methods for cascading failure analysis," in *Proc. 46th Hawaii Intl. Conf. System Sciences*, (Maui, HI, USA), Jan. 2013.

[15] J. Qi, K. Sun, and S. Mei, "An interaction model for simulation and mitigation of cascading failures," *IEEE Trans. Power Syst.*, vol. 30, pp. 804–819, Mar. 2015.

[16] P. Hines, I. Dobson, and P. Rezaei, "Cascading power outages propagate locally in an influence graph that is not the actual grid topology," *IEEE Trans. Power Syst.*, vol. 32, pp. 958–967, Mar. 2017.

[17] K. Zhou, I. Dobson, P. Hines, and Z. Wang, "Can an influence graph driven by outage data determine transmission line upgrades that mitigate cascading blackouts?," in *Prob. Methods Applied Power Syst.*, (Boise, ID, USA), June 2018.

[18] J. Qi, J. Wang, and K. Sun, "Efficient estimation of component interactions for cascading failure analysis by EM algorithm," *IEEE Trans. Power Syst.*, vol. 33, pp. 3153–3161, May 2018.

[19] X. Zhang, F. Liu, R. Yao, *et al.*, "Identification of key transmission lines in power grid using modified k-core decomposition," in *Proc. 3rd Int. Conf. Electric Power and Energy Conversion Systems*, (Istanbul, Turkey), Oct. 2013.

[20] H. M. Merrill and J. W. Feltes, "Cascading blackouts: Stress, vulnerability, and criticality," *preprint*, 2016.

[21] Z. Ma, C. Shen, F. Liu, and S. Mei, "Fast screening of vulnerable transmission lines in power grids: A pagerank-based approach," *IEEE Trans. Smart Grid*, vol. 10, pp. 1982–1991, Mar. 2019.

[22] Y. Yang, T. Nishikawa, and A. E. Motter, "Vulnerability and cosusceptibility determine the size of network cascades," *Phys. Rev. Lett.*, vol. 118, no. 4, p. 048301, 2017.

[23] B. A. Carreras, D. E. Newman, and I. Dobson, "Determining the vulnerabilities of the power transmission system," in *Proc. 45th Hawaii Int. Conf. System Sciences*, (Maui, HI, USA), pp. 2044–2053, Jan. 2012.

[24] U. Nakarmi, M. Rahnamay-Naeini, and H. Khamfroush, "Critical component analysis in cascading failures for power grids using community structures in interaction graphs," *IEEE Trans. Netw. Sci. Eng.*, 2019.

[25] W. Ju, K. Sun, and J. Qi, "Multi-layer interaction graph for analysis and mitigation of cascading outages," *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, vol. 7, pp. 239–249, June 2017.

[26] C. Chen, W. Ju, K. Sun, and S. Ma, "Mitigation of cascading outages using a dynamic interaction graph-based optimal power flow model," *IEEE Access*, vol. 7, pp. 168637–168648, 2019.

[27] X. Wei, J. Zhao, T. Huang, and E. Bompard, "A novel cascading faults graph based transmission network vulnerability assessment method," *IEEE Trans. Power Syst.*, vol. 33, pp. 2995–3000, May 2018.

[28] K. Sun, Y. Hou, W. Sun, and J. Qi, *Power System Control Under Cascading Failures: Understanding, Mitigation, and System Restoration.* Wiley-IEEE Press, 2019.

[29] A. Wang, Y. Luo, G. Tu, *et al.*, "Vulnerability assessment scheme for power system transmission networks based on the fault chain theory," *IEEE Trans. Power Syst.*, vol. 26, pp. 442–450, Feb. 2011.

[30] C. Luo, J. Yang, Y. Sun, *et al.*, "Identify critical branches with cascading failure chain statistics and hypertext-induced topic search algorithm," in *IEEE PES General Meeting*, (Chicago, IL, USA), 2017.

[31] L. Li, H. Wu, and Y. Song, "Temporal difference learning based critical component identifying method with cascading failure data in power systems," in *IEEE PES General Meeting*, (Portland, OR, USA), Aug. 2018.

[32] L. Li, H. Wu, Y. Song, and Y. Liu, "A state-failure-network method to identify critical components in power systems," *Electric Power Systems Research*, vol. 181, p. 106192, 2020.

[33] Z. Wang, A. Scaglione, and R. J. Thomas, "A Markov-transition model for cascading failures in power grids," in *Proc. 45th Hawaii Int. Conf. System Sciences*, (Maui, HI, USA), pp. 2115–2124, Jan. 2012.

[34] M. Rahnamay-Naeini, Z. Wang, N. Ghani, *et al.*, "Stochastic analysis of cascading-failure dynamics in power grids," *IEEE Trans. Power Syst.*, vol. 29, pp. 1767–1779, July 2014.

[35] M. Rahnamay-Naeini and M. M. Hayat, "Cascading failures in interdependent infrastructures: An interdependent Markov-chain approach," *IEEE Trans. Smart Grid*, vol. 7, pp. 1997–2006, July 2016.

[36] Z. Wang, M. Rahnamay-Naeini, J. M. Abreu, *et al.*, "Impacts of operators' behavior on reliability of power grids during cascading failures," *IEEE Trans. Power Syst.*, vol. 33, pp. 6013–6024, Nov. 2018.

[37] U. Nakarmi, M. R. Naeini, M. J. Hossain, and M. A. Hasnat, "Interaction graphs for cascading failure analysis in power grids: A survey," *Energies*, vol. 13, no. 9, p. 2219, 2020.

[38] M. Vaiman *et al.*, "Risk assessment of cascading outages: methodologies and challenges," *IEEE Trans. Power Syst.*, vol. 27, pp. 631–641, May 2012.

[39] K. Zhou, I. Dobson, Z. Wang, A. Roitershtein, and A. P. Ghosh, "A Markovian influence graph formed from utility line outage data to mitigate large cascades," *IEEE Transactions on Power Systems*, vol. 35, no. 4, pp. 3224–3235, 2020.

[40] J. Qi, "Utility outage data driven interaction networks for cascading failure analysis and mitigation," *IEEE Transactions on Power Systems*, vol. 36, no. 2, pp. 1409–1418, 2020.

[41] N. Romero, L. K. Nozick, I. Dobson, N. Xu, and D. A. Jones, "Seismic retrofit for electric power systems," *Earthquake Spectra*, vol. 31, pp. 1157–1176, May 2015.

[42] M. R. Kelly-Gorham, P. Hines, and I. Dobson, "Using historical utility outage data to compute overall transmission grid resilience," in *Mod. Electr. Power Syst. Conf.*, (Wrocław Poland), September 2019.

[43] M. R. Kelly-Gorham, P. D. Hines, K. Zhou, and I. Dobson, "Using utility outage statistics to quantify improvements in bulk power system resilience," *Electric Power Systems Research*, vol. 189, p. 106676, 2020.

[44] B. P. Carlin and T. A. Louis, *Bayesian methods for data analysis*. Boca Raton, FL, USA: CRC Press, 2008.

[45] A. Gelman et al., *Bayesian data analysis*. Boca Raton, FL, USA: CRC Press, 2013.

[46] M. Omlin and P. Reichert, "A comparison of techniques for the estimation of model prediction uncertainty," *Ecological modelling*, vol. 115, pp. 45–59, Feb. 1999.

[47] D. Stegmueller, "How many countries for multilevel modeling? a comparison of frequentist and bayesian approaches," *American Journal of Political Science*, vol. 57, pp. 748–761, July 2013.

[48] H. Li, L. A. Treinish, and J. R. M. Hosking, "A statistical model for risk management of electric outage forecasts," *IBM J. Res. Dev.*, vol. 54, pp. 8:1–8:11, May 2010.

[49] T. Iešmantas and R. Alzbutas, "Bayesian spatial reliability model for power transmission network lines," *Electr. Power Syst. Res.*, vol. 173, pp. 214–219, 2019.

[50] A. Moradkhani *et al.*, "Failure rate estimation of overhead electric distribution lines considering data deficiency and population variability," *Int. Trans. Electr. Energ. Syst.*, vol. 25, pp. 1452–1465, Apr. 2015.

[51] Y. Zhou, A. Pahwa, and S. Yang, "Modeling weather-related failures of overhead distribution lines," *IEEE Trans. Power Syst.*, vol. 21, pp. 1683–1690, Nov. 2006.

[52] M. Yang *et al.*, "Interval estimation for conditional failure rates of transmission lines with limited samples," *IEEE Trans. Smart Grid*, vol. 9, pp. 2752–2763, Jul. 2018.

[53] L. N. Dunn, I. Kavvada, M. D. Badoual, and S. J. Moura, "Bayesian hierarchical methods for modeling electrical grid component failures," *Electric Power Systems Research*, vol. 189, p. 106789, 2020.

[54] T. Dokic et al., "Spatially aware ensemble-based learning to predict weather-related outages in transmission," in *Proc. 52th Hawaii Intl. Conf. System Science*, (Maui, HI, USA), Jan. 2019.

[55] R. Yao and K. Sun, "Toward simulation and risk assessment of weather-related outages," *IEEE Trans. Smart Grid*, vol. 10, pp. 4391–4400, Jul. 2019.

[56] K. Alvehag and L. Soder, "A reliability model for distribution systems incorporating seasonal variations in severe weather," *IEEE Trans. Power Del.*, vol. 26, pp. 910–919, Apr 2011.

[57] Y. Wang *et al.*, "Evaluating weather influences on transmission line failure rate based on scarce fault records via a bi-layer clustering technique," *IET Gener. Transm. Distrib.*, vol. 13, pp. 5305–5312, Nov. 2019.

[58] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.

[59] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[60] L. Stone, D. Simberloff, and Y. Artzy-Randrup, "Network motifs and their origins," *PLoS computational biology*, vol. 15, no. 4, p. e1006749, 2019.

[61] Q. Chen, H. Ren, C. Sun, Z. Mi, and D. Watts, "Network motif as an indicator for cascading outages due to the decrease of connectivity," in *2017 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2017.

[62] A. K. Dey, Y. R. Gel, and H. V. Poor, "Motif-based analysis of power grid robustness under attacks," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1015–1019, IEEE, 2017.

[63] A. K. Dey, Y. R. Gel, and H. V. Poor, "What network motifs tell us about resilience and reliability of complex networks," *Proceedings of the National Academy of Sciences*, vol. 116, no. 39, pp. 19368–19373, 2019.

[64] A. Abedijaberi and J. Leopold, "Motif-level robustness analysis of power grids," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 276–283, IEEE, 2018.

[65] A. Tajer, S. M. Perlaza, and H. V. Poor, *Advanced Data Analytics for Power Systems.* Cambridge University Press, 2021.

[66] M. Papic, K. Bell, Y. Chen, *et al.*, "Survey of tools for risk assessment of cascading outages," in *IEEE PES General Meeting*, (Detroit, MI, USA), July 2011.

[67] M. Papic and I. Dobson, "Comparing a transmission planning study of cascading with historical line outage data," in *Prob. Methods Applied to Power Syst.*, (Beijing, China), Oct. 2016.

[68] I. Dobson, "Estimating the propagation and extent of cascading line outages from utility data with a branching process," *IEEE Trans. Power Syst.*, vol. 27, pp. 2146–2155, Nov. 2012.

[69] A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application (Vol. 1).* Cambridge, U.K.: Cambridge Univ. Press, 1997.

[70] I. Dobson, "Finding a Zipf distribution and cascading propagation metric in utility line outage data," *arXiv preprint arXiv:1808.08434 [physics.soc-ph]*, 2018.

[71] S. D. Guikema, "Formulating informative, data-based priors for failure probability estimation in reliability analysis," *Reliability Engineering & System Safety*, vol. 92, pp. 490–502, Apr. 2007.

[72] B. P. Carlin and T. A. Louis, *Bayesian methods for data analysis*. Boca Raton, FL, USA: CRC Press, 2008.

[73] E. T. Jaynes, "Bayesian methods: General background," in *Maximum Entropy and Bayesian Methods in Applied Statistics, Cambridge, U.K.*, Cambridge Univ. Press, 1986.

[74] S.-C. Fang, J. R. Rajasekera, and H.-S. J. Tsao, *Entropy optimization and mathematical programming*. Springer, 2012.

[75] K. Zhou, I. Dobson, Z. Wang, and A. L. Wilson, "Can the Markovian influence graph simulate cascading resilience from historical outage data?," in *Prob. Methods Applied Power Syst.*, (Liege, Belgium), Aug. 2020.

[76] B. A. Carreras *et al.*, "Complex dynamics of blackouts in power transmission systems," *Chaos*, vol. 14, no. 3, pp. 643–652, 2004.

[77] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos*, vol. 17, no. 2, p. 026103, 2007.

[78] H. Ren, I. Dobson, and B. A. Carreras, "Long-term effect of the N-1 criterion on cascading line outages in an evolving power transmission grid," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1217–1225, 2008.

[79] S. Mei, X. Zhang, and M. Cao, *Power grid complexity*. Beijing, China: Tsinghua University Press, Springer, 2011.

[80] B. A. Carreras *et al.*, "Validating OPA with WECC data," in *Proc. 46th Hawaii Int. Conf. System Sciences*, (Maui, HI, USA), Jan. 2013.

[81] B. A. Carreras *et al.*, "Validating the OPA cascading blackout model on a 19402 bus transmission network with both mesh and tree structures," in *Proc. 52th Hawaii Int. Conf. System Sciences*, (Maui, HI, USA), Jan. 2019.

[82] I. Dobson and D. E. Newman, "Cascading blackout overall structure and some implications for sampling and mitigation," *Int. J. Electr. Power Energy Syst.*, vol. 86, pp. 29–32, Mar. 2017.

[83] J. Kim, J. A. Bucklew, and I. Dobson, "Splitting method for speedy simulation of cascading blackouts," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 3010–3017, 2012.

[84] M. J. Eppstein and P. D. Hines, "A "random chemistry" algorithm for identifying collections of multiple contingencies that initiate cascading failure," *IEEE Transactions on Power Systems*, vol. 27, no. 3, pp. 1698–1705, 2012.

[85] B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman, "Complex dynamics of blackouts in power transmission systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 14, no. 3, pp. 643–652, 2004.

[86] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 17, no. 2, p. 026103, 2007.

[87] S. Mei, X. Zhang, and M. Cao, *Power grid complexity*. Springer Science & Business Media, 2011.

[88] S. Mei, X. Weng, A. Xue, *et al.*, "Blackout model based on OPF and its self-organized criticality," in *2006 Chinese Control Conference*, pp. 1673–1678, IEEE, 2006.

[89] C. D. Meyer, *Chapter 8 Perron-Frobenius Theory of Nonnegative Matrices, in Matrix analysis and applied linear algebra*. SIAM, 2000.

[90] J. Roth, D. A. Barajas-Solano, P. Stinis, J. Weare, and M. Anitescu, "A kinetic Monte Carlo approach for simulating cascading transmission line failure," *arXiv preprint arXiv:1912.08081*, 2019.

[91] K. Zhou, J. R. Cruise, C. J. Dent, I. Dobson, L. Wehenkel, Z. Wang, and A. L. Wilson, "Bayesian estimates of transmission line outage rates that consider line dependencies," *IEEE Trans. Power Syst.*, vol. 36, no. 2, pp. 1095–1106, 2020.

[92] A. Gelman, "Prior choice recommendations.." Github, https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. (accessed Nov. 19, 2019).

[93] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[94] M. H. Kutner *et al.*, *Applied linear statistical models*. Boston, USA: McGraw-Hill Irwin, 2005.

[95] A. Gelman, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[96] M. Betancourt, "A conceptual introduction to Hamiltonian Monte Carlo," *arXiv preprint arXiv:1701.02434*, 2017.

[97] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1593–1623, 2014.

[98] K. Zhou, J. R. Cruise, C. J. Dent, I. Dobson, L. Wehenkel, Z. Wang, and A. L. Wilson, "Applying Bayesian estimates of individual transmission line outage rates," in *Prob. Methods Applied Power Syst.*, (Liege, Belgium), Aug. 2020.

[99] "NOAA national centers for environmental information storm events database." https://www.ncdc.noaa.gov/stormevents.

[100] I. Dobson *et al.*, "Exploring cascading outages and weather via processing historic data," in *51st Hawaii Intl. Conf. Sys. Science*, Jan. 2018.

[101] S. Kancherla and I. Dobson, "Heavy-tailed transmission line restoration times observed in utility data," *IEEE Trans. Power Syst.*, vol. 33, pp. 1145–1147, Jan. 2018.

[102] S. Wernicke and F. Rasche, "Fanmod: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[103] "TPL-001-5-NERC." https://www.nerc.com/pa/Stand/Reliability%20Standards/TPL-001-5.pdf. Accessed: 2010-06-30.

[104] I. Dobson, A. Flueck, S. Aquiles-Perez, S. Abhyankar, and J. Qi, "Towards incorporating protection and uncertainty into cascading failure simulation and analysis," in *2018 IEEE international conference on probabilistic methods applied to power systems (PMAPS)*, pp. 1–5, IEEE, 2018.

[105] M. Papic, K. Awodele, R. Billinton, C. Dent, D. Eager, G. Hamoud, C. Jirutitijaroen, M. Kumbale, J. Mitra, N. Samaan, *et al.*, "Overview of common mode outages in power systems," in *2012 IEEE Power and Energy Society General Meeting*, pp. 1–8, IEEE, 2012.

[106] Wikipedia, "Distance (graph theory)." https://en.wikipedia.org/wiki/Distance_(graph_theory), Accessed: Dec., 2021.

[107] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 41, no. 2, pp. 255–258, 2004.

[108] M. Noebels, I. Dobson, and M. Panteli, "Observed acceleration of cascading outages," *IEEE Transactions on Power Systems*, 2021.

[109] J. N. Darroch and E. Seneta, "On quasi-stationary distributions in absorbing discrete-time finite Markov chains," *J. Appl. Probab.*, vol. 2, pp. 88–100, June 1965.

[110] E. V. Doorn and P. Pollett, "Quasi-stationary distributions for discrete-state models," *European J. Operat. Res.*, vol. 230, pp. 1–14, 2013.

[111] W. J. Stewart, *Introduction to the numerical solution of Markov chains*. Princeton, NJ, USA: Princeton Univ. Press, 1994.

[112] B. Lambert, *A student's guide to Bayesian statistics*. Sage, 2018.

[113] M. Betancourt and M. Girolami, "Hamiltonian Monte Carlo for hierarchical models," *Current trends in Bayesian methodology with applications*, vol. 79, no. 30, pp. 2–4, 2015.

[114] R. M. Neal *et al.*, "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, vol. 2, no. 11, p. 2, 2011.

[115] V. Roy, "Convergence diagnostics for Markov Chain Monte Carlo," *Annu. Rev. Stat. Appl.*, vol. 7, pp. 387–412, 2020.

APPENDIX A. Deriving the quasi-stationary distribution

The quasi-stationary distribution can be derived in a standard way [109,110]. Let $\boldsymbol{d}_k$ be a vector with entry $d_k[i]$ which is the probability that a cascade is in nonempty state s_i at generation k given that the cascade is propagating, that is

$$d_k[i] = \frac{\mathrm{P}[X_k = s_i]}{\mathrm{P}[X_k \neq s_0]} = \frac{\pi_k[i]}{1 - \pi_k[0]}, \qquad i = 1, ..., |\mathcal{S}|$$

Then the quasi-stationary distribution is $\boldsymbol{d}_\infty = \lim_{k\to\infty} \boldsymbol{d}_k$.

Diagonal entries of $\bar{\boldsymbol{Q}}_{1+}$ corresponding to $\bar{\boldsymbol{P}}_{1+}$ are all zero and all other entries are positive. According to the Perron-Frobenius theorem [111], $\bar{\boldsymbol{Q}}_{1+}$ has a unique maximum modulus eigenvalue μ, which is real, positive and simple with left eigenvector $\boldsymbol{v}'$. By normalizing $\boldsymbol{v}'$, we make $\boldsymbol{v}'$ a probability vector. We write $\boldsymbol{w}$ for the corresponding right eigenvector. Moreover, $0 < \mu < 1$ and μ is strictly greater than the modulus of the other eigenvalues of $\bar{\boldsymbol{Q}}_{1+}$. Suppose the cascade starts with probability distribution $\boldsymbol{\pi}_0$ (note that $\pi_0[0] = 0$). According to (3.5), the probability of being in state i at generation k is $\pi_k[i] = (\boldsymbol{\pi}_0\boldsymbol{P}_0\boldsymbol{P}_1...\boldsymbol{P}_{k-2}\boldsymbol{P}_{k-1})[i] = (\boldsymbol{\pi}_0\boldsymbol{P}^{(k-1)})[i]$. In particular, the probability that the cascade terminates by generation k is $\pi_k[0] = \boldsymbol{\pi}_0\boldsymbol{P}^{(k)}[0] = \boldsymbol{\pi}_0\boldsymbol{P}^{(k)}\boldsymbol{e}_0$. Then for $i = 1, ..., |\mathcal{S}|$,

$$d_{k+1}[i] = \frac{\pi_{k+1}[i]}{1 - \pi_{k+1}[0]} = \frac{(\boldsymbol{\pi}_0\boldsymbol{P}^{(k)})[i]}{1 - \boldsymbol{\pi}_0\boldsymbol{P}^{(k)}\boldsymbol{e}_0} = \frac{(\boldsymbol{\pi}_0\boldsymbol{P}^{(k)})[i]}{\boldsymbol{\pi}_0\boldsymbol{P}^{(k)}(\mathbf{1} - \boldsymbol{e}_0)}$$

The first row of $\boldsymbol{P}_k$ is always $[1 \quad 0 \quad ... \quad 0]$. Since $\pi_0[0] = 0$, let $\boldsymbol{\pi}_0 = [0 \quad \bar{\boldsymbol{\pi}}_0]$. Then $\boldsymbol{\pi}_0\boldsymbol{P}^{(k)}(\mathbf{1} - \boldsymbol{e}_0) = \bar{\boldsymbol{\pi}}_0\boldsymbol{Q}^{(k)}\mathbf{1}$ and $(\boldsymbol{\pi}_0\boldsymbol{P}^{(k)})[i] = (\bar{\boldsymbol{\pi}}_0\boldsymbol{Q}^{(k)})[i]$ for $i = 1, ..., |\mathcal{S}|$. And $\boldsymbol{Q}^{(k)} = \bar{\boldsymbol{Q}}_0\bar{\boldsymbol{Q}}_{1+}^{k-1}\prod_{m=0}^{k}(1 - \alpha_m)$, so that $\boldsymbol{d}_\infty = \lim_{k\to\infty} \boldsymbol{d}_{k+1}$ is

$$\begin{aligned}\boldsymbol{d}_\infty &= \lim_{k\to\infty} \frac{\bar{\boldsymbol{p}}_0\boldsymbol{Q}^{(k)}}{\bar{\boldsymbol{p}}_0\boldsymbol{Q}^{(k)}\mathbf{1}} = \lim_{k\to\infty} \frac{\bar{\boldsymbol{p}}_0\bar{\boldsymbol{Q}}_0\bar{\boldsymbol{Q}}_{1+}^{k-1}\prod_{m=0}^{k}(1 - \alpha_m)}{\bar{\boldsymbol{p}}_0\bar{\boldsymbol{Q}}_0\bar{\boldsymbol{Q}}_{1+}^{k-1}\prod_{m=0}^{k}(1 - \alpha_m)\mathbf{1}} \\ &= \frac{\bar{\boldsymbol{p}}_0\bar{\boldsymbol{Q}}_0\mu^{k-1}\boldsymbol{w}\boldsymbol{v}'}{\bar{\boldsymbol{p}}_0\bar{\boldsymbol{Q}}_0\mu^{k-1}\boldsymbol{w}\boldsymbol{v}'\mathbf{1}} = \boldsymbol{v}'\end{aligned}$$

where $\bar{\boldsymbol{Q}}^{(k-1)} \to \mu^{k-1}\boldsymbol{w}\boldsymbol{v}'$ as $k \to \infty$. Therefore, the dominant left eigenvector of $\bar{\boldsymbol{Q}}_{1+}$ is $\boldsymbol{d}_\infty$.

For our data, the top three eigenvalues

in modulus are $\mu = 0.502$ and $-0.136 \pm 0.122\,\mathrm{i}$ with corresponding moduli 0.502 and 0.381.

APPENDIX B. Why use varying transition matrices?

The propagation rate increases as the generation increases because the situation is worse as more outages occur. If the transition matrix is constant, however, the propagation rate converges to a constant fast. Therefore, we need to use a variant transition matrix to capture this characteristic.

Suppose $\boldsymbol{P}$ is a constant matrix. Then from (3.18), the propagation rate for generation k is

$$\rho_k = \frac{\boldsymbol{\pi}_0 \boldsymbol{P}^k (\mathbf{1} - \boldsymbol{e}_0)}{\boldsymbol{\pi}_0 \boldsymbol{P}^{k-1} (\mathbf{1} - \boldsymbol{e}_0)} \tag{B.1}$$

Since the model is an absorbing Markov chain, the stationary distribution is $\boldsymbol{e}_0' = [1 \quad 0 \quad ... \quad 0]$. As $k \to \infty$, $\boldsymbol{\pi}_0 \boldsymbol{P}^k \to \boldsymbol{e}_0' + \mu_2^k c_2 \boldsymbol{v}_2$, where μ_2 is the second largest eigenvalue of $\boldsymbol{P}$ and $\boldsymbol{v}_2$ is the corresponding left eigenvector, c_2 is a constant depending on $\boldsymbol{\pi}_0$. The convergence rate depends on the third largest eigenvalue μ_3. The gap between the true value and the limit is proportional to $\||\mu_3|^k\|$ as $\|\boldsymbol{\pi}_0 \boldsymbol{P}^k (\mathbf{1} - \boldsymbol{e}_0) - \mu_2^k c_2 \boldsymbol{v}_2 (\mathbf{1} - \boldsymbol{e}_0)\| = o(\||\mu_3|^k\|)$. Then,

$$\lim_{k \to \infty} \rho_k = \frac{\mu_2^k c_2 \boldsymbol{v}_2 (\mathbf{1} - \boldsymbol{e}_0)}{\mu_2^{k-1} c_2 \boldsymbol{v}_2 (\mathbf{1} - \boldsymbol{e}_0)} = \mu_2 \tag{B.2}$$

The four largest eigenvalues are: $1,\ 0.50,\ -0.36 + \mathrm{i}0.12,\ -0.36 - \mathrm{i}0.12$ ($|-0.36 - \mathrm{i}0.12| = 0.38$). After five generations, the gap to the limit is below 0.01. So the propagation rate is nearly constant after five generations. This is verified in the data by calculating the propagation rate for each generation.

Therefore, if the Markov chain has a constant transition matrix, ρ_k is not increasing. So we need variant $\boldsymbol{P}$ for different generations to have increasing generation propagation rates.

APPENDIX C. Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC) is a sophisticated sampling algorithm combining ideas from Markov chains, rejection sampling, differential geometry, and numerical integration of Hamiltonian dynamics. This appendix reproduces the HMC Algorithm 1 from [97], briefly outlines how the algorithm works differently than other Markov chain Monte Carlo (MCMC) methods, and then recommends both tutorial and advanced references to HMC. Some general familiarity with MCMC is assumed.

Algorithm 1 Hamiltonian Monte Carlo

Given $\theta^0, \epsilon, L, \mathcal{L}, M$:
for $m = 1$ to M **do**
  Sample $r^0 \sim \mathcal{N}(0, I)$.
  Set $\theta^m \leftarrow \theta^{m-1}, \tilde{\theta} \leftarrow \theta^{m-1}, \tilde{r} \leftarrow r^0$.
  for $i = 1$ to L **do**
    Set $\tilde{\theta}, \tilde{r} \leftarrow \text{Leapfrog}(\tilde{\theta}, \tilde{r}, \epsilon)$.
  end for
  With probability $\alpha = \min\{1, \frac{\exp\{\mathcal{L}(\tilde{\theta}) - 0.5\tilde{r}\cdot\tilde{r}\}}{\exp\{\mathcal{L}(\theta^{m-1}) - 0.5 r^0 \cdot r^0\}}\}$,
    set $\theta^m \leftarrow \tilde{\theta}, r^m \leftarrow -\tilde{r}$.
end for
function Leapfrog(θ, r, ϵ)
$\tilde{r} \leftarrow r + (\epsilon/2)\nabla_\theta \mathcal{L}(\theta)$.
$\tilde{\theta} \leftarrow \theta + \epsilon\tilde{r}$.
$\tilde{r} \leftarrow r + (\epsilon/2)\nabla_\theta \mathcal{L}(\tilde{\theta})$.
return $\tilde{\theta}, \tilde{r}$

HMC has similar overall form as other Metropolis-Hastings Monte Carlo methods in that it proposes and probabilistically accepts successive samples of parameters to sample effectively from the posterior probability density. The successive samples are transitions in an ergodic Markov chain designed so that its final steady state distribution is the posterior probability density. However, HMC samples differently than other methods in an enlarged space. In the notation of Algorithm 1, the parameter vector θ of "position" variables is augmented with a vector of

"momentum" variables r to form an enlarged space of twice the dimension in which the successive samples are taken. The enlarged space enables Hamiltonian dynamics, where the "potential energy" $\mathcal{L}$ is the negative logarithm of the joint pdf of θ, and the "kinetic energy" is $\frac{1}{2}r \cdot r$.

Suppose the sampler is at (θ^0, r^0) in Algorithm 1. To propose a new sample at $(\tilde{\theta}, \tilde{r})$, the initial momentum r^0 is sampled from a Gaussian distribution, and then the Hamiltonian dynamics is integrated for L integration steps with integration step size ϵ. A symplectic leap-frog integrator that interleaves integration steps is used in order to preserve the Hamiltonian structure. Then the proposed sample is probabilistically accepted or rejected in a way similar to the Metropolis algorithm. Hoffman [97] proposed the No-U-Turn Sampler to avoid hand tuning the parameters L and ϵ controlling the integration.

To understand why HMC works, we refer readers to the approachable and intuitive expositions in [96] and [112, Cha.15] for expert explanations of the algorithm and to [97, 113–115] for further technical analysis. In particular, Betancourt discusses how HMC is "uniquely suited to the high-dimensional problems of applied interest." [96] and how HMC can tackle the correlations induced by hierarchical models [113]. The No-U-Turn Sampler has at least the same efficiency as a well-tuned HMC algorithm [113]. The convergence is usually checked by empirical diagnostic tools [115]. Also, we carefully set the initial values of the parameters to make the convergence faster by exploring the outage data in Section III.

APPENDIX D. Convergence of sampling algorithm

This appendix uses four methods to check the convergence of the Hamiltonian Monte Carlo algorithm used to sample the posterior distributions, including potential scale reduction factors, effective sample size diagnostics, trace plots, and autocorrelation plots. In addition, we check that the algorithm is not getting stuck in a local mode in the posterior distributions.

The Gelman-Rubin potential scale reduction factor diagnostic $\hat{R}$ is often used to check Markov chain Monte Carlo convergence [115]. $\hat{R}$ is defined as the ratio of the estimated pooled variance to the estimated within-chain variance (see [45, Sec. 11.4] for the equations of $\hat{R}$). Figure D.1 plots the iterates of $\hat{R}$ for all parameters at increments of 20 iterations from four parallel Markov chains. Figure D.1 shows that all $\hat{R}$s converge and are less than 1.1 after 400 iterations.

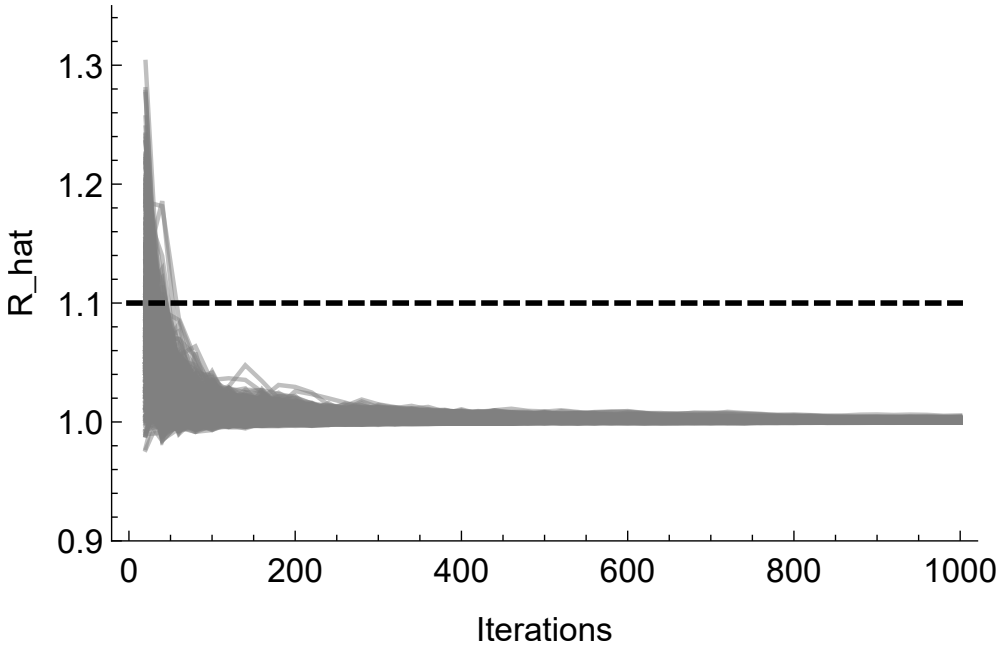


Figure D.1: Iterates of $\hat{R}$ for all parameters computed from four parallel Markov chains at increments of 20 iterations.

As suggested by Gelman [45], we also compute the effective sample size $\hat{n}_{eff}$, which is the equivalent number of independent samples that have the same standard error of the sample mean of the parameter as the Markov chain samples (see [45, Sec. 11.4] for the equations of $\hat{n}_{eff}$). It turns out that $\hat{n}_{eff}$s for all λs are greater than 100 per chain after 300 iterations, which shows that the estimates are reliable.

Graphical methods provide another way to check convergence. We make trace plots and autocorrelation function plots for each variable to check whether the chains are mixing and have large autocorrelation. It is not practical to show all the plots here. Instead, we randomly select four parameters to show the trace plots (Figure D.2) and autocorrelation function plots (Figure D.3). The two chains have mixed, and the autocorrelation decreases quickly and tends to zero.

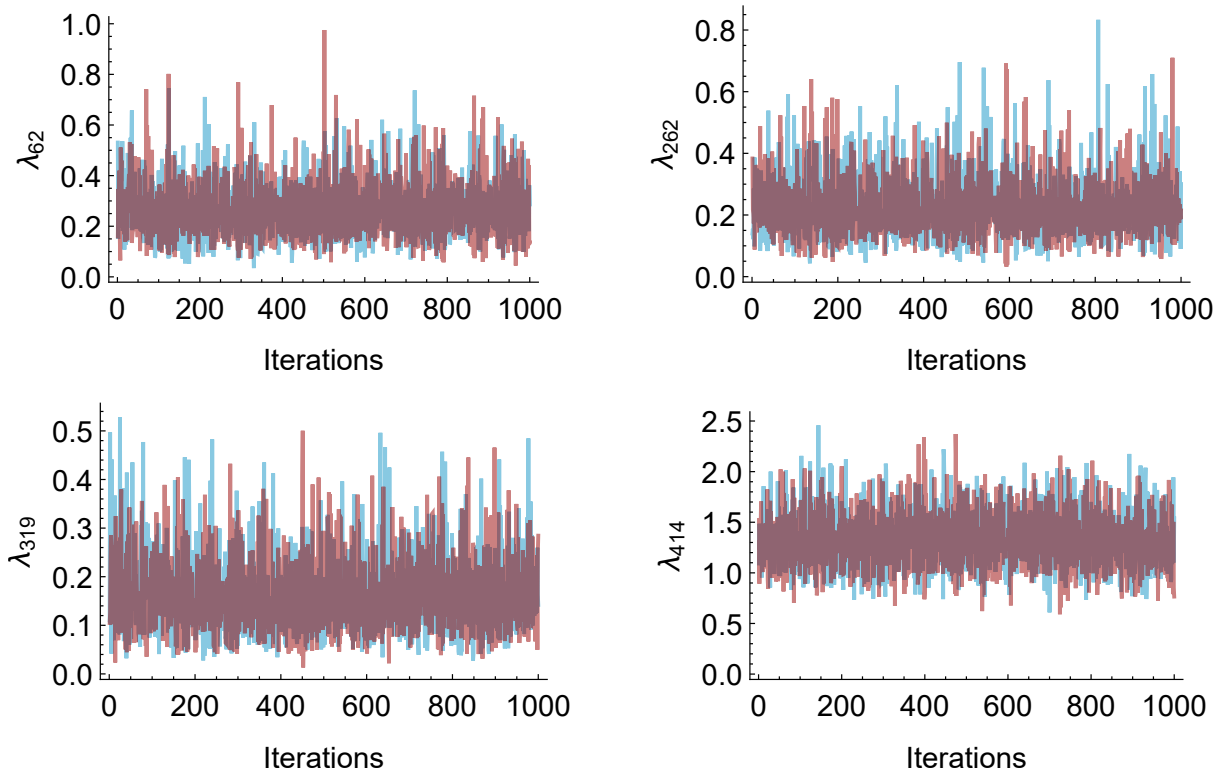


Figure D.2: Trace plots of two chains of four randomly selected λs.

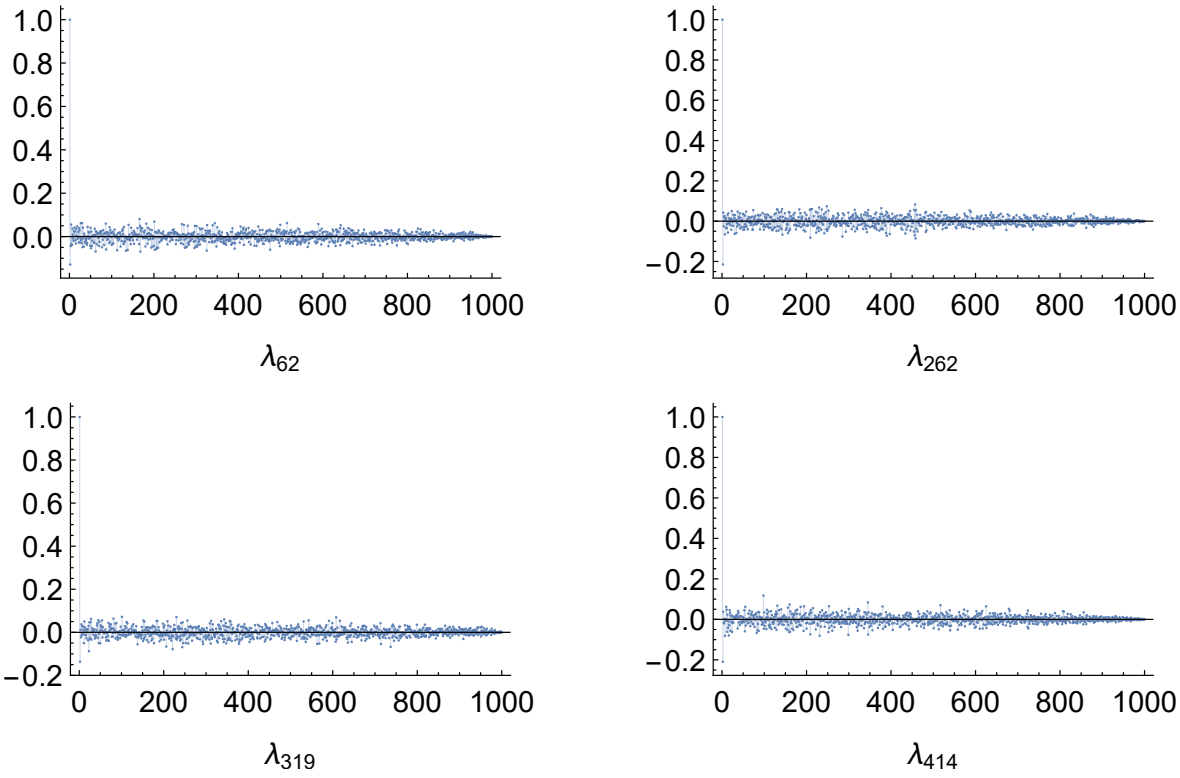


Figure D.3: Autocorrelation function plots of four randomly selected λs.

Based on the results of the four methods of checking convergence, we conclude that there is no evidence of nonconvergence.

To check that the algorithm is not getting stuck in a local mode in the posterior probability distribution, we simulate two additional Markov Chains with random initial values sampled from a uniform distribution over the support of parameters. Each of these additional Markov Chains has 3000 iterations in which the first 2500 samples are burn-in and are thrown away. We compare the posterior distributions of all parameters estimated from the additional chains and the original chain with the initial values in the body of the paper, and we find no convergence issues. Moreover, as we are most interested in the outage rates λ, we implement a Kolmogorov-Smirnoff test on the corresponding distributions of outage rates of the two chains that start with random values. All the λs except two are judged to be from the same distribution with a significance level

0.01. And these two λs have close means (0.32 and 0.33, 0.14 and 0.15) and close standard deviations (0.14 and 0.13, 0.08 and 0.08).