The following publication Li, X., Pan, T., & Pan, K. (2023). Identification of Key Components After Unintentional Failures for Cascading Failure Protection. IEEE Transactions on Network Science and Engineering, 10(2), 1003-1014 is available at https://doi.org/10.1109/TNSE.2022.3225459.

Identification of Key Components after Unintentional Failures for Cascading Failure Protection

Xiang Li* *Member, IEEE*, Tianyi Pan†, Kai Pan‡

*Department of Computer Engineering, Santa Clara University, Santa Clara, CA, USA

†Google Inc, Mountain View, CA, USA

‡Department of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Email: xli8@scu.edu, tianyipan@google.com, kai.pan@polyu.edu.hk

***Abstract*—Cascading failure can aggravate the vulnerability of power grids, which brings attention to cascading failure protection research. Existing works focus on either finding the critical components whose failure can cause large-scale blackouts or methods to mitigate failures after they have happened. However, they are not able to proactively protect against real-world failures, which may not only happen at the critical components. In this paper, we study the problem of finding components that will be impacted the most after unintentional initial failures, which suits the need for practical scenarios. The problem is challenging since approaches like simulating a large number of cascading failures cannot scale and they must be redone when power network parameters change . To tackle the problem, we derive a line importance metric based on all paths and illustrate how it is correlated with highly impacted lines after unintentional failure both intuitively and with an IEEE test case. Further, we design a path sampling algorithm to estimate the metric with provable guarantee and achieve scalability. We evaluate the performance of the proposed method within a protection scenario using various IEEE test cases and demonstrate its superiority against several baseline methods.**

***Index Terms*—Cascading Failure, Vulnerability Analysis, Path Sampling, Protection**

I. Introduction

Smart grids are now essential parts of the modern society. The integration of cyber and physical processes has many benefits [1], however, it also opens up possibilities of attacks and accidents from the cyber surface ([2] and references therein) and makes smart grids more vulnerable. What aggravates the vulnerability of smart grids is cascading failure, where the failure of a component (e.g. a transmission line) in power grids can cause successive failures and eventually lead to a large blackout [3]. Many blackouts in real life are related to cascading failures ([4] and references therein), including three major blackouts in 2003 [5]. Because of the importance of smart grids and catastrophic impact of cascading failures, protection against cascading failures has been studied in various settings [3], [6]–[13].

The existing works toward cascading failure protection mainly answer two questions. Firstly, what are the top components in a power grid to fail such that the disruption to the network is maximized after cascading failure? Secondly, how to reduce the damage to the power grid when failures already happened? The first question is answered through vulnerability analysis [7]–[9], [14], which works well against intentional attacks. Since the malicious attacks usually aim at maximizing damage, the attack decisions overlap well with the top components revealed with vulnerability analysis. Typical answers to the second question are through load shedding [3] or controlled islanding [15]. They are reactive ways that can eliminate transmission line overloading, usually at the cost of lower overall yield, via ramping down loads or intentionally trip transmission lines, respectively.

Based on the characteristics of real-world cascading failures and protection requirements, we observe a new question that has not been answered yet by existing works: what are the lines in a power grid that need protection the most after an unintentional failure? The reason why answering this question is needed is two-folds. On one hand, most of the cascading failures in real life are due to accidents like fallen trees, severe weather conditions, human error, etc [4], hence they may not just cause failures to the critical components and vulnerability analysis that focuses on intentional attacks may not be as effective against such accidents. On the other hand, if load shedding or controlled islanding decisions are not made in a timely manner, cascading failure can still happen [16]. Hence, a proactive approach that makes decisions prior to actual failures is necessary. When we identify the components that may fail right after unintentional failures, we can enhance those components, for example, increase transmission line capacity, and allocate more resources to monitor and control such lines.

An intuitive way for answering the new question is to leverage cascading failure data [6], [17]–[22]. [6] tries to find the interaction of components via data analysis and propose to protect against cascading failure by applying certain control measures to throttle critical component to component interactions. However, since real-world cascading failure data is limited, in most of the cases (except for [22]), analysis requires extensive simulation of the cascading failure process. It can be very costly to simulate cascading failure for all size-k initial random component failures even with moderate values of k. For example, in a power grid with 200 transmission

lines, collecting all cascading failure data when $k = 5$ requires 2.5 billion simulations. Also, in all cases, the data analysis must be redone whenever power network parameters change. Therefore, it is challenging to scale this method and hence a light-weight approach is needed. Also, the approach is preferably reusable with different power network parameters.

In this paper, we propose a light-weight, effective and proactive approach that can protect against or mitigate the impact of cascading failures from unintentional initial failures. It is achieved through two steps. For the first step, we utilize all paths between generators and loads to derive a line importance metric and experimentally show that it aligns well with metrics derived from cascading failure data. Although generating all paths are much more cost effective than generating all cascades, it is still not practical for large power grids. Hence, for the second step, we design a path sampling approach and prove theoretically that the line importance metric generated using path samples and that generated using all path are close. Also, since the path samples are dependent only to the topology of the power network, they don't have to be recalculated when power supplies/demands change. The proposed line importance metric generated with path samples is compared with a few other network topology based metrics in a protection scenario. We experimentally verified that protections based on the proposed metric has superior performance against other metrics in various IEEE test cases, in terms of both number of line failures as well as total load shed.

Our main contributions are summarized as follows:

- We propose an all path based line importance metric and show that lines with high values in this metric are critical in mitigating cascading failures from unintentional initial failures, in other words, they are more vulnerable to fail after unintentional initial failures.
- We design a sampling approach that can efficiently and accurately estimate the proposed line importance metric with theoretically bounded error margin.
- We evaluate the proposed metric in a protection scenario using power network test cases and show that it outperforms a few topology based line importance metrics.

Organization. The rest of the paper is organized as follows. We first summarize the related works in Section. II. In Sect. III, we propose a line importance metric based on all paths, as well as demonstrate that the proposed all path based line importance metric captures the critical lines in cascading failure after unintentional initial failure. Then, a path sampling method for estimating the all path based line importance metric is proposed in Section. IV and the error bounds are proved. The performance of cascading failure protection using the new line importance metric are evaluated within various IEEE test cases in Section. V. We conclude the paper in Sect. VI.

II. Related Works

Cascading failure in power grids is an important practical problem since it is a major cause of power system blackouts [4], [5]. Also, it is a complicated problem in theory. Cascading failures are usually modeled by iteratively removing overloaded lines, islanding and load re-balancing and the process involves solving power equations for each iteration [3]. The cascading failure model involves multiple steps and is hard to be analyzed under a unified theoretical framework. One additional complexity of cascading failure in power grid is that the evolution of the cascades may not be contiguous [3], [23], which limits the application of typical diffusion models into cascading failure analysis.

One major line of research towards cascading failure protection is vulnerability analysis. The goal is to identify the critical components whose failure can lead to large-scale blackouts. Because of the complexity of the problem, existing works mostly focus on heuristics combined with optimization techniques [7]–[14] or game theoretical approaches [24]–[26]. Another line of research uses load shedding or controlled islanding to mitigate the impact of cascading failure [3], [15], [27]–[29]. However, the works cannot be applied to proactively protect against cascading failure after unintentional initial failures, since vulnerability analysis is more effective towards intentional attacks aiming at the critical components and load shedding/controlled islanding approaches are reactive.

Some recent works rely on simulated cascading failure data to obtain interactions of components in cascading failures [6], [17], [20], [22], make predictions of cascading failures [18], [21] or reveal characteristics of cascading failures [19]. These works, together with another line of work that builds simpler cascading failure models [30], [31], can shed light on the importance of components in a cascading failure. However, the goal of most of the works (except for [6]) are similar to vulnerability analysis, hence the identified critical lines may not align well for the task of protecting against unintentional initial failures. Also, since the approaches based on simulated cascading failure data would have to restart simulation from scratch whenever the power network parameters change and most of them (except for [22]) require a large amount of cascading failure data to begin with, the approaches may not be scalable. Therefore, there is a need for an approach to identify components that are vulnerable after unintentional initial failures, which should be light-weight and is not necessarily relying on cascading failure data, so the result would not be impacted by power network parameters changes.

III. Quantifying Line Importance

In this section, we first discuss a view of line importance after unintentional initial failures using cascading failure data. Then, we derive a line importance metric based on paths and show the relation to cascading failure based line importance. Especially, we give examples on why the proposed metric based on all paths is better than existing metrics based on shortest paths, and validate the results with an IEEE test case.

A. Line Importance from Cascading Failure Data

There are different ways to quantify line importance in cascading failure. The most prominent approach is based on the number of subsequent line failures when a certain line fails [7], [9], [13]. In the context of cascading failure protection against unintentional initial line failures, a more preferable alternative is to consider the first round failures: the lines that

fail right after initial failures. We adopt the cascading failure model described in [3], which is based on DC power flow, and a round means one iteration of islanding, supply-demand balancing, solving power equations and removal of overheated lines. Conceptually, this view emphasizes the importance of lines that are the most susceptible to failures of other lines, comparing to the approach that deem the lines that can cause more other lines to fail as more important. Also, there are a variety of ways that lines may fail in an unintentional failure (natural disaster, human error, etc.). Although some types of unintentional failures, like natural disasters, are more likely to happen in certain regions of the power system, the exact locations that such failures can happen are still random. Other types of failures, like faulty equipment, human errors, vehicle accidents and trees [4], are more random in nature. Hence, thorough protection of those unintentional failures can be quite complicated, while subsequent line failures in cascading failure can be described by models and the protection of which is more approachable. The reason for choosing only the first round failures but not all failures is that focusing on first round failures is more beneficial to mitigating cascading failures as early as possible.

Denote C as the collection of cascades and let $c \in C$ denote one cascade. For each cascade c, let $r_c^0, r_c^1, \ldots$ denote the set of failed lines in each round and r_c^0 are the initial failures. We define the cascading failure based line importance $I_l^c, l \in L$ as the line's frequency of appearance in first round failures r_c^1:

$$I_l^c = \frac{|\{c|l \in r_c^1\, c \in C\}|}{|C|}, \forall l \in L \tag{1}$$

where L is the set of transmission lines.

B. Line Importance from Paths

There are a few existing line importance metrics that are based on paths. The examples are edge betweenness centrality [32] and edge current flow betweenness centrality [33]. Since edge betweenness centrality is based on shortest paths, it is less useful in the context of cascading failure. Failing lines with largest betweenness centrality may merely increase the length of shortest paths between generator/load pairs, with no impact on connectivity and power network dynamics. Edge current flow betweenness centrality utilizes unit current injection to decide the involvement of lines in the current flow from a source/destination node pair. One downside of using this metric in power networks is that the characteristics of nodes are not considered. Electrical betweenness centrality [13] was proposed to include power of generators and load of loads, it is proposed as a node centrality metric but can be extended to lines. However, although both edge current flow betweenness and electrical betweenness implicitly considers all paths, such line importance metrics do not reveal how lines will respond when other lines fail, which is the primary focus in this paper since we target the first round failures for protection.

The proposed line importance metric is based on an intuitive interpretation about the impact of the failure of one line to another line: Before the failure, power flow between a generator/load pair can utilize all paths connecting them, while after the failure, there will be more pressure on the paths that do not include the failed lines. Consider a generator s, a load d and denote all paths between them as $\mathcal{P}^{sd}$. For a line l, let $\mathcal{P}_l^{sd}$ be the set of paths between s, d that contains l. Based on the intuition, we can approximate a line l's pressure η after the failure of line l' as:

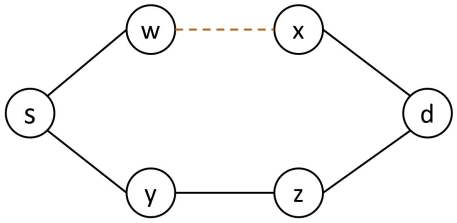


Fig. 1: Non-local Cascade

$$\eta_l^{l'} = \sum_{s\in S}\sum_{d\in D} \frac{|\mathcal{P}_l^{sd} - \mathcal{P}_{l'}^{sd}|}{|\mathcal{P}^{sd} - \mathcal{P}_{l'}^{sd}|} \tag{2}$$

where S and D are the sets of generators and loads, respectively and "$-$" is the set difference operation.

(2) aggregates the fraction of survived paths after failure of line l' that contains line l, for all generator/load pairs, which is directly related to the evaluation of first round failures. One extra benefit of this interpretation is that it explains the non-local cascade behavior with respect to the topology [3], [23] with only topological information. Consider the example in Figure 1 with one generator s and one load d. There are two paths between the generator/load pair, when the line (w, x) fails, the flow on path (s, w, x, d) will be redistributed to path (s, y, z, d) and may cause non-local cascading failure on line (y, z).

In the case of unintentional failures, we can aggregate $\eta_l^{l'}$ for line l over all $l' \neq l$ to quantify the impact line l receives. Also, similar to the electrical betweenness centrality metric [13], to let lines connecting larger generators and loads having higher importance, we weigh each generator/load pair by $\sqrt{w_s w_d}$ where w_s and w_d are the power for generators and load for loads, respectively. Additionally, we let line importance to be reversely proportional to line capacity, since based on the cascading failure model in [3], a line fails when its flow is over the capacity by a certain margin, which means that lines with lower capacities are more likely to fail.

In summary, the proposed all path based line importance metric can be written as:

$$I_l^p = \sum_{s\in S}\sum_{d\in D} \frac{\sqrt{w_s w_d}}{r_l} \sum_{l'\in L^{sd}, l'\neq l} \frac{|\mathcal{P}_l^{sd} - \mathcal{P}_{l'}^{sd}|}{|\mathcal{P}^{sd} - \mathcal{P}_{l'}^{sd}|} \tag{3}$$

where r_l is the capacity of line l and L^{sd} is the collection of all lines that are included in any paths between generator s and load d. In the special case where $\mathcal{P}^{sd} - \mathcal{P}_{l'}^{sd} = \emptyset$, we let $\frac{|\mathcal{P}_l^{sd} - \mathcal{P}_{l'}^{sd}|}{|\mathcal{P}^{sd} - \mathcal{P}_{l'}^{sd}|} = 0$. Note that it is implicitly assumed that all lines fail with the same probability with an unweighted aggregation. If more information on the power system is available, it is straightforward to aggregate $\eta_l^{l'}$ based on individual failure probability of lines. In future renewable grids, a node can be

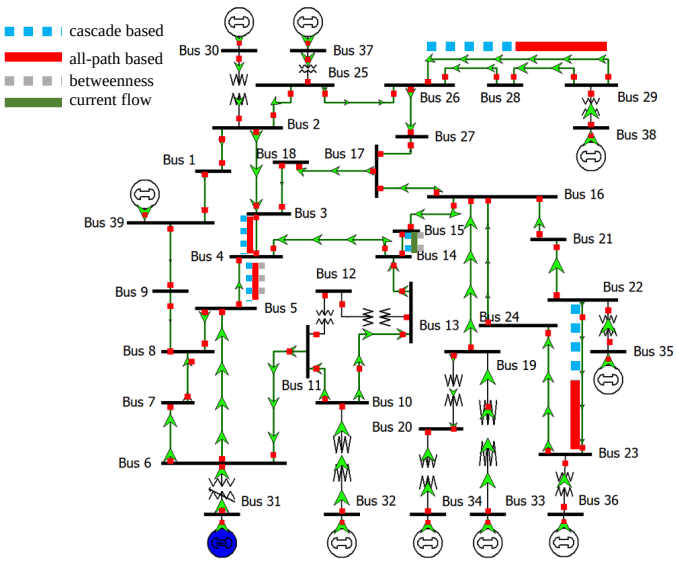


Fig. 2: Top-5 Important Lines in IEEE 39-Bus System

a generator or a load at different times (e.g. a house with solar panels can be a generator at daytime and a load at night time). In such cases, we can calculate I_l^p for different snapshots and take the average.

In practice, factors like transmission line material and age can also be considered in I_l^p. We emit those factors in this theoretical model to focus on network topology and power flow and the associated line importance metric for cascading failure. Note that identifying important lines with a theoretical model can aid the decision making in system planing considering the practical factors, for example, having more regular maintenance or rebuild the important lines.

Since in the worst case, enumerating all paths in a connected network between one pair of nodes takes exponential time (with respect to the number of lines), the proposed metric cannot scale to larger networks. Instead of calculating I_l^p directly, we will show in Section IV that we can approximate it using path samples with guaranteed accuracy. In the remaining of this section, we will demonstrate the effectiveness of the metric in a small IEEE test case.

C. Effectiveness of the All Path Based Line Importance Metric

In order to examine the effectiveness of the proposed metric, we compute I_l^c and I_l^p from the IEEE 39-bus system [34]. To compute I_l^c, we simulate all scenarios in n-2 and n-3 contingency analyses (hence $\binom{n}{2} + \binom{n}{3}$ cascades for a network with n nodes). As a baseline, we also compute edge betweenness centrality I_l^b and edge current flow betweenness centrality I_l^f for all lines. Then, we order all lines in descending order by the four metrics respectively. Since the main purpose is to find out whether I_l^p can identify the lines that are impacted the most by unintentional initial failures, we focus on the top-ranked lines in the comparison. Specifically, how many lines with top-5 I_l^c are ranked as top-5 by I_l^p (and I_l^b, I_l^b respectively). The result is visualized in Fig. 2 (original figure is from [35]).

It is clear that all path based line importance resonates well with cascading failure data based line importance, if we consider the top-5 ranked lines. Four out of five top lines by highest I_l^c are also top-5 in I_l^p. On the contrary, edge betweenness centrality and edge current flow betweenness centrality cannot capture the important lines in a cascade: their top-5 ranked lines only overlap with two and one lines with top-5 lines by I_l^c.

IV. Probabilistic Approximation of All Path Based Line Importance

In this section, we introduce the sampling approach that can be used to approximate all path based line importance I_l^p.

A. Path Sampling

Although there are multiple unknown metrics in (3), it is only necessary to estimate one of them. Let

$$\mathcal{P}_{l\neg l'}^{sd} = \mathcal{P}_{l}^{sd} - \mathcal{P}_{l'}^{sd}, \quad \mathcal{P}_{\neg l'}^{sd} = \mathcal{P}^{sd} - \mathcal{P}_{l'}^{sd}$$

the metric we need is

$$\frac{|\mathcal{P}_{l\neg l'}^{sd}|}{|\mathcal{P}_{\neg l'}^{sd}|}$$

which is the fraction of (s, d) paths not including line l' that include line l.

In order to obtain the estimate of the metric, we extend the path sampling methods discussed in [36], [37]. The main difference is that in this work, we introduce a new pmf for path sampling. Also, we add explicit restrictions that certain lines should not be included in the sample, which was not the case for neither [36] nor [37]. In the algorithm, the probability of a path is denoted as $g(p)$ and $g(p)$ has domain $D(s, d, L, l')$.

Algorithm 1 Path Sampling Algorithm

Input: Power network $G = (V, E)$, generator $s \in V$, load $d \in V$, path length threshold L, line to avoid l'.
Output: A path p and probability $g(p)$
  $a_0 = s$, $i = 0$, $p = a_0$, $f(p) = 1$
  while $a_i \neq d$ **do**
    Let $N(a_i) = \{b_i | (a_i, b_i) \in E\}$ be neighbors of a_i that are not in p, adding b_i to p will keep path length within L and $(a_i, b_i) \neq l'$.
    Choose a_{i+1} according to pmf $p(a_{i+1}|a_i, N(a_i), L - i)$ from $N(a_i)$.
    if no a_{i+1} found **then**
      break
    $i = i + 1$, $p = pa_i$, $g(p) = g(p) \times \frac{1}{|N(a_i)|}$
  return $p = a_0 a_1 ..., g(p)$

The path sampling algorithm is described in Alg. 1. It starts from the generator node s and iteratively pick nodes from a set of candidate nodes in the path until either load node d is reached or the candidate node set is empty. The candidate node set contains all nodes that are neighbors of the current node and 1) adding the node will not form a cycle, 2) adding the node will not cause the path length to be greater than a predefined threshold L and 3) the line connecting the current node and new node is not l', the line to avoid. We add condition 2) to control the impact of long paths. In our

problem, the extremely long paths have low contribution since they are likely to be disconnected whenever unintentional line failure happens, hence have little impact in (3).

In both [36] and [37], the next node is selected uniformly randomly from the set of candidate nodes, so the probability mass function (pmf) is

$$P(a_{i+1} = v) = \frac{1}{N(a_i)}, \forall v \in N(a_i) \tag{4}$$

However, there are two problems if we use this pmf. Firstly, as mentioned in [36], the shorter paths will have much higher probability. Secondly, many sample paths may be wasted as they cannot reach to the load d when the length already hits L. For the first problem, [36] proposed to first estimate the length distribution of paths and then update the probability of selecting node d (when it is among the candidates) according to the length distribution. The approach may not work well in our problem since a large fraction of samples will be wasted and it may take too many samples to estimate a good enough length distribution. [37] mitigated the second problem by generating a shortest path between (s, d) when Alg. 1 failed to find a path for a predetermined number of times, but it still has a large amount of wasted samples.

In order to solve both problems, we propose to first obtain a crude estimate of $\mathcal{P}^{sd}$ with pmf (4), denote as ρ^{sd} and then utilize ρ^{sd} as well as shortest path distances among nodes in the main sampling process. Denote $sp(u, v)$ as the shortest path distance between two nodes u, v, we have

$$p(a_{i+1}) = \begin{cases} \min(\frac{1}{g(p)\rho^{sd}}, 1), & d \in N(a_i), a_{i+1} = d \\ \frac{1}{|N(a_i)|}, & sp(a_i, d) < L - i, d \notin N(a_i) \\ \frac{1-\min(\frac{1}{g(p)\rho^{sd}},1)}{|N(a_i)-1|}, & sp(a_i, d) < L - i, \\ & d \in N(a_i), a_{i+1} \neq d \\ \frac{1}{|N^-(a_i)|}, & sp(a_i, d) >= L - i \end{cases}$$

The new pmf depends on two main factors: 1) whether the load d is in $N(a_i)$ and 2) whether the shortest path distance to d from the current node a_i is shorter than the remaining length. Whenever $d \in N(a_i)$, we will select d as the next node with a probability that the overall probability of the path is as close to $\frac{1}{\rho^{sd}}$ as possible. When $sp(a_i, d) < L-i$, it is still possible to reach d eventually no matter what node is selected, so we uniformly randomly select nodes in $N(a_i)$ as the next node. Note that the probability is adjusted when $d \in N(a_i)$. $N^-(a_i) = \{v | v \in N(a_i), sp(v, d) = sp(a_i, d) - 1\}$, it means that when the shortest path distance to d from the current node is no smaller than the remaining length, the next node will be selected only from nodes that are closer to d than the current node. Otherwise, the sampled path will never reach d.

B. The Estimators

For notation convenience, let $f^{sd}_{l\neg l'} = \frac{|\mathcal{P}^{sd}_{l\neg l'}|}{|\mathcal{P}^{sd}_{\neg l'}|}$. Also, for a random path p, let $I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'})$ be an indicator function that takes value 1 if $p \in \mathcal{P}^{sd}_l$ and $p \notin \mathcal{P}^{sd}_{l'}$, and takes value 0 otherwise. In the following, we first assume the knowledge of $|\mathcal{P}^{sd}_{\neg l'}|$ and prove the following unbiased estimators of $f^{sd}_{l\neg l'}$

Lemma 1. *Let R be the set of path samples, then $\bar{f}^{sd}_{l\neg l'} = \frac{1}{|R|}\sum_{p\in R} \frac{I(p\in\mathcal{P}^{sd}_l \wedge p\notin\mathcal{P}^{sd}_{l'})}{g(p)|\mathcal{P}^{sd}_{\neg l'}|}$ is an unbiased estimator for $f^{sd}_{l\neg l'}$.*

Proof. Let $X^{sd}_{l\neg l'}(p)$ be the random variable

$$X^{sd}_{l\neg l'}(p) := \frac{I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'})}{g(p)|\mathcal{P}^{sd}_{\neg l'}|} \tag{5}$$

The expectation of $X^{sd}_{l\neg l'}(p)$ is:

$$\begin{aligned} \mathbf{E}(X^{sd}_{l\neg l'}(p)) &= \sum_{p\in D(s,d,L,l')} \frac{I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'})}{g(p)|\mathcal{P}^{sd}_{\neg l'}|} \times g(p) \\ &= \sum_{p\in D(s,d,L,l')} \frac{I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'})}{|\mathcal{P}^{sd}_{\neg l'}|} \\ &= \frac{1}{|\mathcal{P}^{sd}_{\neg l'}|} \sum_{p\in D(s,d,L,l')} I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'}) \\ &= \frac{|\mathcal{P}^{sd}_{l\neg l'}|}{|\mathcal{P}^{sd}_{\neg l'}|} = f^{sd}_{l\neg l'} \quad \square \end{aligned}$$

Since it is #P-complete to compute $|\mathcal{P}^{sd}_{\neg l'}|$ [38], we will not be able to use the estimators in Lemma 1 directly. Instead, we will first estimate $|\mathcal{P}^{sd}_{\neg l'}|$ with an unbiased estimator and use the estimate $\bar{\mathcal{P}}^{sd}_{\neg l}$ in the estimator of $f^{sd}_{l\neg l'}$.

Lemma 2. $\frac{1}{|R|}\sum_{p\in R} \frac{I(p\in\mathcal{P}^{sd} \wedge p\notin\mathcal{P}^{sd}_{l'})}{g(p)}$ *is an unbiased estimator of $|\mathcal{P}^{sd}_{\neg l'}|$.*

The proof is similar to the one for Lemma 1 and hence omitted.

Let $\bar{\mathcal{P}}^{sd}_{\neg l'}$ be the estimator of $\mathcal{P}^{sd}_{\neg l'}$ where $\bar{\mathcal{P}}^{sd}_{\neg l'} := \frac{1}{|R|}\sum_{p\in R} \frac{I(p\in\mathcal{P}^{sd} \wedge p\notin\mathcal{P}^{sd}_{l'})}{g(p)}$, we have the following corollary of Lemma 1.

Corollary 1. $\tilde{f}^{sd}_{l\neg l'} = \frac{1}{|R|}\sum_{p\in R} \frac{I(p\in\mathcal{P}^{sd}_l \wedge p\notin\mathcal{P}^{sd}_{l'})}{g(p)\bar{\mathcal{P}}^{sd}_{\neg l'}}$ *is an unbiased estimator for $\frac{|\mathcal{P}^{sd}_{\neg l'}|}{\bar{\mathcal{P}}^{sd}_{\neg l'}} f^{sd}_{l\neg l'}$.*

From Corollary 1, it is clear that (3) can be rewritten with the unbiased estimators defined above:

$$I^p_l = \sum_{s\in S}\sum_{d\in D} \frac{\sqrt{w_s w_d}}{r_l} \sum_{l'\in L^{sd}, l'\neq l} \frac{\bar{\mathcal{P}}^{sd}_{\neg l'}}{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})} \mathbf{E}(\tilde{f}^{sd}_{l\neg l'}) \tag{6}$$

C. Accuracy of the Estimators

In this section, we derive theoretical bounds on the number of path samples required to obtain estimators for $|\mathcal{P}^{sd}_{\neg l'}|$ and $f^{sd}_{l\neg l'}$ with performance guarantee.

We start with the estimator $\bar{\mathcal{P}}^{sd}_{\neg l'}$ of $|\mathcal{P}^{sd}_{\neg l'}|$. Let $X^{sd}_{\neg l'}(p)$ be the random variable $X^{sd}_{\neg l'}(p) := \frac{I(p\in\mathcal{P}^{sd} \wedge p\notin\mathcal{P}^{sd}_{l'})}{g(p)}$. It is clear that $\frac{1}{g(p)} \leq d^L_{max}$ where d_{max} is the maximum degree in the network, so $X^{sd}_{\neg l'}(p) \in [0, d^L_{max}]$.

Lemma 3 (Hoeffding's Inequality)**.** *[39] Let $X_1, X_2, ..., X_n$ are independent random variables and X_i is strictly bounded by $[a_i, b_i]$. Denote $\bar{X} = \frac{1}{n}\sum_{i=1,...,n} X_i$, we have*

$$\boldsymbol{P}(|\bar{X} - \boldsymbol{E}(\bar{X})| \geq t) \leq 2\exp(\frac{-2n^2t^2}{\sum_{i=1,...,n}(a_i - b_i)^2})$$

Theorem 1. *With $|R_1| = \frac{\ln 2 - \ln \delta_1}{2\epsilon_1^2}$ path samples, we have*

$$\boldsymbol{P}(|\bar{\mathcal{P}}^{sd}_{\neg l'} - \boldsymbol{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})| \geq t_1) \leq \delta_1 \tag{7}$$

where $\epsilon_1 > 0, \delta_1 \in (0,1)$ and $t_1 = \epsilon_1 d^L_{max}$.

Proof. By Hoeffding's inequality, we have

$$\begin{aligned}
&\mathbf{P}(|\bar{\mathcal{P}}^{sd}_{\neg l'} - \mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})| \geq t_1) \\
&= \mathbf{P}(|\frac{1}{|R_1|}\sum_{p\in R_1} X^{sd}_{\neg l'}(p) - \mathbf{E}(\frac{1}{|R_1|}\sum_{p\in R_1} X^{sd}_{\neg l'}(p))| \geq t_1) \\
&\leq 2\exp(\frac{-2|R_1|^2 t_1^2}{\sum_{i=1,...,|R_1|}(d^L_{max})^2}) \\
&\leq 2\exp(\frac{-2|R_1|t_1^2}{(d^L_{max})^2})
\end{aligned}$$

Solving $2\exp(\frac{-2|R|t_1^2}{(d^L_{max})^2}) = \delta_1$ gives $|R_1| = \frac{\ln 2 - \ln \delta_1}{2\epsilon_1^2}$. □

Let $Y^{sd}_{l\neg l'}(p)$ be a random variable

$$Y^{sd}_{l\neg l'}(p) := \frac{I(p \in \mathcal{P}^{sd}_l \wedge p \notin \mathcal{P}^{sd}_{l'})}{g(p)\bar{\mathcal{P}}^{sd}_{\neg l'}}$$

It is clear that $Y^{sd}_l(p) \in [0, \frac{d^L_{max}}{\bar{\mathcal{P}}^{sd}_{\neg l'}}]$. Using similar proof techniques as in Theorem 1, we have the following theorem for the number of samples to guarantee a bounded estimate of $\tilde{f}^{sd}_{l\neg l'}$.

Theorem 2. *With $|R_2| = \frac{\ln 2 - \ln \delta_2}{2\epsilon_2^2}$ path samples, we have*

$$\boldsymbol{P}(|\tilde{f}^{sd}_{l\neg l'} - \boldsymbol{E}(\tilde{f}^{sd}_{l\neg l'})| \geq t_2) \leq \delta_2 \tag{8}$$

where where $\epsilon_2 > 0, \delta_2 \in (0,1)$ and $t_2 = \epsilon_2 \frac{d^L_{max}}{\bar{\mathcal{P}}^{sd}_{\neg l'}}$.

D. Performance Bound and the Complete Algorithm

In this section, we will propose the algorithm that yields estimations of the path importance metric defined in (6) with performance guarantee, utilizing the building blocks obtained from the previous sections.

We first extend Theorem 1 to obtain a data-dependent bound of $|\bar{\mathcal{P}}^{sd}_{\neg l'} - \mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})|$. Specifically, we want to generate enough samples such that $\bar{\mathcal{P}}^{sd}_{\neg l'} \geq k\epsilon_1 d^L_{max}$, where k is a predefined positive number. We can achieve that by keep increasing the number of samples to estimate $\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})$. This way, we have

$$\mathbf{P}(|1 - \frac{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})}{\bar{\mathcal{P}}^{sd}_{\neg l'}}| \geq \frac{1}{k}) \leq \delta_1 \tag{9}$$

Then, we prove the performance bound for estimating the path importance metric I^p_l for all lines $l \in E$ with both multiplicative and additive error terms.

Theorem 3.

$$\boldsymbol{P}(\frac{k-1}{k}I^p_l - \frac{\theta}{k+1} \leq \bar{I}^p_l \leq \frac{k+1}{k}I^p_l + \frac{\theta}{k-1}) \geq 1 - \delta, \forall l \in E$$

where $\theta = \frac{\epsilon_2|S||D||E|}{\epsilon_1}$ and $\delta \in (0,1)$.

Proof. For each quadruple (s, d, l, l'), the error it brings to (6) is

$$\frac{\bar{\mathcal{P}}^{sd}_{\neg l'}}{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})}\mathbf{E}(\tilde{f}^{sd}_{l\neg l'}) - \tilde{f}^{sd}_{l\neg l'}$$

With (9), we have

$$\begin{aligned}
&\mathbf{P}(\frac{k-1}{k} \leq \frac{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})}{\bar{\mathcal{P}}^{sd}_{\neg l'}} \leq \frac{k+1}{k}) \geq 1 - \delta_1, \text{ so} \\
&\mathbf{P}(\frac{k}{k+1} \leq \frac{\bar{\mathcal{P}}^{sd}_{\neg l'}}{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})} \leq \frac{k}{k-1}) \geq 1 - \delta_1
\end{aligned} \tag{10}$$

Using the fact that $\bar{\mathcal{P}}^{sd}_{\neg l'} \geq k\epsilon_1 d^L_{max}$, the parameter t_2 defined in Theorem 2 is now $t_2 = \frac{\epsilon_2}{\epsilon_1 k}$ and we have

$$\begin{aligned}
&\mathbf{P}(|\tilde{f}^{sd}_{l\neg l'} - \mathbf{E}(\tilde{f}^{sd}_{l\neg l'})| \geq \frac{\epsilon_2}{\epsilon_1 k}) \leq \delta_2 \\
&\mathbf{P}(\tilde{f}^{sd}_{l\neg l'} - \frac{\epsilon_2}{\epsilon_1 k} \leq \mathbf{E}(\tilde{f}^{sd}_{l\neg l'}) \leq \tilde{f}^{sd}_{l\neg l'} + \frac{\epsilon_2}{\epsilon_1 k}) \geq 1 - \delta_2
\end{aligned} \tag{11}$$

Combining (10) and (11), we have

$$\begin{aligned}
\frac{\bar{\mathcal{P}}^{sd}_{\neg l'}}{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})}\mathbf{E}(\tilde{f}^{sd}_{l\neg l'}) - \tilde{f}^{sd}_{l\neg l'} &\geq \frac{k}{k+1}(\tilde{f}^{sd}_{l\neg l'} - \frac{\epsilon_2}{\epsilon_1 k}) - \tilde{f}^{sd}_{l\neg l'} \\
&= -\frac{\tilde{f}^{sd}_{l\neg l'}}{k} - \frac{\epsilon_2}{\epsilon_1(k+1)}
\end{aligned} \tag{12}$$

$$\begin{aligned}
\frac{\bar{\mathcal{P}}^{sd}_{\neg l'}}{\mathbf{E}(\bar{\mathcal{P}}^{sd}_{\neg l'})}\mathbf{E}(\tilde{f}^{sd}_{l\neg l'}) - \tilde{f}^{sd}_{l\neg l'} &\leq \frac{k}{k-1}(\tilde{f}^{sd}_{l\neg l'} + \frac{\epsilon_2}{\epsilon_1 k}) - \tilde{f}^{sd}_{l\neg l'} \\
&= \frac{\tilde{f}^{sd}_{l\neg l'}}{k} + \frac{\epsilon_2}{\epsilon_1(k-1)}
\end{aligned} \tag{13}$$

Also, the probability for having both (12) and (13) is $(1 - \delta_1)(1 - \delta_2)$.

Applying the bounds to all quadruples (s, d, l, l') gives

$$\begin{aligned}
\bar{I}^p_l - I^p_l &\geq -\frac{\sum_{s\in S}\sum_{d\in D}\frac{\sqrt{w_s w_d}}{r_l}k\tilde{f}^{sd}_{ll'}}{k} \\
&\quad - \frac{\epsilon_2|S||D|\sum_{s\in S}\sum_{d\in D}(|L^{sd}| - 1)}{\epsilon_1(k+1)} \\
&\geq -\frac{I^p_l}{k} - \frac{\epsilon_2|S||D||E|}{\epsilon_1(k+1)}
\end{aligned}$$

and

$$\begin{aligned}
\bar{I}^p_l - I^p_l &\leq \frac{\sum_{s\in S}\sum_{d\in D}\frac{\sqrt{w_s w_d}}{r_l}k\tilde{f}^{sd}_{ll'}}{k} \\
&\quad + \frac{\epsilon_2|S||D|\sum_{s\in S}\sum_{d\in D}(|L^{sd}| - 1)}{\epsilon_1(k-1)} \\
&\leq \frac{I^p_l}{k} + \frac{\epsilon_2|S||D||E|}{\epsilon_1(k-1)}
\end{aligned}$$

Let $\theta = \frac{\epsilon_2|S||D||E|}{\epsilon_1}$ and set $\delta_1 = \delta_2 = \frac{\delta}{|S||D||E|^2}$, we have

$$\mathbf{P}(\frac{k-1}{k}I_l^p - \frac{\theta}{k+1} \leq \bar{I}_l^p \leq \frac{k+1}{k}I_l^p + \frac{\theta}{k-1}) \geq 1-\delta$$

for all $l \in E$ by union bound. □

The complete algorithm is detailed in Alg. 2. Note that when the sample paths are generated and $\tilde{f}_{l\neg l'}^{sd}$ are calculated, Alg. 2 can directly calculate $\bar{I}_l^p$ as long as the topology of the power network stays the same.

Algorithm 2 All Path Based Line Importance Estimation

Input: Power network $G = (V, E)$, generator set S, load set D, path length threshold L, $k, \epsilon_2, \delta_1, \delta_2 > 0$
Output: $\bar{I}_l^p, \forall l \in E$.
if $\tilde{f}_{l\neg l'}^{sd}$ are not calculated or topology of G changed **then**
  for $\forall l \in E$ **do**
    for $\forall s \in S, d \in D, l' \in E$ **do**
      Run Alg. 1 to generate sample paths to estimate $\bar{\mathcal{P}}_{\neg l'}^{sd}$ until $\bar{\mathcal{P}}_{\neg l'}^{sd} = k\epsilon_1 d_{max}^L$.
      Generate $\frac{\ln 2 - \ln \delta_2}{2\epsilon_2^2}$ samples paths with Alg. 1 to estimate $\tilde{f}_{l\neg l'}^{sd}$ and store it.
for $\forall l \in E$ **do**
  $\bar{I}_l^p = \sum_{s\in S}\sum_{d\in D}\frac{\sqrt{w_s w_d}}{r_l}\sum_{l'\in L^{sd}, l'\neq l}\tilde{f}_{l\neg l'}^{sd}$.
return $\bar{I}_l^p, \forall l \in E$.

E. Approximate Algorithms for Better Efficiency

To further boost the efficiency of Alg. 2, we consider to only generate path samples for generator/load pairs with high $\sqrt{w_s w_d}$ values since other pairs have low contribution to I_l^p. Figure 3 shows the contribution to $\sum_{s\in S}\sum_{d\in D}\sqrt{w_s w_d}$ from top ranked generator/load pairs ordered by $\sqrt{w_s w_d}$ in descending order, for the IEEE 300-Bus System [40]. We can observe that top 50% of generator/load pairs contribute 84% of the sum, meaning we can cut required calculation by half and would not lose much accuracy.

V. Performance Evaluation

In previous sections, we figured out how to obtain the most important lines after unintentional initial failures and further developed sampling techniques to obtain such lines in large power grids efficiently. In this section, we utilize the identified important lines in cascading failure protection and evaluate the effectiveness of protection against unintentional initial failures.

A. Experiment Setup

1) Protection Strategy: In the experiments, in order to show the direct impact from protecting the identified important lines, we adopt a simple protection strategy: double the capacity of the lines based on their original capacity. This way, we can clearly observe the status of the individual lines and overall network stats before/after the protection. In practice, a large increase in transmission line capacity in a short period of time may not always be feasible. Instead, the capacities of lines can be increased over a longer time frame and the identified important lines can have higher priority in the process. For example, in the OPA models (e.g. [41]), line capacities are increased in the slow dynamics based on their vulnerability. It is possible to consider transmission line vulnerability after unintentional failures in such models when deciding the lines that should be improved.

For both estimations in Alg. 2, we generate a fixed number of sample paths for each triple of generator, load and the line to avoid. The number changes with data sets and is summarized in Table I. Also, the sample paths length limit L is set to be twice the diameter of each network.

2) Algorithms to Compare: We consider the following algorithms as baselines:

1) **Edge Betweenness**: pick top-K lines with highest edge betweenness centrality.
2) **Current Flow Betweenness**: pick top-K lines with highest edge current flow betweenness centrality.
3) **Cascade Importance**: pick top-K lines with highest I_l^c (defined in (1)) values, calculated from all $N-2$ and $N-3$ cascades.
4) **PageRank** [22]: pick top-K lines with highest PageRank value from a modified PageRank model in the interaction graph (where transmission lines are nodes in the graph), generated using all $N-1$ failures.
5) **GLODF**: In this algorithm, we calculate the flow change on each line without solving power flow equations with generalized line outage distribution factors (GLODF) [42]. Specifically, we first calculate the flow change on each line for each of the $N-2$ and $N-3$ cascades, and pick the K lines with highest $\frac{\text{average flow change}}{\text{line capacity}}$.
6) **Random**: pick K random lines to protect (result is averaged over 10 runs).
7) **No Protection**: do not apply any protection.

Most of the algorithms for finding critical lines in vulnerability analysis cannot scale to neither larger networks nor more number of critical lines. For example, [9] only considers the IEEE 300-Bus System with at most two critical lines. Hence, we only apply Cascade Importance and GLODF for the IEEE 118-Bus System. Also, as the size of the power grid networks can be very different, the value K depends on the network size instead of being fixed across all networks.

3) Evaluation Methods: For each power grid network, we randomly generate a fixed collection of initial failures, each set of initial failures contains 3 to 8 failed lines to simulate unintentional initial failures. Then, we will simulate cascading failure with each set of initial failures in power grid networks with the important lines generated by each algorithm being protected. As pointed out in [43], topological measurements may not be suitable to evaluate the impact of cascading failure, hence, we rely on measurements such as total number of line failures and total load shed, from the simulation result. Both measurements give idea of the blackout size and provide an approximation to the impact of unintentional failures to each protected network.

4) Data Sets: Throughout the experiments, we consider three data sets: IEEE 118-Bus System, IEEE 300-Bus System and the Polish 3120-Bus System. The data sets are all from the pandapower package [40]. In Table I, we summarize the

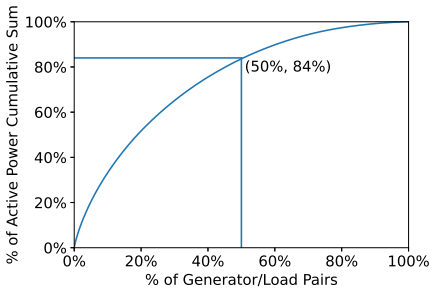


Fig. 3: CDF of $\sqrt{w_s w_d}$ for IEEE 300-Bus System

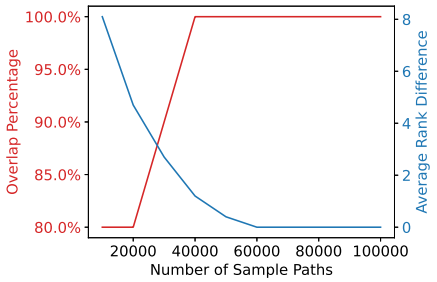


Fig. 4: Overlap and Rank Difference of Top-10 Lines with Varying # of Sample Paths. IEEE 118-Bus System.

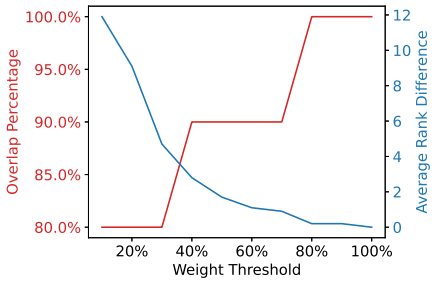


Fig. 5: Overlap and Rank Difference of Top-10 Lines with Varying Weight Threshold. IEEE 118-Bus System.

basic stats of the data sets as well as the number of top lines K used in each data set. Note that # lines in the table includes both transmission lines and transformers. Also, we ignore generators/loads whose active power is zero.

TABLE I: Stats of Data Sets

Data Set	# Lines	# Generators	# Loads	K	# Samples
118-Bus	186	23	80	10	100K
300-Bus	411	51	176	20	50K
3120-Bus	3693	124	2156	100	1K

B. Stability of the Important Lines

The first question we want to answer with experiments is: how much will the set of important lines change with the sampling approach? Stability is a key question since although Theorem 3 bounds line importance for all lines, there's no guarantee on the order of line importance values of the lines. For protection, it is crucial to ensure the set of most important lines stays stable in different runs so that the protection strategy is consistent. To answer this question, we calculate all path based line importance for all transmission lines in the IEEE 118-Bus System, varying the number of sample paths for each triple of generator, load and the line to avoid. Also, we examine the variation of the set of generator/load pairs to consider by varying weight threshold. A weight threshold of $X\%$ means we consider only top $X\%$ of generator/load pairs with highest $\sqrt{w_s w_d}$ values.

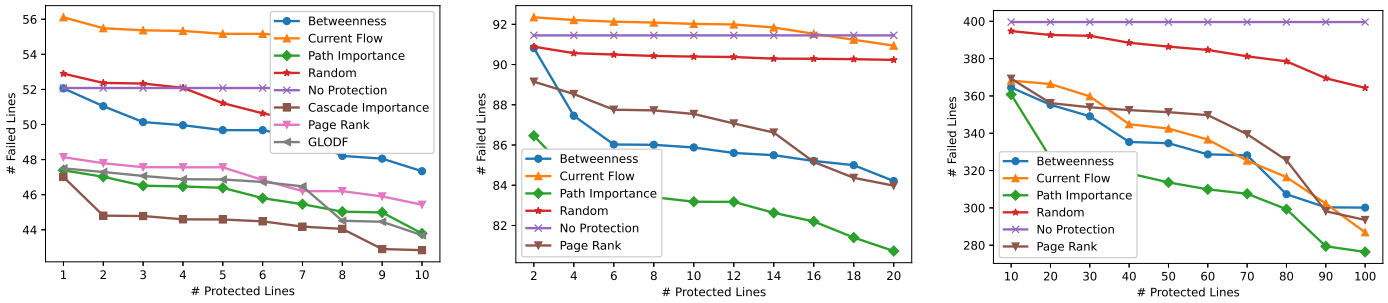
Stability is evaluated with a baseline and two metrics. The baseline is the line ranking with $100K$ samples and 100% weight threshold. The two metrics are: 1) **Overlap**: fraction of top-10 lines that also exist in top-10 of the baseline ranking. 2) **Average rank difference**: let the ranks of the top-10 lines in the baseline be $a_1, ..., a_{10}$, average rank difference is defined as $\frac{\sum_{i=1}^{10} |a_i - i|}{10}$.

The results are illustrated in Figs. 4 and 5. It is clear that the top-10 lines are quite stable: from Fig. 4, the top-10 lines have 100% overlap since $40K$ samples and their rankings are stabilized at $60K$ samples. From Fig. 5, the top-10 lines have 100% overlap since a weight threshold of 80%. Results for 300-Bus and 3120-Bus data sets are similar and are omitted for conciseness. To have better efficiency, we choose to use the weight threshold 50% for those two data sets in all path based line importance calculation.

C. Protection Performance

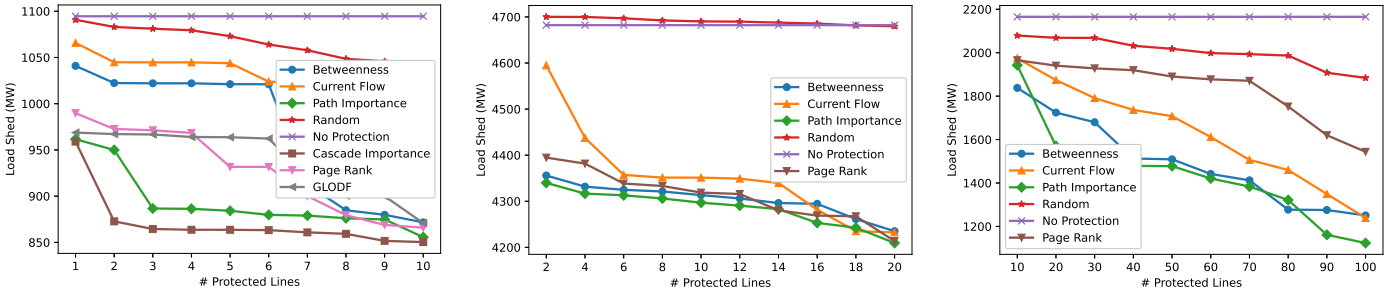
To evaluate protection performance of the algorithms, we randomly generate another collection of $20,000$ sets of initial failures for each network and the number of failed lines in each set is between 3 and 8. We then simulate cascading failure using the same collection of initial failure sets against all algorithms. Note that for each algorithm, we double the capacity of the identified lines, hence the cascading behavior will be different for different algorithms even though the initial failures are the same. The performance is measured using two metrics, average number of failed lines and average load shed over the set of cascading failures and the results are summarized in Table II. It is clear that Path Importance has better performance than almost all other algorithms according to both metrics. The only exceptions are Cascade Importance and GLODF. Cascade Importance chooses lines to protect according stats from complete N-2 and N-3 cascading failure results, which is reasonable to be better than other algorithms. GLODF can be seen as an approximation of Cascade Importance, as it calculates line flow changes without solving power flow equations. Hence, the performance of GLODF is inferior to Cascade Importance, but quite similar to Path Importance. However, even for the small IEEE-118 bus system, both Cascade Importance and GLODF need to consider more than a million cascades for $N-2$ and $N-3$ scenarios, which would not scale. Also, it's worth noting that having protection may result in more failed lines than having no protection, like in the case of using edge current flow betweenness centrality to identify important lines in the IEEE 118-Bus System. However, average load shed does decrease when having protection.

Next, we vary the number of lines to protect. In each data set, we simulated cascading failure with top 10%, 20% ..., 100% of the K identified important lines protected. The results are summarized in Figures 6 and 7. From the results, the effectiveness of the proposed path based line importance metric is further proved, as it outperformed most of the baselines (except for cascade importance) regardless of the number of critical lines protected. What can also be observed from the figures is that the protection gain is the highest by



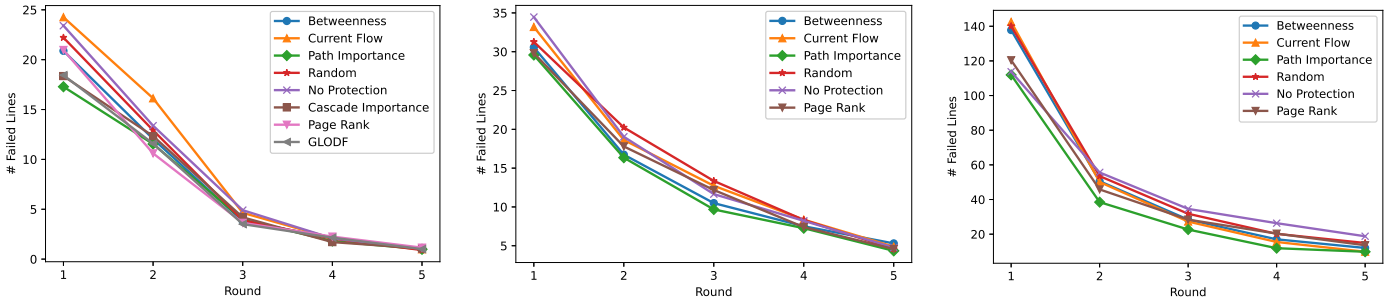
(a) IEEE 118 (b) IEEE 300 (c) Polish 3120

Fig. 6: Number of failed lines with 10%, 20%, ..., 100% of the identified critical lines protected



(a) IEEE 118 (b) IEEE 300 (c) Polish 3120

Fig. 7: Amount of load shed with 10%, 20%, ..., 100% of the identified critical lines protected



(a) IEEE 118 (with 10 Lines Protected) (b) IEEE 300 (with 20 Lines Protected) (c) Polish 3120 (with 100 Lines Protected)

Fig. 8: Number of failed lines per round

protecting the first 10%-20% of the critical lines and drops when protecting more lines.

We further break down the number of line failures by round of failure, in the scenario that K lines are protected in each data set. For readability, we only show the results from first 5 rounds, although in some cases a cascading failure can last for more than 15 rounds. Figure 8 demonstrates the correlation between all path based line importance and first round failures, since protection using that importance metric results in the least number of first round failures. However, number of failures for that importance metric may be higher than other algorithms for later rounds.

Identified Critical Lines: We list the idendified top-10 critical lines in the IEEE 118-Bus System by the four best-performing algorithms. Since the algorithms rank the lines by very different criteria, it is not surprising that there exists limited overlap among critical lines identified by different algorithms. However, the lines are all critical to some extent, so the protection performances are similar.

Running Time: We are also interested in the running time of the best-performing algorithms: Path Importance, Cascade Importance, PageRank and GLODF. The other simple heuristics run faster but didn't perform well, so they are ignored from the analysis. Since the running time of the algorithms are all highly dependent on the parameters, having a thorough comparison of their running time can be hard due to the large parameter space. Here, we only compare the running times of the algorithms with the parameters used for the experiments

TABLE II: Protection Performance Summary

Data Set	Algorithm	Avg. # Failed Lines	Avg. Load Shed (MW)
118-Bus	Betweenness	47.34	871.66
	Current Flow	54.70	1020.61
	Path Importance	43.79	855.67
	Random	48.51	1035.96
	Cascade Importance	**42.84**	**850.31**
	Page Rank	45.43	865.84
	GLODF	43.68	870.76
	No Protection	52.08	1094.56
300-Bus	Betweenness	84.21	4235.07
	Current Flow	90.94	4232.40
	Path Importance	**80.73**	**4209.37**
	Random	90.23	4679.75
	Page Rank	83.98	4213.56
	No Protection	91.45	4682.13
3120-Bus	Betweenness	300.14	1250.80
	Current Flow	286.89	1239.18
	Path Importance	**276.39**	**1123.50**
	Random	364.32	1883.98
	Page Rank	293.53	1543.52
	No Protection	399.62	2165.25

TABLE III: Top-10 Critical Lines in IEEE 118-Bus System

Rank	Path Importance	Cascade Importance	PageRank	GLODF
1	L81	L111	L38	L12
2	L31	L73	L7	L115
3	L161	L31	L96	L120
4	L117	L30	L8	L124
5	L111	L32	L36	L14
6	L5	L15	L51	L117
7	L62	L148	L9	L123
8	L40	L158	L33	L5
9	L104	L165	L31	L108
10	L50	L143	L29	L109

above. For Path Importance, we report its running time both with and without pre-processing, where pre-processing means to pre-generate all the samples pre-calculate all $\tilde{f}^{sd}_{l\neg l'}$ values. This is to simulate the scenario that the critical lines need to be found again when power network supplies/demands change. For all other algorithms, pre-processing is not possible and the running time of this scenario is not different from finding the critical lines of a new network, so we only report one time per algorithm per data set.

From Table IV, we can see that Path Importance without pre-processing is faster than Cascade Importance and GLODF, which shows that sampling paths is more efficient than extensively solving linear systems in the case of Cascade Importance (the power flow system) and GLODF (the linear system for GLODF calculation). Also, path sampling is highly parallelizable, so the running time can be much shorter in practice with multi-threading. PageRank is much faster than the above three since it only involves LODF calculation from N-1 failures and matrix operations. The fastest is Path Importance with pre-processing, which is at least 17 times faster than PageRank, showing it's advantage to avoid most of the recalculation when supplies/demands change.

TABLE IV: Running Times (in Seconds)

	Datasets		
Algorithms	118-Bus	300-Bus	3120-Bus
Path Importance (with pre-processing)	0.003	0.026	8.22
Path Importance (without pre-processing)	23,544	69,426	431,146
Cascade Importance	148,932	N/A	N/A
PageRank	2.85	4.01	141.16
GLODF	134,136	N/A	N/A

VI. Conclusion

In this paper, we studied the cascading failure protection problem from a new angle: identify the lines that are the most likely to fail after unintentional initial failures. We first proposed an all path based line importance metric and illustrated that there is a strong correlation between lines with high values of the metric and lines who are likely to fail after unintentional initial failures. Then, we designed a path sampling algorithm to estimate the proposed metric efficiently with theoretically bounded error margin. The proposed metric was evaluated in a protection scenario using multiple IEEE power system test cases and we demonstrated that the proposed metric outperformed several baseline metrics.

Acknowledgement

This work was supported by NSF CNS-1948550.

References

[1] Edward A Lee. Cyber physical systems: Design challenges. In *2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC)*, pages 363–369. IEEE, 2008.

[2] Wenye Wang and Zhuo Lu. Cyber security in the smart grid: Survey and challenges. *Computer networks*, 57(5):1344–1371, 2013.

[3] Andrey Bernstein, Daniel Bienstock, David Hay, Meric Uzunoglu, and Gil Zussman. Power grid vulnerability to geographically correlated failures—analysis and control implications. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pages 2634–2642. IEEE, 2014.

[4] Hassan Haes Alhelou, Mohamad Esmail Hamedani-Golshan, Takawira Cuthbert Njenda, and Pierluigi Siano. A survey on power system blackout and cascading events: Research motivations and challenges. *Energies*, 12(4):682, 2019.

[5] Pouyan Pourbeik, Prabha S Kundur, and Carson W Taylor. The anatomy of a power grid blackout-root causes and dynamics of recent major blackouts. *IEEE Power and Energy Magazine*, 4(5):22–29, 2006.

[6] Junjian Qi, Kai Sun, and Shengwei Mei. An interaction model for simulation and mitigation of cascading failures. *IEEE Transactions on Power Systems*, 30(2):804–819, 2014.

[7] Subhankar Mishra, Xiang Li, Alan Kuhnle, My T Thai, and Jungtaek Seo. Rate alteration attacks in smart grid. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2353–2361. IEEE, 2015.

[8] Subhankar Mishra, Xiang Li, Tianyi Pan, Alan Kuhnle, My T Thai, and Jungtaek Seo. Price modification attack and protection scheme in smart grid. *IEEE Transactions on Smart Grid*, 8(4):1864–1875, 2016.

[9] Kaarthik Sundar, Mallikarjuna Vallem, Russell Bent, Nader Samaan, Bharat Vyakaranam, and Yury Makarov. Nk failure analysis algorithm for identification of extreme events for cascading outage pre-screening process. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2019.

[10] Andrés Delgadillo, José Manuel Arroyo, and Natalia Alguacil. Analysis of electric grid interdiction with line switching. *IEEE Transactions on Power Systems*, 25(2):633–641, 2009.

[11] Subhankar Mishra, Xiang Li, My T Thai, and Jungtaek Seo. Cascading critical nodes detection with load redistribution in complex systems. In *International Conference on Combinatorial Optimization and Applications*, pages 379–394. Springer, 2014.

[12] Jungtaek Seo, Subhankar Mishra, Xiang Li, and My T Thai. Catastrophic cascading failures in power networks. *Theoretical Computer Science*, 607:306–319, 2015.

[13] Bin Liu, Zhen Li, Xi Chen, Yuehui Huang, and Xiangdong Liu. Recognition and vulnerability analysis of key nodes in power grid based on complex network centrality. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(3):346–350, 2017.

[14] Dirk Witthaut, Martin Rohden, Xiaozhu Zhang, Sarah Hallerberg, and Marc Timme. Critical links and nonlocal rerouting in complex supply networks. *Physical review letters*, 116(13):138701, 2016.

[15] Hrudaya Manjari Dola and Badrul H Chowdhury. Intentional islanding and adaptive load shedding to avoid cascading outages. In *2006 IEEE power engineering society general meeting*, pages 8–pp. IEEE, 2006.

[16] Erick E Aponte and J Keith Nelson. Time optimal load shedding for distributed power systems. *IEEE Transactions on Power Systems*, 21(1):269–277, 2006.

[17] Paul DH Hines, Ian Dobson, Eduardo Cotilla-Sanchez, and Margaret Eppstein. "Dual Graph" and "Random Chemistry" methods for cascading failure analysis. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 2141–2150. IEEE, 2013.

[18] Sudha Gupta, Ruta Kambli, Sushama Wagh, and Faruk Kazi. Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework. *IEEE Transactions on Industrial Electronics*, 62(4):2478–2486, 2014.

[19] Yang Yang, Takashi Nishikawa, and Adilson E Motter. Small vulnerable sets determine large network cascades in power grids. *Science*, 358(6365):eaan3184, 2017.

[20] Tianyi Pan, Alan Kuhnle, Xiang Li, and My Thai. Vulnerability of interdependent networks with heterogeneous cascade models and timescales. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pages 290–299. IEEE, 2018.

[21] Renjian Pi, Ye Cai, Yong Li, and Yijia Cao. Machine learning based on bayes networks to predict the cascading failure propagation. *IEEE Access*, 6:44815–44823, 2018.

[22] Zhiyuan Ma, Chen Shen, Feng Liu, and Shengwei Mei. Fast screening of vulnerable transmission lines in power grids: A pagerank-based approach. *IEEE Transactions on Smart Grid*, 10(2):1982–1991, 2017.

[23] Saleh Soltan, Dorian Mazauric, and Gil Zussman. Analysis of failures in power grids. *IEEE Transactions on Control of Network Systems*, 2015.

[24] Quanyan Zhu and Tamer Basar. Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems. *IEEE Control Systems Magazine*, 35(1):46–65, 2015.

[25] Maggie X Cheng, Mariesa Crow, and Quanmin Ye. A game theory approach to vulnerability analysis: Integrating power flows with topological analysis. *International Journal of Electrical Power & Energy Systems*, 82:29–36, 2016.

[26] Weixian Liao, Sergio Salinas, Ming Li, Pan Li, and Kenneth A Loparo. Cascading failure attacks in the power system: a stochastic game perspective. *IEEE Internet of Things Journal*, 4(6):2247–2259, 2017.

[27] Qin Ba and Ketan Savla. A dynamic programming approach to optimal load shedding control of cascading failure in dc power networks. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3648–3653. IEEE, 2016.

[28] Sakshi Pahwa, C Scoglio, Sanjoy Das, and N Schulz. Load-shedding strategies for preventing cascading failures in power grid. *Electric Power Components and Systems*, 41(9):879–895, 2013.

[29] Mahshid Rahnamay-Naeini, Zhuoyao Wang, Nasir Ghani, Andrea Mammoli, and Majeed M Hayat. Stochastic analysis of cascading-failure dynamics in power grids. *IEEE Transactions on Power Systems*, 29(4):1767–1779, 2014.

[30] Jiajia Song, Eduardo Cotilla-Sanchez, Goodarz Ghanavati, and Paul DH Hines. Dynamic modeling of cascading failure in power systems. *IEEE Transactions on Power Systems*, 31(3):2085–2095, 2015.

[31] Fan Wenli, Liu Zhigang, Hu Ping, and Mei Shengwei. Cascading failure model in power grids using the complex network theory. *IET Generation, Transmission & Distribution*, 10(15):3940–3949, 2016.

[32] Ulrik Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008.

[33] Ulrik Brandes and Daniel Fleischer. Centrality measures based on current flow. In *Annual symposium on theoretical aspects of computer science*, pages 533–544. Springer, 2005.

[34] T Athay, R Podmore, and S Virmani. A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems*, (2):573–584, 1979.

[35] IEEE 39-Bus System. https://icseg.iti.illinois.edu/ieee-39-bus-system/. Accessed: 2020-12-22.

[36] Ben Roberts and Dirk P Kroese. Estimating the number of st paths in a graph. *J. Graph Algorithms Appl.*, 11(1):195–214, 2007.

[37] Alan Kuhnle, Tianyi Pan, Victoria G Crawford, Md Abdul Alim, and My T Thai. Pseudo-separation for assessment of structural vulnerability of a network. *ACM SIGMETRICS Performance Evaluation Review*, 45(1):13–14, 2017.

[38] Leslie G Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[39] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[40] Leon Thurner, Alexander Scheidler, Florian Schäfer, Jan-Hendrik Menke, Julian Dollichon, Friederike Meier, Steffen Meinecke, and Martin Braun. pandapower—an open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Transactions on Power Systems*, 33(6):6510–6521, 2018.

[41] Shengwei Mei, Fei He, Xuemin Zhang, Shengyu Wu, and Gang Wang. An improved opa model and blackout risk assessment. *IEEE Transactions on Power Systems*, 24(2):814–823, 2009.

[42] Teoman Guler, George Gross, and Minghai Liu. Generalized line outage distribution factors. *IEEE Transactions on Power systems*, 22(2):879–881, 2007.

[43] Paul Hines, Eduardo Cotilla-Sanchez, and Seth Blumsack. Do topological models provide good information about electricity infrastructure vulnerability? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 20(3):033122, 2010.

Xiang Li (M'18) is an Assistant Professor at the Department of Computer Science and Engineering of Santa Clara University. She received her Ph.D. degree in Computer and Information Science and Engineering department of the University of Florida. Her research interests are centered on the large-scale optimization and its intersection with cyber-security of networking systems, big data analysis, and cyber physical systems. She has published 32 articles in various prestigious journals and conferences such as IEEE Transactions on Mobile Computing, IEEE IEEE Transactions on Smart Grids, IEEE INFOCOM, IEEE ICDM, including one Best Paper Award in IEEE MSN 2014, Best Paper Nominee in IEEE ICDCS 2017, and Best Paper Award in IEEE International Symposium on Security and Privacy in Social Networks and Big Data 2018. She is an associate editor of the *Computational Social Networks* journal and the *Journal of Combinatorial Optimization*.

Tianyi Pan received his Ph.D. degree in computer engineering from the University of Florida. His research focuses on approximation algorithms of optimization problems and vulnerability analysis in interdependent networks, including online social networks, smart grid and communication networks.

Kai Pan (S'13-M'16) received the B.S. degree in industrial engineering from Zhejiang University, Hangzhou, China, in 2010, and the M.S. and Ph.D. degrees in industrial and systems engineering from the University of Florida, Gainesville, FL, USA, in 2014 and 2016, respectively. He is currently an assistant professor in the Department of Logistics and Maritime Studies at the Hong Kong Polytechnic University, Hung Hom, Hong Kong. His research interests include stochastic integer programming, data-driven optimization, and energy system applications.