Link Disruption Scenario Generation for Transportation Network Criticality Analysis

A Technical Report Submitted to the Rural Safe Efficient Advanced Transportation (R-SEAT) Center and United States Department of Transportation

FINAL REPORT

Principal Investigator:
Anil Yazici, Ph.D.
Associate Professor
Department of Civil Engineering
Stony Brook University
2425 Computer Science,
Stony Brook, NY 11790
Phone: (631) 632-9349
E-mail: Anil.Yazici@stonybrook.edu

Research Assistant:
Amir Hossein Ali Khan
Research Assistant
Department of Civil Engineering
Stony Brook University
1212 Computer Science,
Stony Brook, NY 11790
E-mail: amirhossein.alikhan@stonybrook.edu

February 2025

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, under the grant 69A3552348321 from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Link Disruption Scenario Generation for Transportation Network Criticality Analysis		5. Report Date 6. Performing Organization Code
7. Author(s) Anil Yazici (https://orcid.org/0000-0001-9613-0464) Amir Hossein Ali Khan (https://orcid.org/0009-0002-6199-9471)		8. Performing Organization Report No.
9. Performing Organization Name and Address Stony Brook University 100 Nicolls Road, Stony Brook, NY 11794 Tel: (631) 632-6000		10. Work Unit No. (TRAIS) 11. Contract or Grant No. 69A3552348321
12. Sponsoring Agency Name and Address Rural Safe Efficient Advanced Transportation (R-SEAT) Center 2525 Pottsdamer Street Tallahassee, FL 32310		13. Type of Report and Period Covered Final Report Period Covered: 03/01/2024 – 02/28/2025 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Transportation network criticality analysis is a crucial step for prioritization of network improvements to create a more robust and resilience transportation network, The common approach in link criticality analysis is to assess the performance of the network under varying disruption scenarios that include link failure(s). Simultaneous multi-link disruptions can capture network interactions and travelers' rerouting responses that are overlooked by conventional single-link analysis. However, the combinatorial growth of possible disruption scenarios makes exhaustive evaluation computationally infeasible. This study develops a computationally efficient sampling framework for approximating network-wide individual-link criticality rankings using a substantially reduced set of multi-link disruption scenarios. For this purpose, a large scenario pool combining random and statistically controlled non-random scenarios were first generated to establish stable Pseudo Ground Truth (Pseudo-GT) rankings that are assumed to estimate the actual link criticality rankings. Then a disruption scenario generation procedure was developed and evaluated across the Sioux Falls, Eastern Massachusetts, and Anaheim networks. Results show that random sampling consistently produces rankings with higher Spearman correlations with the Pseudo-GT ranking at smaller sample sizes than non-random sampling. The greater representation of higher-order disruptions in random sampling also suggests that scenarios involving more simultaneous failures can capture link interactions more efficiently. Independent evaluation on the Berlin-Friedrichshain network further supported the applicability of the proposed guidelines. Overall, the framework substantially reduces the computational effort required to obtain network-wide link criticality rankings while accounting for multi-link interactions.

17. Key Words	18. Distribution Statement No restrictions		
19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price

ACKNOWLEDGEMENTS

This project was sponsored by the Rural Safe Efficient Advanced Transportation (R-SEAT) Center and United States Department of Transportation. The Principal Investigators would like to thank the representatives of the R-SEAT Center for their valuable feedback throughout the project activities.

Table of Contents

List of Tables

List of Figures

ABSTRACT

Transportation network criticality analysis is a crucial step for prioritization of network improvements to create a more robust and resilience transportation network, The common approach in link criticality analysis is to assess the performance of the network under varying disruption scenarios that include link failure(s). Simultaneous multi-link disruptions can capture network interactions and travelers' rerouting responses that are overlooked by conventional single-link analysis. However, the combinatorial growth of possible disruption scenarios makes exhaustive evaluation computationally infeasible. This study develops a computationally efficient sampling framework for approximating network-wide individual-link criticality rankings using a substantially reduced set of multi-link disruption scenarios. For this purpose, a large scenario pool combining random and statistically controlled non-random scenarios were first generated to establish stable Pseudo Ground Truth (Pseudo-GT) rankings that are assumed to estimate the actual link criticality rankings. Then a disruption scenario generation procedure was developed and evaluated across the Sioux Falls, Eastern Massachusetts, and Anaheim networks. Results show that random sampling consistently produces rankings with higher Spearman correlations with the Pseudo-GT ranking at smaller sample sizes than non-random sampling. The greater representation of higher-order disruptions in random sampling also suggests that scenarios involving more simultaneous failures can capture link interactions more efficiently. Independent evaluation on the Berlin-Friedrichshain network further supported the applicability of the proposed guidelines. Overall, the framework substantially reduces the computational effort required to obtain network-wide link criticality rankings while accounting for multi-link interactions.

1. INTRODUCTION

Transportation network criticality analysis seeks to identify and rank network components according the consequences of their degradation or failure. Policymakers need the criticality ranking of transportation network links to support investment and infrastructure improvement decisions. Link criticality is assessed using selected criticality metrics that quantify variations in system performance under different disruption scenarios. A common approach is to evaluate individual link criticality by considering single-link degradation or failure, i.e., running traffic assignment for each link removal scenario, calculating the difference in a network performance measure (e.g., total system travel time) between the unaffected network and the failure scenario, and ranking the links based on the resulting functionality loss. In this approach, the greater the functionality loss, the higher the criticality ranking of the link. However, under disaster conditions such as hurricanes and snowstorms, multiple links are more likely to fail simultaneously. Consequently, single-link removals do not adequately capture the actual criticality of network links because of network dependencies. For example, a link that may not cause a substantial deterioration in network performance when removed individually, and therefore may not be identified as critical, can nevertheless cripple the transportation system when it fails in conjunction with other links [1], [2], [3], [4], [5], [6]. Therefore, multiple simultaneous link removal scenarios are required to capture the interactions among network links in the criticality analysis.

Although considering multiple simultaneous link failures provides a more realistic assessment of network criticality, it also introduces a significant computational challenge. Because disruption scenarios are defined by combinations of failed links, the number of possible scenarios grows combinatorially with network size. As a result, exhaustively evaluating all possible multiple-link failure scenarios to determine the true ground-truth criticality ranking quickly becomes computationally infeasible. For example, even a relatively small network with only 50 links has ($2^{50} - 1 = 1{,}125{,}899{,}906{,}842{,}624$) possible link-failure combinations, each requiring a separate traffic assignment analysis. Although not all combinations are practically meaningful (e.g., scenarios involving the failure of nearly every link or resulting in disconnected networks), the number of realistic disruption scenarios remains prohibitively large. Consequently, obtaining the true ground-truth criticality ranking for real-world transportation networks is computationally infeasible.

To address this challenge, this study develops a computationally efficient sampling framework for transportation network criticality analysis that approximates the ground-truth link criticality ranking using only a fraction of the disruption scenarios required by exhaustive evaluation. To achieve this objective, the proposed framework consists of two stages. First, to overcome the lack of the ground-truth ranking as a benchmark, a sufficiently large and representative disruption scenario pool is generated to derive a stable reference ranking, referred to as the Pseudo Ground Truth (Pseudo-GT) ranking. Although generating the Pseudo-GT ranking is itself computationally intensive, once its robustness and stability are established, it provides the benchmark needed to evaluate subsamples in the second stage. The second stage employs a stepwise sampling procedure

with a stopping rule to identify representative subsamples whose resulting link criticality rankings closely approximate the Pseudo-GT ranking. The characteristics of these high-performing subsamples are then analyzed to develop practical sampling guidelines.

This proposed framework is evaluated on three benchmark transportation networks with distinct topological characteristics selected from the Transportation Network Online Repository [7]: Sioux Falls, Eastern Massachusetts, and Anaheim. Based on the results obtained from these benchmark networks, practical sampling guidelines are developed. To evaluate the generalizability of these guidelines, they are subsequently applied and validated on the Berlin-Friedrichshain transportation network.

The remainder of this report is organized as follows. The next section reviews the relevant literature on transportation network criticality analysis under multiple-link disruption scenarios and identifies the research gap addressed in this study. The methodology section presents the proposed sampling framework, including the generation of the representative disruption scenario pool and the stepwise subsampling procedure. The results and discussion section evaluates the proposed framework, presents the resulting findings, develops practical sampling guidelines, and validates their applicability using the Berlin-Friedrichshain network.

2. LITERATURE REVIEW

2.1. Limitations of Single-Link Disruption Analysis

Although numerous studies have employed single-link disruption scenarios to assess transportation link criticality, this approach provides a limited network-wide perspective because it cannot capture the interdependencies and flow interactions among simultaneously disrupted links. A link that causes only a minor performance loss when disrupted independently may become highly consequential when it fails in combination with other links [1], [2], [3], [4], [5], [6]. Consequently, single-link disruption procedures cannot fully represent the interaction effects arising from simultaneous link failures.

2.2. Hazard-Informed and Area-Based Multi-Link Disruptions

Recognizing these limitations, a growing body of literature has examined the simultaneous degradation of multiple links. Because enumerating all possible multi-link disruption scenarios is computationally infeasible, several studies have sought to reduce the scenario space. For example, Bagloee et al. [4] restricted the scenario space by identifying a limited set of highly used candidate roads, within which a selected number of high-demand roads are identified as “candidate” critical links.

Beyond restricting the set of candidate links, researchers have reduced the scenario space by focusing on hazard-specific or area-covering disruptions. Jenelius and Mattsson [8] developed a grid-based approach in which the study area is divided into spatial cells and all road links intersecting a given cell are assumed to fail simultaneously. Similarly, Erath and Axhausen [9] constrained the search according to the geographic extent of the hazard-affected area and employed

optimization heuristics to identify combinations of failed links that produce the greatest increase in user costs. Focusing on earthquake risk and probabilities, Haghighi et al. [10] developed a probabilistic multi-scenario framework that integrates Monte Carlo simulation, network-wide traffic assignment, and regression analysis to rank critical links according to their damage probabilities and contributions to travel-time increases across multiple earthquake-disruption scenarios. In the context of flood risk and probabilities, Sullivan et al. [11] developed a hazard-informed Network Criticality Index that combines link-specific flood probabilities, Monte Carlo-generated multi-link disruption scenarios, and traffic-assignment outcomes to rank individual road segments. Johnson et al. [12] integrated area-based freight-routing simulations, gradient boosting, and partial dependence plots to map spatial vulnerability under simultaneous disruptions of geographically correlated road and railway links. Finally, Yang et al. [6] proposed a mixed-integer optimization framework that avoids explicit scenario enumeration and identifies worst-case and near-worst-case combinations of area-covering disruptions based on network connectivity and route redundancy.

Although these approaches provide systematic mechanisms for generating hazard-informed or area-based multi-link disruption scenarios, their emphasis on spatially concentrated failures overlooks consequential disruptions involving geographically dispersed links. Sohouenou et al. [13] found that localized failures were only slightly more damaging than randomly distributed failures, with differences in mean robustness ranging from 1.6 to 4.3 percentage points. Furthermore, Wang et al. [3] and Jin et al. [5] demonstrated that the links forming a critical combination are not necessarily connected or even located near one another. These findings establish the need for hazard-independent approaches capable of identifying consequential link combinations regardless of their spatial proximity.

2.3. Hazard-Independent Identification of Critical Link Combinations

Accordingly, several studies have examined geographically unconstrained multi-link disruptions, generally with the objective of identifying the most damaging combinations of failed links. Sohouenou et al. [13], for example, adopted a hazard-independent approach to evaluate all 584,934 combinations of one to five simultaneous bidirectional link failures in the Sioux Falls network using user-equilibrium traffic assignment. To reduce the computational burden of exhaustive enumeration, they proposed an iterative heuristic for identifying the most damaging link combinations. Guo et al. [14] developed a graph-autoencoder-based importance-sampling method that learns link importance from topological and traffic-flow attributes and biases scenario generation toward rare, extreme multi-link failure combinations. Pursuing the same objective, Wang et al. [3] proposed an enumeration-free global optimization framework; Jin et al. [5] formulated the identification of critical road combinations as a bilevel mixed-integer nonlinear programming problem; Gu et al. [15] developed a combinatorial optimization approach; Gu et al. [16] introduced an enumeration-free random-key genetic algorithm; and Bhavathrathan and Patil [17] formulated a minimax optimization model. Tu et al. [18] developed a path-flow-weighted betweenness-centrality measure that integrates network topology and traffic flows to identify the top-K critical multi-link combinations. Finally, Jiang et al. [1] developed a probabilistic bilevel

optimization framework that integrates partial capacity degradation, disruption probabilities, and user-equilibrium traffic assignment to identify the simultaneous multi-link failure scenario with the greatest network impact. Collectively, however, these studies primarily aim to identify the most damaging combinations of simultaneous link disruptions rather than rank individual links while accounting for joint disruption effects.

2.4. Individual-Link Ranking Under Multi-Link Disruptions

In contrast to this combination-oriented objective, a smaller body of research has used simultaneous link-disruption scenarios to produce network-wide rankings of individual links while accounting for joint disruption effects. Barati et al. [19] developed an enumeration-free integrated MLP–CNN model that uses network topology and link-flow information to predict network-wide individual-link criticality rankings. However, the final ranking reliability depends on the representativeness of the networks and disruption scenarios used for training. For each network in the training set, the study employed a scenario pool of (25n), considered disruption orders ranging from 2 to 8, and imposed a minimum link-representation threshold of 2%. These settings were adopted as rules of thumb rather than validated within the study. Whereas Barati et al. [19] predicted network-wide individual-link criticality rankings using both traffic-flow and topological information, Jana et al. [20] developed an edge-based graph neural network to approximate road-link rankings based primarily on edge betweenness under travel-time perturbations and the random simultaneous removal of up to 1% of network links. Although the method enables rapid ranking updates for large disrupted networks, the resulting rankings primarily reflect network topology and shortest-path characteristics rather than incorporating travel demand and travelers' rerouting behavior.

2.5. Congestion-Induced Cascading Disruptions

Related research has examined interactions among multiple link disruptions through congestion-induced cascading processes rather than through predefined simultaneous failure scenarios. In the study by Chen et al. [21], link disruptions emerge endogenously through a recursive traffic-percolation process in which congestion-impaired links are successively removed and traffic is redistributed. This redistribution may overload additional links and thereby trigger cascading failures. In another study, Owais and Ramadan [22] integrated Monte Carlo demand simulation, stochastic user-equilibrium assignment, stacked-autoencoder modeling, and global sensitivity analysis to rank links according to their direct and nonlinear effects on average user delay. Although their deep-learning framework substantially reduces the computational burden, the resulting rankings represent demand-induced flow sensitivity rather than link occurrence and network-performance consequences under explicit multi-link failure scenarios.

2.6. Research Gap and Study Contribution

Taken together, the reviewed studies address the computational complexity of network criticality analysis by restricting candidate links or disruption scenarios, focusing on hazard-specific or spatially concentrated events, searching only for the most damaging link combinations, or

replacing repeated traffic assignments with surrogate models. Related cascading and demand-sensitivity approaches capture additional forms of network interaction but do not directly address the sampling of explicit multi-link disruption scenarios for network-wide individual-link ranking. Although these approaches serve important purposes, they provide limited guidance on how simultaneous multi-link disruption scenarios should be sampled when the objective is to obtain a stable, network-wide ranking of individual links that captures their interactions with other disrupted links. Overall, the literature has not systematically determined how disruption scenarios should be allocated across failure orders or what sample size or sampling method is efficient to produce a stable individual-link criticality ranking. Accordingly, this study seeks to develop a computationally efficient procedure for sampling a representative subset of the multi-link disruption scenario space that preserves interaction effects and reliably approximates the network-wide criticality ranking of individual links.

3. METHODOLOGY

The objective of this study is to develop a computationally efficient sampling framework for transportation network criticality analysis that accurately approximates link criticality ranking using only a small fraction of the disruption scenarios required by exhaustive evaluation. This objective is achieved in two main steps:

I. Disruption Scenario Pool and Pseudo-Ground Truth Ranking

First, to overcome the absence of the ground-truth ranking as a benchmark, a stable reference ranking, referred to as the *Pseudo Ground Truth (Pseudo-GT)* ranking, is derived from a *large disruption scenario pool*. The disruption scenario pool is assumed to be representative because it contains a sufficiently large number of scenarios and captures interactions among network links by incorporating disruption scenarios with varying orders of simultaneous link failures. This assumption is subsequently validated through stability analyses of the resulting Pseudo-GT ranking. Since generating the Pseudo-GT ranking is computationally intensive, once its robustness and stability are established, it can serve as the benchmark for evaluating the representativeness of the subsamples generated in the second stage. Section 3.2 provides a detailed description of the disruption scenario generation procedure.

II. Subsampling of Disruption Scenario Pool:

To identify a computationally efficient subsample (i.e., one that produces a ranking highly correlated with Pseudo-GT while requiring as low subsample size as possible) from the disruption scenario pool, a stepwise sampling procedure with a stopping rule is formulated. At each run, batches of (S) disruption scenarios (with (S) varying according to network size) are incrementally sampled from the scenario pool, and the criticality ranking is updated after each addition. The correlation between two consecutive updated rankings is calculated at each step. Once this correlation reaches a predetermined, high stopping threshold, additional sampling is considered unlikely to produce meaningful changes in the estimated ranking. This point is referred to as the stopping point. At the stopping point, the correlation between the subsample ranking and the

Pseudo-GT ranking is calculated as a measure of the subsample's representativeness in terms of the produced link criticality ranking. To account for sampling variability, the procedure is repeated 200 times using different randomly selected scenarios from the disruption scenario pool. The characteristics of the resulting high-performing subsamples (e.g., stopping sample size, disruption-order composition, etc.) form the basis for the development of the practical sampling guidelines. Additional details of this procedure are provided in Section 3.3.

3.1. Experimental Setup:

The following experimental setup were used to test the performance of generated disruption scenarios:

- The agreement between rankings was evaluated using Spearman's rank correlation coefficient, which is based on differences in rank positions and penalizes larger deviations more strongly. For example, comparing (1, 2, 3, 4) with (1, 3, 2, 4) versus (1, 4, 3, 2), the latter yields a lower correlation due to larger rank differences.
- The Efficiency Index metric proposed by [23] is used as the criticality metric in this study since it incorporates travel demand and travelers' rerouting behavior and handles dysconnectivity caused by simultaneous link failures. However, the proposed procedure is metric-independent, and the same analytical steps can be applied to other criticality metrics without the loss of generality.
- Since there are many disruption scenarios that include multiple links, the criticality score for each scenario relates to the criticality of all the links involved in the scenario rather than the individual links. Hence, the criticality ranking cannot be obtained directly by sorting criticality scores because each link has a distribution of criticality scores coming from all the scenarios that the link was in. Accordingly, a distribution-based method proposed by Barati et al. [24] was employed to compute the criticality ranking for individual links. In this method, links are first ordered according to the mean of their criticality score distributions. Consecutive links are then compared using the Anderson–Darling test to determine whether their distributions differ significantly. If a significant difference is detected, the link with the higher mean is assigned the higher rank. Otherwise, skewness and the coefficient of variation are used as tie-breaking criteria or to assign equal ranks.
- Each stage of the analysis was conducted on three case-study transportation networks, Sioux Falls, Eastern Massachusetts, and Anaheim, selected from the Transportation Network Online Repository [7]. These networks were chosen because of their distinct topological characteristics and sizes. Figure 1 presents the topological representations of these networks. The disruption scenario pool sizes (B) assigned to the Sioux Falls, Eastern Massachusetts, and Anaheim test networks are considered as 460K, 780K, and 500K, respectively.

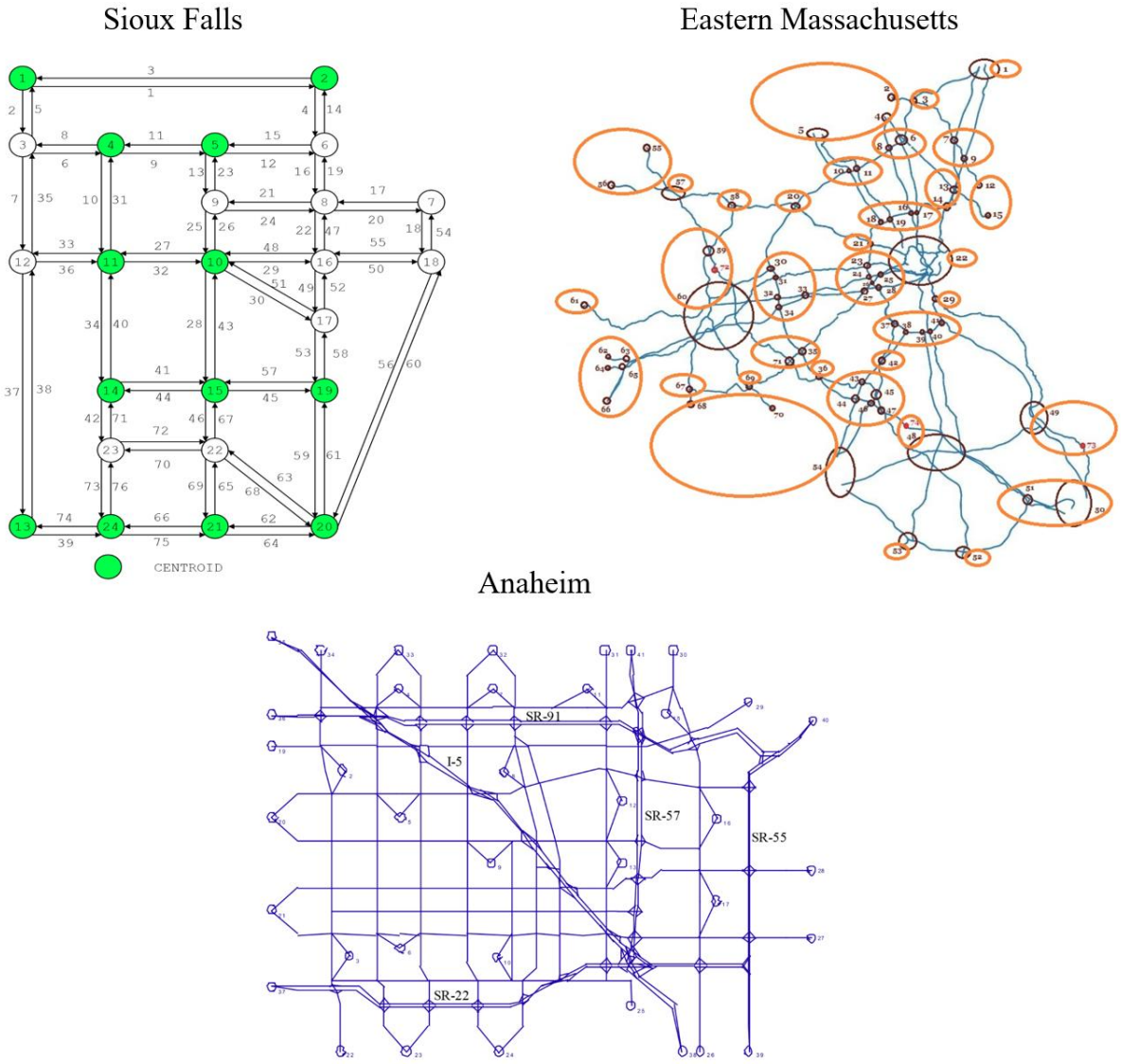

Figure 1 - Topological Representation of Test Networks

3.2. Generation of the Representative Disruption Scenario Pool and Pseudo-Ground Truth Ranking

To enable a computationally efficient approximation of the ground-truth link criticality ranking, the first stage of the proposed framework constructs a representative disruption scenario pool from which an approximate reference ranking, referred to as the Pseudo Ground Truth (Pseudo-GT) ranking, is derived. The Pseudo-GT ranking is assumed to closely approximate the actual ground-truth ranking because it is derived from a sufficiently large disruption scenario pool. To ensure that the disruption scenario pool is representative, it also should capture interactions among

network links by incorporating disruption scenarios with varying orders of simultaneous link failures. Accordingly, the scenario pool includes disruption scenarios ranging from single link failures up to 10 simultaneous link failures, with each link appearing multiple times in combination with different links.

One straightforward approach for generating the disruption scenario pool is to create purely random generated disruption scenarios that include multiple link failures. However, purely random sampling tends to overrepresent high-order and lead to sparse or even omitted representation of lower-order disruption scenarios. To illustrate this, consider disruption scenarios involving 1 to 10 simultaneous link failures in the relatively small Sioux Falls network, which contains 76 links. The number of possible scenarios grows rapidly with the order of failure: there are 76 single-link removal scenarios, 2,850 two-link simultaneous failures, 70,300 three-link failures, and up to 954,526,728,530 scenarios involving ten simultaneous failures. Consequently, based on classical probability principles, the probability of randomly selecting lower-order removal scenarios becomes extremely small (e.g., $7 \times 10^{-8}$for three-link simultaneous failures in this case). As a result, these lower-order failures may be underrepresented or even omitted in the sampled pool. This imbalance makes it questionable to rely exclusively on random scenario generation and highlights the need to complement and compare random sampling with a non-random or statistically controlled sampling approach. This way, more adequate coverage across disruption orders can be achieved and the impact of removal order (i.e., how many links fail simultaneously) compositions can be revealed. Therefore, to incorporate both random and non-random sampling approaches within the disruption scenario pool and to enable a controlled comparison between them, a non-random sampling framework is proposed.

3.2.1. Non-Random Sampling Framework

In the proposed non-random/statistically controlled sampling approach, each link is required to appear at least (r) times within each disruption order under a specified level of statistical confidence. This requirement ensures that all disruption orders are adequately represented while simultaneously guaranteeing sufficient representation of each link within every disruption order.

To elaborate on this approach, consider a network with n links. The minimum number of link appearances in each disruption order is denoted by (r). At failure order k, each sampled scenario corresponds to the simultaneous failure of exactly k links, drawn uniformly from the n available links. The probability that a given link appears in a randomly selected order-k scenario is:

$$P = k/n$$

If n_k denotes the number of sampled scenarios at order k, then the number of times a particular link appears in these samples follows a binomial distribution:

$$X \sim \text{Binomial}(n_k, p)$$

For sufficiently large n_k, this binomial distribution can be approximated by a normal distribution. The requirement that each link appears at least r times with confidence level $1 - \alpha$ can therefore be expressed as:

$$P(X \geq r) \geq 1 - \alpha$$

Using the normal approximation:

$$\Pr\left(\frac{X-\mu}{\sigma} \geq \frac{r-\mu}{\sigma}\right) \approx 1 - \Phi\left(\frac{r-\mu}{\sigma}\right)$$

Let $z = \Phi^{-1}(1-\alpha)$. Then the above condition is equivalent to:

$$\Phi\left(\frac{r-\mu}{\sigma}\right) \leq \alpha \Leftrightarrow \frac{\mu - r}{\sigma} \geq z$$

With $\mu = n_k p$ and $\sigma = \sqrt{n_k p(1-p)}$, this becomes:

$$n_k p - z\sqrt{n_k p(1-p)} \geq r$$

Solving for n_k:

$$n_k = \left\lceil \left(\frac{z\sqrt{p(1-p)} + \sqrt{z^2 p(1-p) + 4pr}}{2p} \right)^2 \right\rceil$$

which is valid when $n_k p(1-p) \gtrsim 10$.

Although this expression guarantees that a single link is observed at least r times with probability $1 - \alpha$, this confidence level does not hold simultaneously for all n links. When multiple events are tested concurrently (one for each link), the probability that at least one link fails to meet the requirement increases.

To maintain a family-wise confidence level of $1 - \alpha$ across all links, the Bonferroni correction distributes the allowable error equally among them:

$$\Pr(X \geq r) \geq 1 - \frac{\alpha}{n}$$

which yields:

$$n_k = \left\lceil \left(\frac{z\sqrt{p(1-p)} + \sqrt{z^2 p(1-p) + 4pr}}{2p} \right)^2 \right\rceil$$

With $z = \Phi^{-1}\left(1 - \frac{\alpha}{n}\right)$.

In this study, half of the total disruption scenario pool is allocated to the non-random/statistically controlled sampling approach. Therefore, if the total disruption scenario pool size is B, total sample size for non-random sampling becomes B/2. Accordingly, the largest feasible value of r is determined. More specifically, the value of r is selected such that:

$$\Sigma\, n_k \leq B/2, \text{ for } k = 1, 2, ..., K$$

For example, consider the Sioux Falls network with n = 76 links, disruption orders ranging from k = 1 to 10, and a confidence level of 95% (α = 0.05). Suppose that the non-random sampling size is B/2 = 250,000 scenarios. Applying the Bonferroni correction yields:

$$\alpha / n \approx 6.7 \times 10^{-4}$$

which corresponds to:

$$z \approx 3.20$$

For each disruption order k, the corresponding probability becomes:

$$p = k / 76$$

The largest feasible repetition value is determined as:

$$r = 2{,}127$$

This means that, with the available non-random sampling size of 250,000 scenarios, disruption scenarios can be distributed across failure orders such that each link appears at least 2,127 times within every disruption order, while maintaining a 95% family-wise confidence level across all links. The resulting distribution of scenarios across disruption orders is presented in Table 3.

Table 1 - Non-Random Sampling: Distribution of Scenarios Across Disruption Orders (n = 76, B/2 = 250,000)

Order (k)	Number of scenarios (n_k)
1	76
2	2,850
3	57,691
4	43,248
5	34,582
6	28,805
7	24,678
8	21,583
9	19,176
10	17,250
Total	**249,939**

3.3. Subsampling of Representative Disruption Scenarios: Stepwise Sampling Procedure

Computing the Pseudo-GT ranking using the procedure described in the previous section requires running traffic assignment for a large number of disruption scenarios, making it computationally intensive. For example, on a computationally powerful workstation equipped with a 24-core Intel® Xeon® W5-2465X CPU (up to 4.8 GHz Turbo) and 64 GB of memory, traffic assignment runs with parallel processing for the Sioux Falls, Eastern Massachusetts, and Anaheim networks required approximately 1, 4, and 4 days, respectively. To improve the computational efficiency of criticality analysis, the second stage of the proposed framework aims to identify representative disruption scenario subsamples that closely approximate the Pseudo-GT ranking while minimizing the required computational effort. Accordingly, a stepwise sampling procedure with a stopping rule is employed, using the Pseudo-GT ranking as the benchmark for evaluating the representativeness of the resulting subsamples.

At each independent run of the subsampling procedure, S (varying based on network size) disruption scenarios are incrementally sampled at each step and the criticality ranking is updated accordingly (referred to as the *cumulative ranking*). At the end of each step, the correlation between the updated ranking at step m and that of the previous step $m-1$ is computed (referred to as the *cumulative correlation*). The process continues iteratively until the cumulative correlation exceeds a predetermined stopping threshold (e.g., a high Spearman correlation of 0.95). In other words, the procedure terminates once additional sampling no longer produces meaningful changes in the cumulative ranking. At the stopping point, the correlation between the stopping cumulative ranking and the Pseudo Ground Truth ranking is calculated, and the corresponding stopping sample size and link combinations in the subsample is recorded. This correlation provides a

measure of the ranking estimation power of the subsample. Note that a high ranking correlation between consecutive samples (i.e., cumulative correlation) does not necessarily indicate a high correlation between ranking at the stopping point with the pseudo-GT ranking. However, the difference in the scenario characteristics between subsamples that have higher and lower correlation with pseudo-GT provides important insights on how the representative samples should be.

To ensure randomness in the subsampling procedure and to compare the performance of random and non-random sampling approaches, the stepwise procedure is repeated independently 200 times using subsamples drawn only from the random scenario pool and another 200 times using subsamples drawn only from the non-random scenario pool. Figure 2 illustrates the overall workflow of the proposed stepwise subsampling procedure.

The stepwise sampling approach requires the following inputs:

- Transportation network size (i.e., total number of links n)
- The step size (S)
- The stopping correlation threshold (T)

Changes in step size (S) and stopping correlation threshold (T) can affect both the stopping sample size and the final correlation with the Pseudo-GT ranking. Therefore, to identify appropriate parameter values that improve both the accuracy of the resulting ranking and the associated computational effort, the following values were considered for the step size and stopping correlation threshold:

- **Step size:** Using a small step size can yield high cumulative correlation between consecutive samples, which leads to a premature termination of the subsampling procedure before the sample reach a high correlation with the pseudo-GT ranking. To address this, varying step sizes proportional to the network size ($S = n, 2n, 3n$) were considered. This allows the step size to scale with network size for a reliable convergence.
- **Stopping correlation threshold:** Three high stopping correlation thresholds were considered: $T = 0.90$, 0.95, and 0.98.

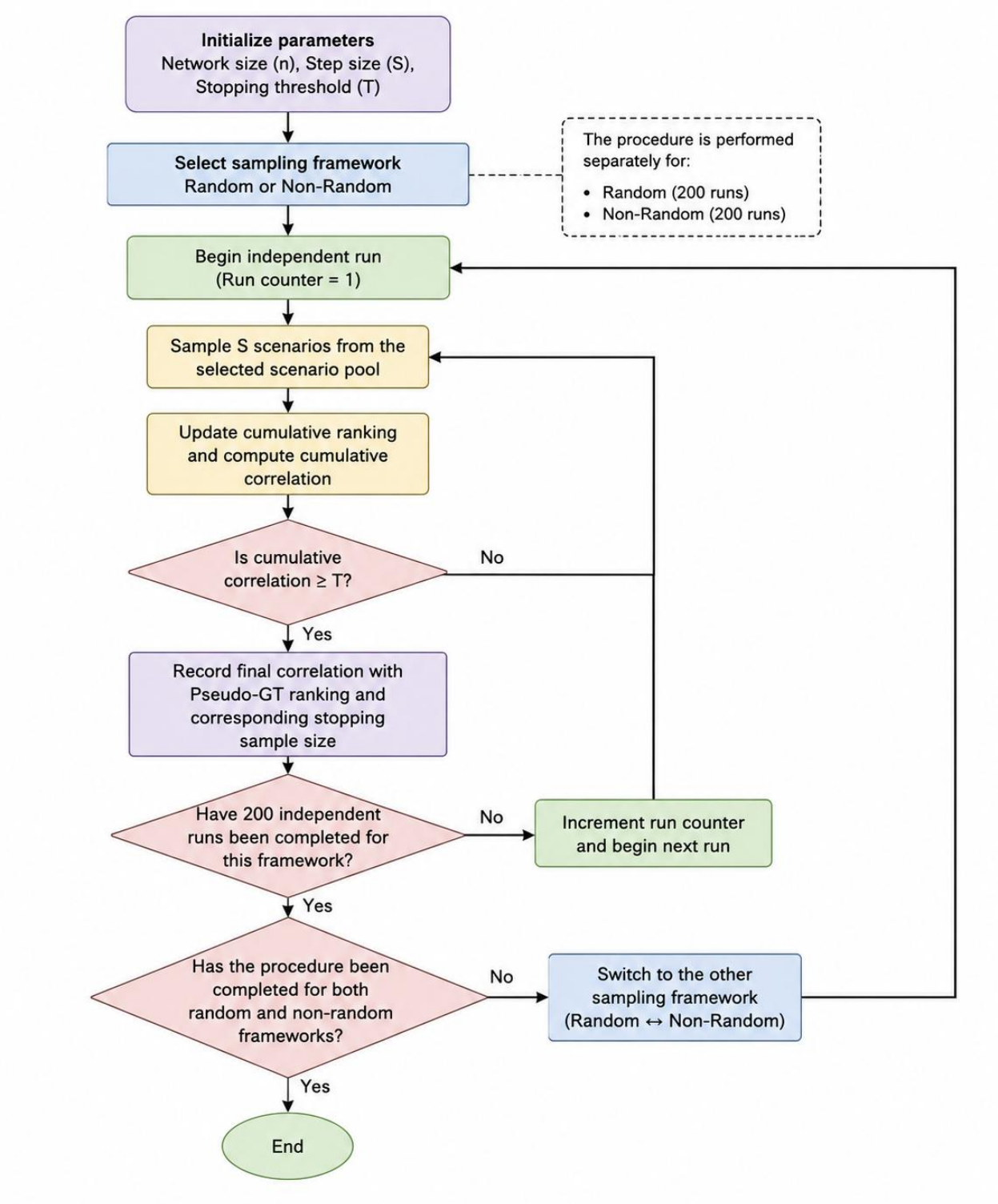

Figure 2 - Flowchart of the Stepwise Sampling Procedure with Stopping Rule

4. RESULTS AND DISCUSSION

4.1. Validation of the Pseudo-GT Ranking

Before evaluating the proposed sampling framework, the stability of the Pseudo-GT ranking was assessed using a repeated split-sample validation procedure. For each network, the complete disruption-scenario pool was randomly divided into two equally sized subsets, and a criticality ranking was independently generated from each subset. Spearman's rank correlation coefficient was then calculated between the two rankings. To reduce sensitivity to any particular partition of the scenario pool, this procedure was repeated 100 times using independently generated random splits. Table 2 reports the mean and standard deviation of the resulting correlations.

Because the Pseudo-GT ranking was derived from the combined random and non-random scenario pools, three additional comparisons were conducted. Specifically, Spearman correlations were calculated between (1) the Pseudo-GT ranking and the ranking derived from the random scenario pool, (2) the Pseudo-GT ranking and the ranking derived from the non-random scenario pool, and (3) the rankings derived separately from the random and non-random scenario pools. These comparisons assess the consistency of the Pseudo-GT ranking with each scenario-generation method and indicate whether the resulting rankings are sensitive to the composition of the scenario pool.

Table 2 - Split-Sample Validation Results for the Pseudo-GT Ranking

Network	Mean Spearman Correlation (ρ)	Standard Deviation	Pseudo-GT vs. Random ρ	Pseudo-GT vs. Non-Random ρ	Random vs. Non-Random ρ
Sioux Falls	0.99	0.003	0.99	0.99	0.98
Eastern Massachusetts	0.94	0.006	0.98	0.94	0.88
Anaheim	0.85	0.006	0.90	0.80	0.74

According to Table 2, the Pseudo-GT ranking is highly stable for the Sioux Falls and Eastern Massachusetts networks and moderately stable for the Anaheim network. The relatively lower stability observed for Anaheim is likely attributable to its larger network size (914 links) and the relatively smaller disruption scenario size in the analysis. About 500,000 scenarios were generated for Anaheim compared whereas 760,000 scenarios were used for Eastern Massachusetts that is approximately four times smaller than Anaheim. Therefore, the stability of the Anaheim Pseudo-GT ranking is expected to improve with a larger disruption scenario size. Nevertheless, the consistently high correlations obtained across repeated random splits indicate that the generated scenario pools, and the Pseudo-GT rankings derived from them, are sufficiently representative to provide reliable baseline rankings for the subsequent evaluation of the proposed subsampling strategy.

4.2. Results for Sioux Falls

Figure 3(a) illustrates the relationship between the stopping sample size and the final correlation between stopping ranking and the Pseudo-GT ranking under random sampling for the Sioux Falls network (n=76). The blue markers correspond to disruption samples for which the stopping threshold (T) for the cumulative correlation is set to 0.90. Considering a step size equal to the network size $S = n = 76$ number of disruption scenarios is incrementally subsampled. After each incremental addition, the criticality ranking is updated, and the Spearman rank correlation between the updated ranking and the ranking before the addition is computed. If this correlation exceeds 0.90, the run is terminated. At stopping point, the sample size and the correlation of stopping ranking with the Pseudo-GT ranking is recorded and represented as a point in Figure 3(a). This procedure is repeated 200 times for each stopping threshold (0.90, 0.95, and 0.98) under both random and non-random sampling. The results of all runs are presented in Figure 3.

4.2.1. *Impact of Stopping Correlation Threshold*

According to Figure 3(a), for Sioux Falls network with the step size of $S = n = 76$ under random sampling, the stopping sample size associated with a stopping threshold of 0.90 ranges approximately from 200 to 400. Within this range, the final correlation with Pseudo-GT ranking varies between approximately 0.62 and 0.85. When the stopping threshold is increased to 0.95, the stopping sample size increases, ranging from approximately 250 to 600. Correspondingly, the final correlation with the Pseudo-GT improves, varying between 0.65 and 0.89. For T = 0.98, the stopping sample size and final correlation ranges vary between [550,1200] and [0.76,0.94], respectively. Based on these results, increasing in the stopping threshold is associated with higher sample sizes and higher correlations with Pseudo-GT. Figure 3(b), which presents the results for non-random sampling, exhibits the same overall trend.

4.2.2. *Comparison of Random vs. Non-Random Sampling*

Figure 3 also provides a comparison of the random and non-random sampling approaches for the Sioux Falls network. For the step size of (S=n), random sampling tends to produce higher, less variable, and more consistent (Figure 4) correlations with the Pseudo-GT ranking at smaller stopping sample sizes. These results suggest that, random sampling can produce more reliable criticality rankings with fewer scenarios.

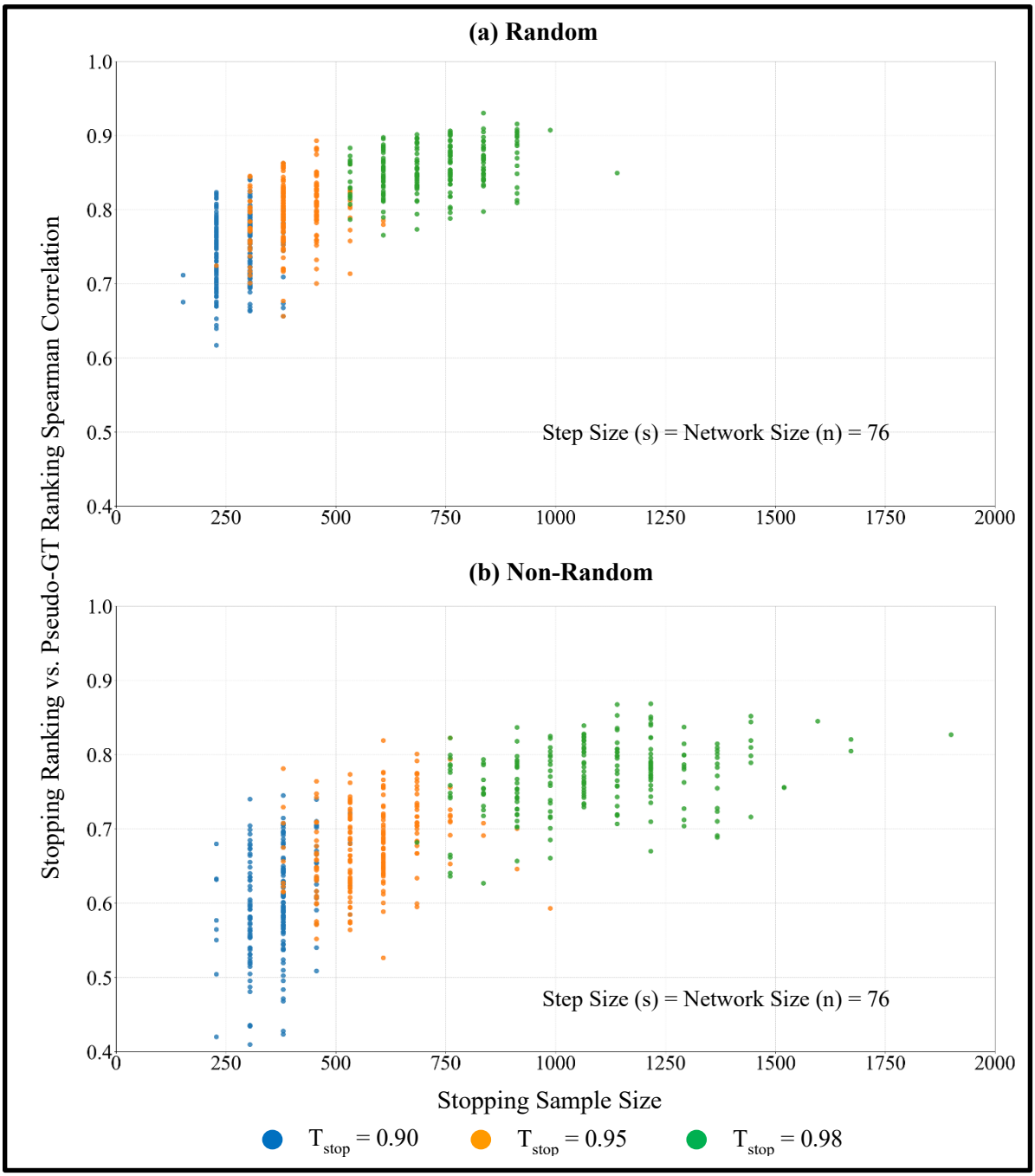

Figure 3 - Final correlation with the Pseudo-GT ranking vs. stopping sample size under Random and Non-Random sampling for Sioux Falls (step size = 76)

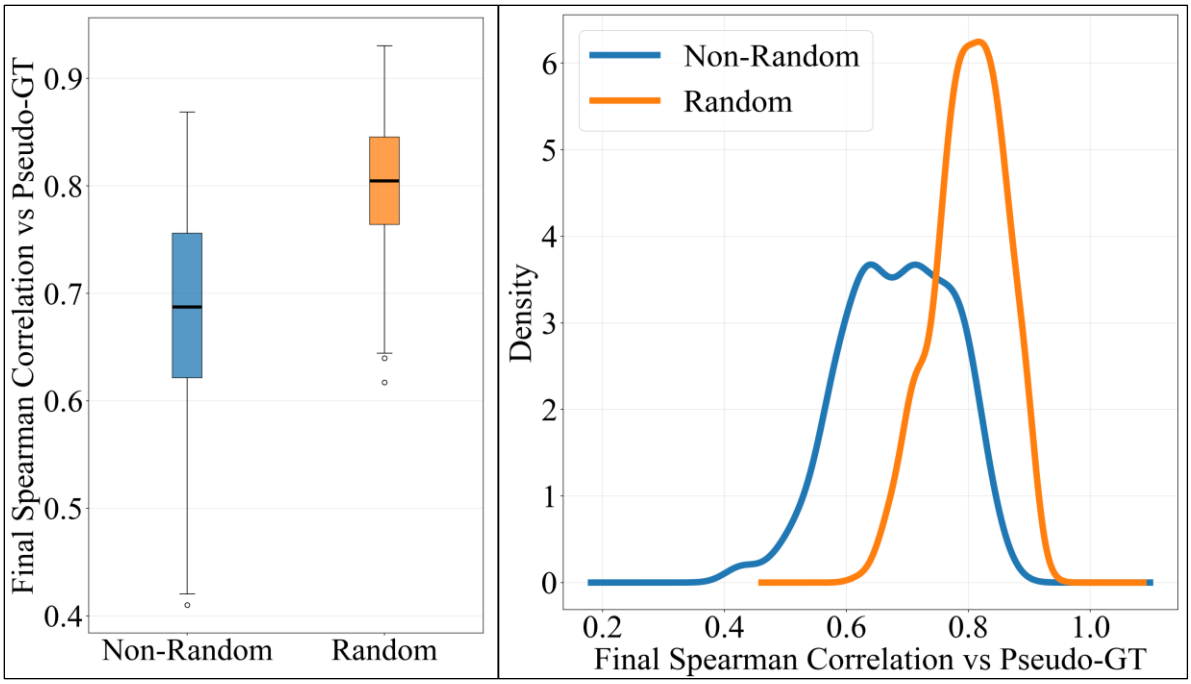

Figure 4 - Distribution of the correlation with the Pseudo-GT ranking for random and non-random sampling in the Sioux Falls network (step size = 76)

To quantify the comparison better, in the next step, subsamples that achieve high correlations with the Pseudo-GT ranking are further analyzed. Three cutoffs for the correlation with the Pseudo-GT ranking, $\rho = 0.7$, 0.8, and 0.9, are considered. Table 3 shows the percentage of runs for the Sioux Falls network with S = n that meet the high-correlation cutoffs under random and non-random sampling. According to the table, when the stopping threshold is set to 0.98, 100% (all 200) of the runs under random sampling achieve a correlation greater than 0.7 with the Pseudo-GT ranking, whereas the corresponding percentage for non-random sampling is 94%. The difference between the two methods becomes more pronounced when the high-correlation cutoff is increased to 0.8. While 95.5% of the runs under random sampling achieve a correlation greater than 0.8, only 25.5% of non-random sampling runs meet this criterion. This again suggests that, in this case, random sampling provides a more reliable ranking. When the average stopping sample size of the high-correlation runs is considered (Table 4), for the same stopping threshold of 0.98 and $\rho = 0.7$, the random sampling method terminates after an average of 719 scenarios, whereas the non-random sampling method terminates after an average of 1,104 scenarios. Overall, these results indicate that random sampling achieves higher correlations with the Pseudo-GT ranking at smaller sample sizes and, in this case, outperforms the non-random sampling method on both measures.

Table 3 - Percentage of runs meeting different final correlation cutoffs under Random (R) and Non-Random (NR) sampling for Sioux Falls (step size = 76)

Network	Step Size	Stopping Threshold	$\rho \geq 0.70$		$\rho \geq 0.80$		$\rho \geq 0.90$	
			% R	% NR	% R	% NR	% R	% NR
Sioux Falls (n = 76)	n = 76	0.9	83.5%	7.0%	12.0%	-	-	-
		0.95	99.0%	38.5%	50.5%	1.5%	-	-
		0.98	100.0%	94.0%	95.5%	25.5%	7.5%	-

Table 4 – Mean stopping sample sizes for different final correlation cutoffs under Random (R) and Non-Random (NR) sampling for Sioux Falls (step size = 76)

Network	Step Size	Stopping Threshold	$\rho \geq 0.70$		$\rho \geq 0.80$		$\rho \geq 0.90$	
			R	NR	R	NR	R	NR
Sioux Falls (n = 76)	n = 76	0.9	278	407	288	-	-	-
		0.95	402	612	413	684	-	-
		0.98	719	1104	721	1200	836	-

4.3. Analysis Across All Test Networks

The analysis presented in the previous section was based on a single network topology and a single step size in the subsampling procedure. This section extends the analysis to two additional networks, Eastern Massachusetts and Anaheim, and two additional step sizes, 2n and 3n, for the stepwise sampling to further investigate the effects of network topology and step size on the proposed procedure and to identify more generalizable findings.

Figure 5 and Figure 6 show the final correlation with the Pseudo-GT ranking versus stopping sample size across all test networks and step sizes under random and non-random sampling, respectively. Table 5 presents the percentage of runs meeting different final correlation cutoffs within each sampling setup, while Table 6 reports the corresponding average stopping sample sizes.

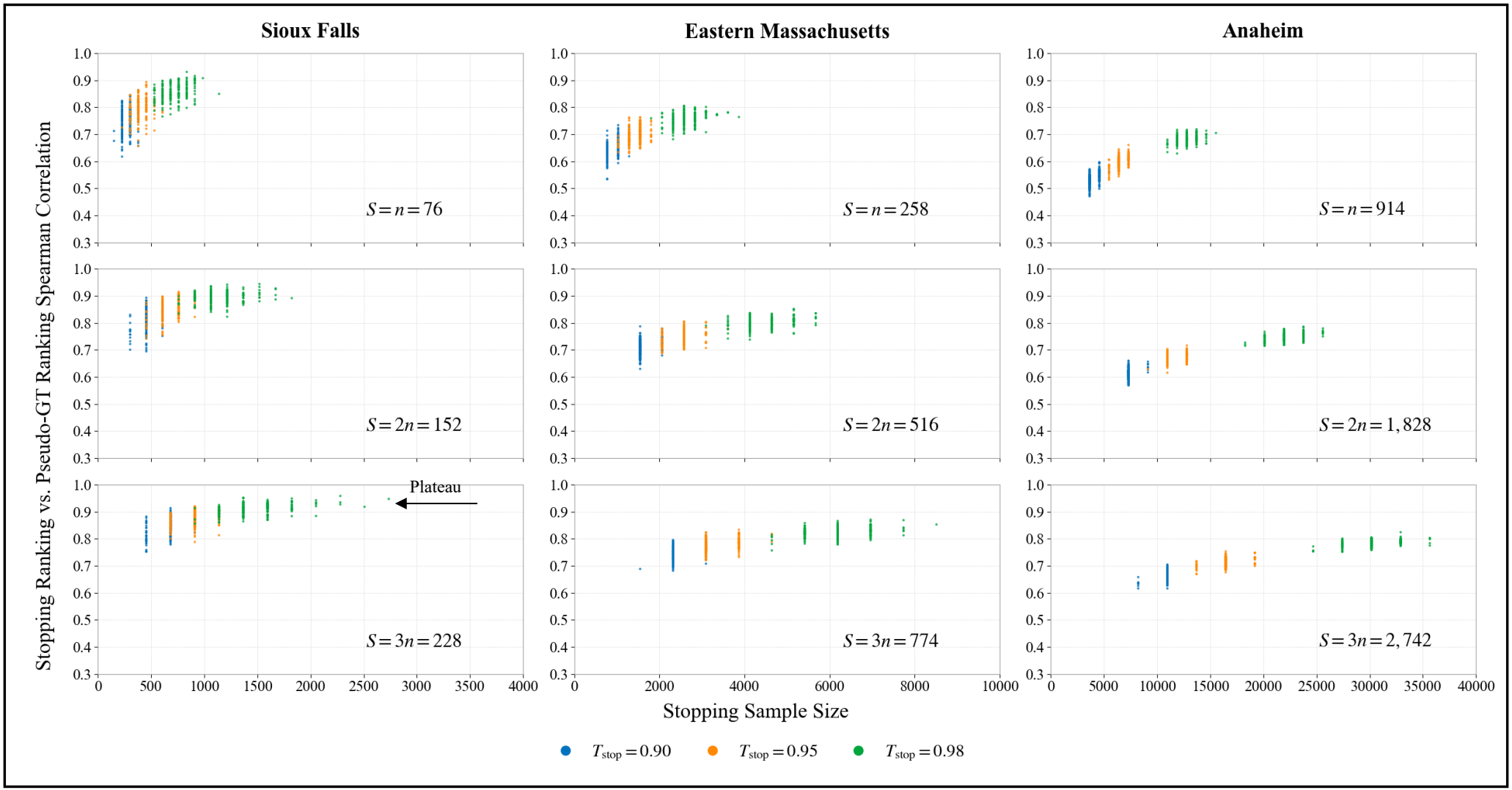

Figure 5 - Final correlation with the Pseudo-GT ranking vs. stopping sample size under Random sampling across all test networks and step sizes

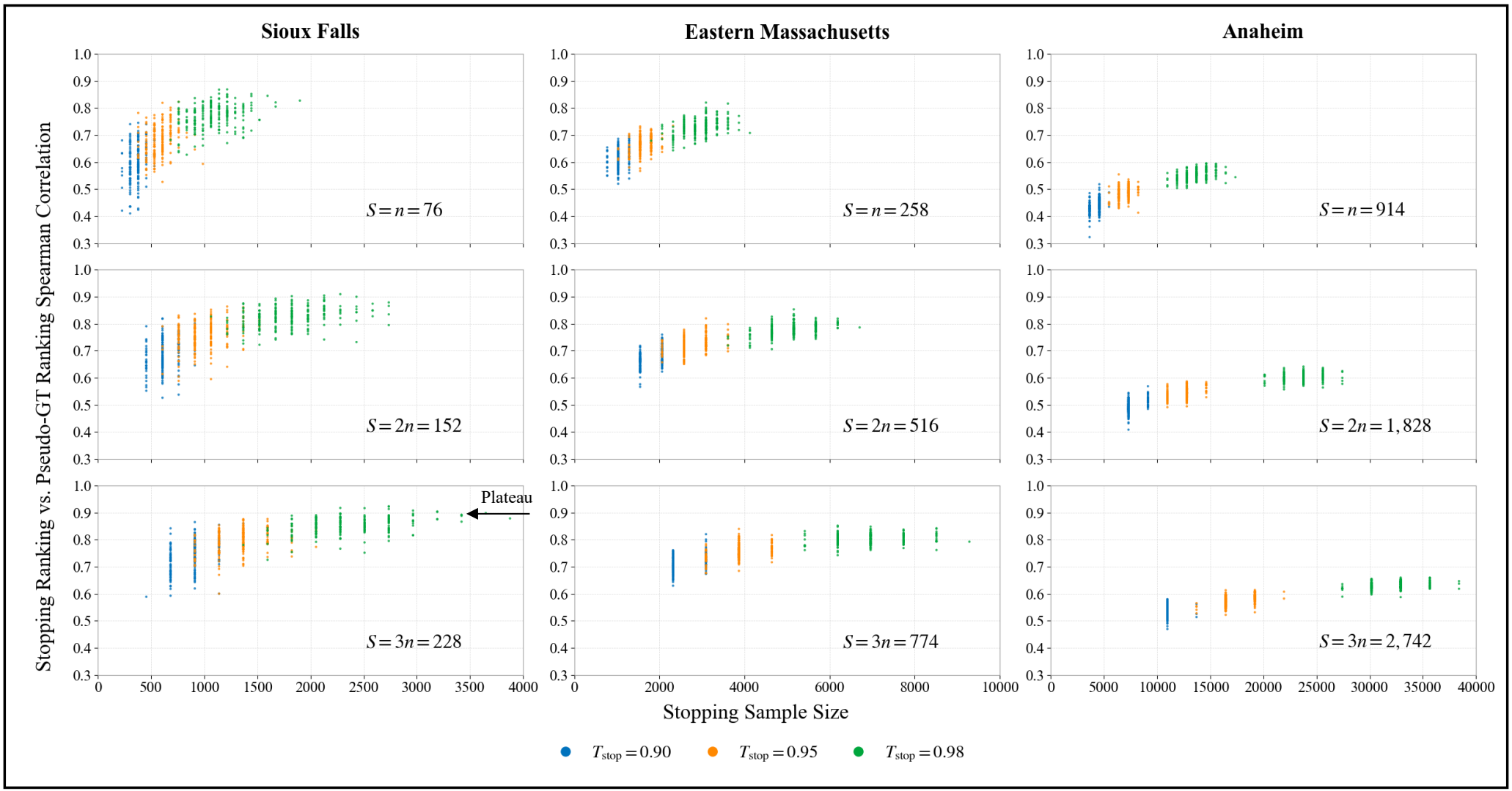

Figure 6 - Final correlation with the Pseudo-GT ranking vs. stopping sample size under Non-Random sampling across all test networks and step sizes

Table 5 - Percentage of runs meeting different final correlation cutoffs under Random (R) and Non-Random (NR) sampling across all test networks and step sizes

Network	Step Size	Stopping Threshold	$\rho \geq 0.70$		$\rho \geq 0.80$		$\rho \geq 0.90$	
			% R	% NR	% R	% NR	% R	% NR
Sioux Falls (n = 76)	n = 76	0.9	83.5%	7.0%	12.0%	-	-	-
		0.95	99.0%	38.5%	50.5%	1.5%	-	-
		0.98	100.0%	94.0%	95.5%	25.5%	7.5%	-
	2n = 152	0.9	99.5%	46.5%	64.0%	2.5%	0.5%	-
		0.95	100.0%	89.0%	92.5%	16.5%	6.0%	-
		0.98	100.0%	100.0%	100.0%	76.0%	41.5%	1.5%
	3n = 228	0.9	100.0%	76.5%	88.0%	10.5%	2.0%	-
		0.95	100.0%	98.5%	99.5%	50.5%	19.0%	-
		0.98	100.0%	100.0%	100.0%	95.0%	71.5%	4.0%
Eastern Massachusetts (n = 258)	n = 258	0.9	5.0%	-	-	-	-	-
		0.95	43.5%	11.5%	-	-	-	-
		0.98	99.0%	83.0%	2.0%	1.5%	-	-
	2n = 516	0.9	62.5%	27.5%	0.0%	-	-	-
		0.95	98.0%	77.5%	3.0%	0.5%	-	-
		0.98	100.0%	100.0%	48.0%	17.5%	-	-
	3n = 774	0.9	95.5%	64.0%	-	1.0%	-	-
		0.95	100.0%	97.0%	18.5%	6.0%	-	-
		0.98	100.0%	100.0%	88.5%	54.0%	-	-
Anaheim (n = 914)	n = 914	0.9	-	-	-	-	-	-
		0.95	-	-	-	-	-	-
		0.98	15.5%	-	-	-	-	-
	2n = 1,828	0.9	-	-	-	-	-	-
		0.95	3.5%	-	-	-	-	-
		0.98	100.0%	-	-	-	-	-
	3n = 2,742	0.9	1.0%	-	-	-	-	-
		0.95	75.5%	-	-	-	-	-
		0.98	100.0%	-	-	-	-	-

Table 6 – Mean stopping sample sizes for different final correlation cutoffs under Random (R) and Non-Random (NR) sampling across all test networks and step sizes

Network	Step Size	Stopping Threshold	$\rho \geq 0.70$		$\rho \geq 0.80$		$\rho \geq 0.90$	
			R	NR	R	NR	R	NR
Sioux Falls (n = 76)	n = 76	0.9	278	407	288	-	-	-
		0.95	402	612	413	684	-	-
		0.98	719	1104	721	1200	836	-
	2n = 152	0.9	473	675	487	821	760	-
		0.95	673	1008	684	1110	811	-
		0.98	1156	1809	1156	1848	1198	2077
	3n = 228	0.9	658	902	674	966	741	-
		0.95	894	1352	894	1404	978	-
		0.98	1496	2381	1496	2400	1550	2793
Eastern Massachusetts (n = 258)	n = 258	0.9	1032	-	-	-	-	-
		0.95	1486	1716	-	-	-	-
		0.98	2626	3004	2709	3268	-	-
	2n = 516	0.9	1668	2036		-	-	-
		0.95	2504	2806	2752	3096	-	-
		0.98	4484	5098	4601	5396	-	-
	3n = 774	0.9	2334	2564	-	3096	-	-
		0.95	3433	3890	3661	4064	-	-
		0.98	6196	6962	6240	7023	-	-
Anaheim (n = 914)	n = 914	0.9	-	-	-	-	-	-
		0.95	-	-	-	-	-	-
		0.98	13268	-	-	-	-	-
	2n = 1,828	0.9	-	-	-	-	-	-
		0.95	12535	-	-	-	-	-
		0.98	22228	-	-	-	-	-
	3n = 2,742	0.9	10968	-	-	-	-	-
		0.95	16579	-	-	-	-	-
		0.98	30231	-	-	-	-	-

4.3.1. *Subsample vs. Pseudo-GT Ranking Correlations*

According to Figure 5 and Figure 6, both random and non-random sampling produce runs that achieve high correlations with the Pseudo-GT ranking. The general trend indicates that larger

stopping sample sizes are associated with higher correlations with the Pseudo-GT ranking. However, some runs achieve similarly high correlations at considerably smaller sample sizes. Therefore, the parameters and characteristics of the subsamples generated by the stepwise sampling procedure are further analyzed in the following subsections to identify sampling strategies that reliably produce rankings with high correlation to the Pseudo-GT ranking while minimizing the required sample size.

4.3.2. *Network Size vs. Ranking Correlation*

Figure 5 and Figure 6 confirm the expectation that as the network size increases (hence, the network dependencies and flows become more complex), it becomes more challenging to achieve high correlations with Pseudo-GT. Nevertheless, as stopping threshold increases, the variability in the correlation with the Pseudo-GT ranking decreases, and the resulting rankings become more stable. The results further indicate that the final correlation plateaus beyond a certain sample size. The sample sizes at these plateau points are arguably the computationally favorable sample sizes because further increases in sample size yield negligible improvements.

4.3.3. *Impact of Stopping Correlation Threshold and Step Size*

Another observation, visualized in Figure 5 and Figure 6 and quantitatively presented in Table 5, is that the stopping threshold of $T = 0.98$ consistently produces higher correlations for both random and non-random sampling, particularly when $S = 3n$ is used as the step size. For example, according to Table 5, for the Sioux Falls network under random sampling with $S = 3n$, all three stopping thresholds perform similarly when $\rho = 0.70$. However, at $\rho = 0.90$, $T = 0.98$ performs substantially better, with 71.5% of runs achieving this threshold, compared with only 2% and 19% for $T = 0.90$ and $T = 0.95$, respectively. Furthermore, Table 6 shows that the difference in the mean sample size between $T = 0.98$ and the other stopping thresholds is relatively small. Therefore, the additional computational effort required for $T = 0.98$ is justified by the improved accuracy and reliability of the resulting link criticality ranking. Therefore, $S = 3n$ and $T = 0.98$ are selected as the step size and stopping threshold, respectively, for an efficient sampling strategy and are used in the subsequent analyses.

4.3.4. *Comparison of Random vs. Non-Random Sampling*

Random sampling was previously shown to be more computationally efficient (i.e., higher pseudo-GT ranking correlation with lower sample size) than non-random sampling in the Sioux Falls network. Figure 5 and Figure 6, together with Table 5 and Table 6, demonstrate that this trend is consistent across all three test networks. In particular, for the Anaheim network, none of the non-random sampling runs achieved a final correlation exceeding the cutoff of $\rho = 0.70$. Furthermore, Table 6 shows that, for every correlation threshold considered, the average sample size required to achieve the same ranking accuracy is consistently smaller under random sampling than under non-random sampling.

Since the primary distinction between the two sampling approaches lies in the distribution of simultaneous link removal orders within the generated disruption scenarios, the next section delves

into the effect of link removal order on sampling efficiency to identify a more computationally efficient sampling strategy.

4.3.5. *Impact of Link Removal Order on Sampling Efficiency*

The superior performance of random sampling suggests that the distribution of simultaneous link removal orders plays an important role in sampling efficiency. As shown in Figure 7, the removal-order composition differs substantially between the random and non-random scenario pools across the three test networks. Random sampling is dominated by higher-order disruption scenarios (primarily orders 9 and 10), whereas lower-order disruptions are comparatively underrepresented. Because higher-order scenarios simultaneously capture interactions among a larger number of links within a single traffic assignment, each scenario may provide richer information for estimating link criticality than lower-order scenarios. This characteristic may explain the superior performance of random sampling observed in the previous sections.

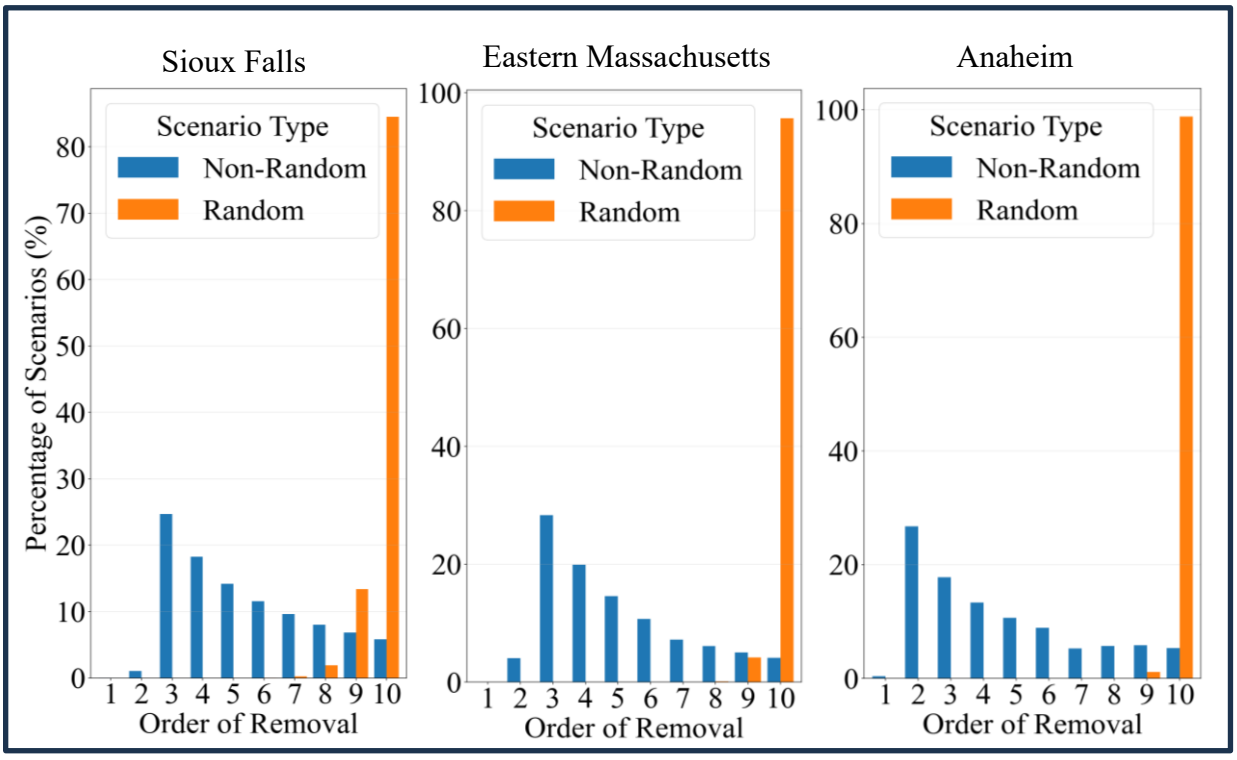

Figure 7 - Distribution of simultaneous link removal orders in the random and non-random scenario pools

Existing network criticality studies typically begin with single-link failures and progressively incorporate higher-order failures as additional computational resources become available. The results suggest that higher-order simultaneous link failures may contribute more effectively to accurate criticality ranking estimation than lower-order failures. Hence, the findings suggest an alternative strategy of using higher-order disruption scenarios from the start due to higher computationally efficiency for approximating the criticality ranking.

Please note that this study considered simultaneous link removals of up to order 10, but the optimal range and distribution of removal orders are expected to depend on network characteristics, topology, and size. Furthermore, excessively high removal orders may introduce disconnectivity that does not improve ranking accuracy. Identifying the optimal removal-order distribution and the maximum beneficial removal order for different classes of transportation networks therefore remains an important direction for future research.

4.4. Recommended Disruption Scenario Generation Strategy

The results obtained from the three case-study networks (Sioux Falls (n = 76), Eastern Massachusetts (n = 258), and Anaheim (n = 914)) provide consistent evidence for selecting an efficient scenario generation strategy for transportation network criticality analysis. Across all three networks, random scenario generation consistently outperformed non-random sampling by producing link criticality rankings with higher correlations to the Pseudo Ground Truth (Pseudo-GT) ranking while requiring smaller stopping sample sizes. In addition, among the three step sizes evaluated, S = 3n consistently yielded the highest ranking accuracy, while the stopping threshold T = 0.98 provided the most reliable convergence with only a modest increase in computational effort relative to lower stopping thresholds.

Based on these findings, the recommended scenario generation strategy is stepwise random sampling with a stopping rule considering disruption scenarios with up to 10 simultaneous link removals, using a step size of S = 3n and a stopping correlation threshold of (T = 0.98). Once the stopping criterion is satisfied, the resulting link criticality ranking provides a reliable approximation of the Pseudo-GT ranking while requiring only a small fraction of the computational effort needed to generate the complete Pseudo-GT scenario pool.

Besides identifying the preferred sampling procedure, the results also can provide general guidance regarding the sample size. To develop a conservative engineering recommendation, the 95th percentile of the stopping sample size for samples that passed high-correlation cutoffs $\rho \geq 0.70$ was considered. Using this criterion, the estimated 95th-percentile stopping sample sizes are approximately 1,600, 6,200, and 33,000 disruption scenarios for Sioux Falls, Eastern Massachusetts, and Anaheim, respectively. Normalizing these values by the number of links in each network yields required sample sizes of approximately 21n, 24n, and 36n, respectively. Accordingly, a practical rule of thumb can be established. For transportation networks containing up to approximately 1,000 links, a random sampling of up to 10 simultaneous orders of removal using a sample size of approximately 40n is recommended. Although smaller sample sizes may be sufficient for smaller networks, adopting 40n provides additional robustness while still requiring substantially lower computational burden than constructing the Pseudo-GT ranking itself. This conservative recommendation can typically be completed within hours on modern computing hardware, making this recommendation computationally feasible for transportation network criticality analysis.

4.4.1. *Validation of the Recommended Sampling Strategy on an Independent Network*

To assess whether the sampling guidelines developed using the Sioux Falls, Eastern Massachusetts, and Anaheim networks generalize to a network not included in their development, the Berlin-Friedrichshain network from the Transportation Network Repository [7] was selected as an independent validation case. As illustrated in Figure 8, the network comprises 224 nodes, 339 physical transportation links, and 184 centroid connector links. Because centroid connectors are artificial links that connect traffic analysis zones to the physical network, they were retained in the traffic assignment model but excluded from the set of disruption candidates and the resulting criticality ranking. Accordingly, the network size used to define the sampling parameters was based on the 339 physical links ($n = 339$).

A disruption scenario pool containing 1.2 million scenarios was generated for the Berlin-Friedrichshain network, comprising 600k randomly generated scenarios and 600k non-random scenarios. The stability of the resulting Pseudo Ground Truth (Pseudo-GT) ranking was evaluated using the split-sample validation procedure described in Section 4.1. Across the repeated splits, the mean Spearman correlation between the two independent rankings was 0.850, with a standard deviation of 0.009. The relatively high mean correlation and low between-split variability indicate the stability of Pseudo-GT ranking for use as the reference ranking in the subsequent validation analysis.

With the same composition of simultaneous-failure orders as before, random sampling continued to produce higher correlations with the Pseudo-GT ranking at smaller sample sizes. This finding indicates that the observed advantage of random sampling was not limited to the three benchmark networks used to develop the sampling guidelines.

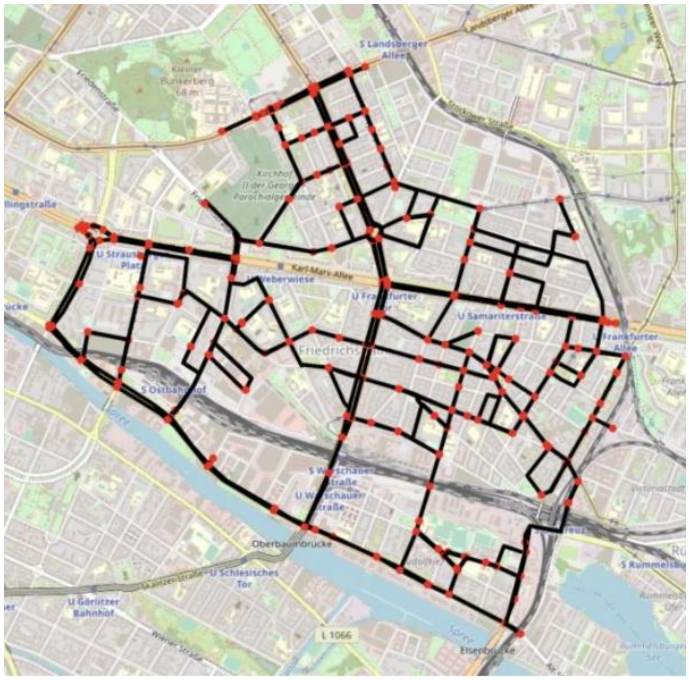

Figure 8 - Berlin-Friedrichshain Transportation Network

The two sampling recommendations established in Section 4.4 were evaluated. First, the stepwise random sampling procedure was implemented using a step size of $S = 3n$ and a stopping threshold of $T = 0.98$. A total of 200 independent runs were performed. At the stopping point, 155 runs (77.5%) produced rankings with a Spearman correlation of at least $\rho = 0.70$ with the Pseudo-GT ranking, whereas none achieved $\rho \geq 0.80$. The mean final correlation was 0.714, with a standard deviation of 0.021. Thus, although the stopping-rule procedure produced relatively consistent results across runs, its final correlation with Pseudo-GT was relatively lower for Berlin-Friedrichshain than for the networks used to develop the guideline. This can be due to different topology or OD structure.

Second, the fixed sample-size recommendation of $40n$ was evaluated. For $n = 339$, this recommendation corresponds to 13,560 randomly generated disruption scenarios per run. Among 200 independent runs, 187 (93.5%) achieved a final correlation of at least $\rho = 0.70$, while none achieved $\rho \geq 0.80$. The mean correlation increased to 0.732, and its standard deviation decreased to 0.019. Relative to the stopping-rule implementation, the $40n$ sample size therefore increased the proportion of runs meeting the $\rho = 0.70$ accuracy criterion by 16 percentage points.

Figure 9 presents the run-level Spearman correlations obtained under the two implementations. The results show that the fixed $40n$ recommendation produced larger and more consistent correlations above $\rho = 0.70$, although the correlations remained below $\rho = 0.80$ under both implementations. Overall, the findings support the practical applicability of the proposed sampling guidelines while highlighting opportunities for future refinement for large-scale transportation networks.

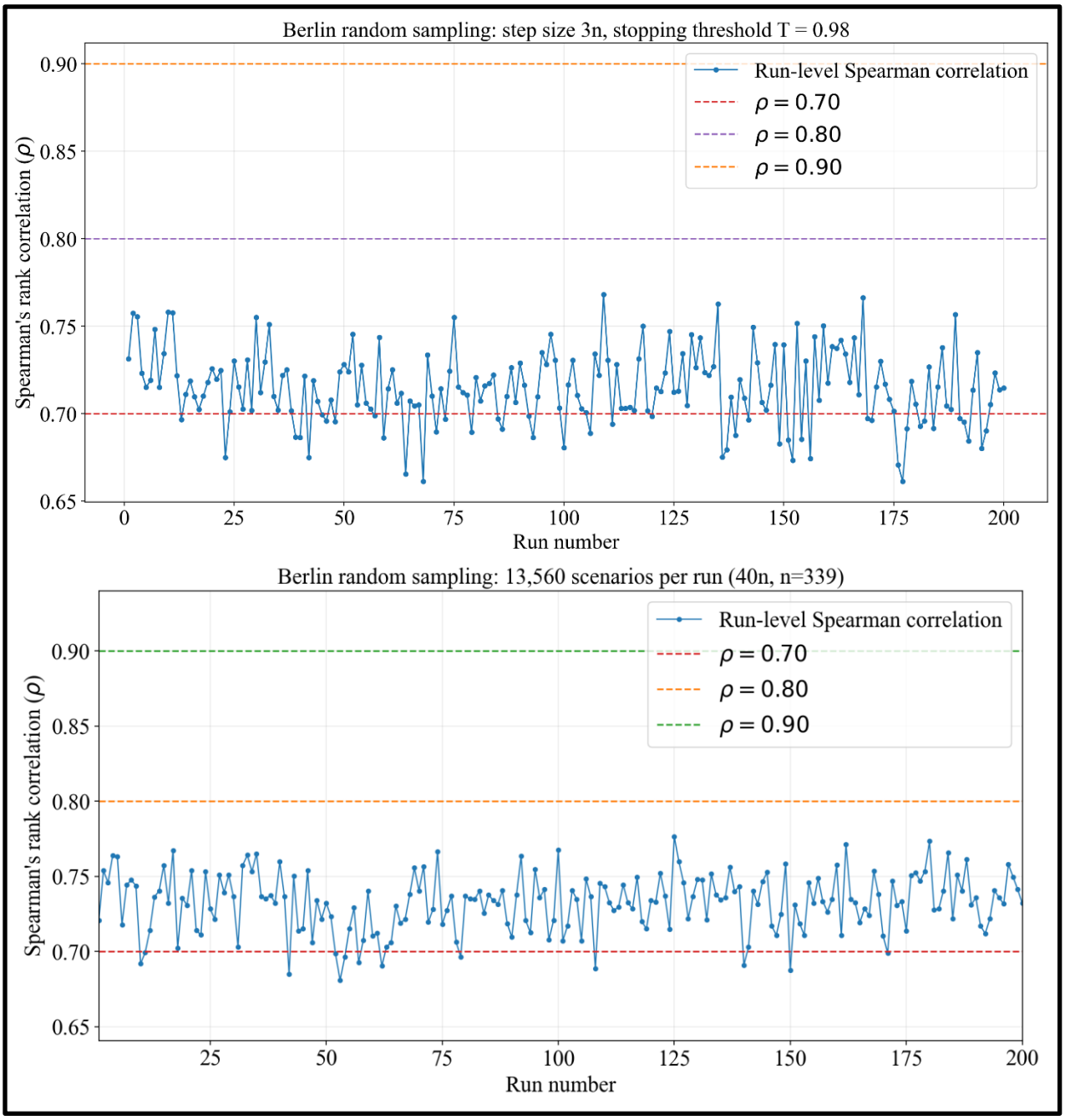

Figure 9 - Run-level Spearman correlations between the sampled and Pseudo-GT link criticality rankings for the Berlin-Friedrichshain network

5. CONCLUSION

Transportation network criticality analysis based on simultaneous multi-link disruptions can provide a more comprehensive representation of network vulnerability than conventional single-

link analysis by capturing interactions among disrupted links and travelers' rerouting responses. However, the combinatorial growth of possible disruption scenarios makes exhaustive evaluation computationally infeasible. This study addressed this challenge by developing a sampling framework for approximating network-wide individual-link criticality rankings using substantially fewer multi-link disruption scenarios.

The framework was evaluated using the Sioux Falls, Eastern Massachusetts, and Anaheim transportation networks. Large disruption scenario pools combining random and statistically controlled non-random scenarios were first generated to establish Pseudo Ground Truth (Pseudo-GT) rankings. Split-sample validation demonstrated sufficient stability of these rankings for use as reference benchmarks. A stepwise sampling procedure was then evaluated using different step sizes, stopping thresholds, and sampling approaches. Across the three networks, random sampling consistently produced higher correlations with the Pseudo-GT ranking at smaller sample sizes than non-random sampling. The results also indicated that the greater representation of higher-order disruptions in random sampling may improve efficiency by capturing interactions among more links within each traffic assignment. Among the tested configurations, a step size of $S = 3n$ and a stopping threshold of $T = 0.98$ provided the most reliable performance.

Based on the results across the three test networks, the recommended scenario generation strategy is stepwise random sampling with a stopping rule considering disruption scenarios with up to 10 simultaneous link removals, using a step size of S = 3n and a stopping correlation threshold of (T = 0.98). Another recommended rule of thumb is sampling of $40n$ randomly generated disruption scenarios, considering simultaneous removals of up to 10 links. Independent evaluation on the Berlin-Friedrichshain network provided additional support for these recommendations. Overall, the findings demonstrate that network-wide link criticality rankings that account for multi-link interactions can be approximated using only a fraction of the scenarios required to construct the Pseudo-GT ranking, substantially reducing computational requirements.

Several limitations provide directions for future research. First, the current framework considers a fixed maximum removal order of 10 links. Future research should investigate adaptive maximum disruption orders based on network size and characteristics to better capture link interactions across different networks. Second, although the framework was developed using three benchmark networks and evaluated on an additional independent network, testing a broader range of networks with diverse sizes and topological characteristics is necessary to improve the generalizability of the findings. Such analyses could support predictive models for estimating recommended sample sizes based on network characteristics. Finally, further research should examine the composition of high-performing subsamples, particularly the contribution of different simultaneous-removal orders to ranking accuracy. Identifying the most influential disruption orders and determining an optimal allocation of scenarios among them could further reduce computational effort while maintaining reliable link criticality rankings.

6. REFERENCES

[1] Y. Jiang, Y. Wang, W. Y. Szeto, A. H. Chow, and A. Nagurney, "Probabilistic assessment of transport network vulnerability with equilibrium flows," *International journal of sustainable transportation*, vol. 15, no. 7, pp. 512–523, 2021.

[2] E. Jenelius, "Redundancy importance: Links as rerouting alternatives during road network disruptions," *Procedia Engineering*, vol. 3, pp. 129–137, Jan. 2010, doi: 10.1016/j.proeng.2010.07.013.

[3] D. Z. W. Wang, H. Liu, W. Y. Szeto, and A. H. F. Chow, "Identification of critical combination of vulnerable links in transportation networks – a global optimisation approach," *Transportmetrica A: Transport Science*, vol. 12, no. 4, pp. 346–365, Apr. 2016, doi: 10.1080/23249935.2015.1137373.

[4] S. A. Bagloee, M. Sarvi, B. Wolshon, and V. Dixit, "Identifying critical disruption scenarios and a global robustness index tailored to real life road networks," *Transportation research part E: logistics and transportation review*, vol. 98, pp. 60–81, 2017.

[5] K. Jin, W. Wang, X. Li, X. Hua, S. Chen, and S. Qin, "Identifying the critical road combination in urban roads network under multiple disruption scenarios," *Physica A: Statistical Mechanics and its Applications*, vol. 607, p. 128192, Dec. 2022, doi: 10.1016/j.physa.2022.128192.

[6] J. Yang, X. Xu, and S. Ryu, "Unveiling network vulnerability under multiple area-covering disruption scenarios: A scenario-enumeration-free model and empirical insights into targeted protection," *Transportation Research Part E: Logistics and Transportation Review*, vol. 206, p. 104559, Feb. 2026, doi: 10.1016/j.tre.2025.104559.

[7] B. Stabler, *bstabler/TransportationNetworks*. (Feb. 18, 2026). Jupyter Notebook. Accessed: Feb. 24, 2026. [Online]. Available: https://github.com/bstabler/TransportationNetworks

[8] E. Jenelius and L.-G. Mattsson, "Road network vulnerability analysis of area-covering disruptions: A grid-based approach with case study," *Transportation research part A: policy and practice*, vol. 46, no. 5, pp. 746–760, 2012.

[9] A. Erath and K. Axhausen, "Joint Failure Vulnerability of Transportation Infrastructure," presented at the European Transport Conference, 2010Association for European Transport (AET), 2010. Accessed: Aug. 17, 2026. [Online]. Available: https://trid.trb.org/View/1115995

[10] N. Haghighi, S. K. Fayyaz, X. C. Liu, T. H. Grubesic, and R. Wei, "A multi-scenario probabilistic simulation approach for critical transportation network risk assessment," *Networks and Spatial Economics*, vol. 18, no. 1, pp. 181–203, 2018.

[11] J. L. Sullivan, K. Sentoff, J. Segale, N. L. Marshall, E. Fitzgerald, and R. Schiff, "A new risk-based measure of link criticality for flood risk planning," *Transportation*, vol. 51, no. 6, pp. 2051–2071, Dec. 2024, doi: 10.1007/s11116-023-10396-y.

[12] P. M. Johnson, W. Barbour, J. V. Camp, and H. Baroud, "Using machine learning to examine freight network spatial vulnerabilities to disasters: a new take on partial dependence plots," *Transportation Research Interdisciplinary Perspectives*, vol. 14, p. 100617, 2022.

[13] P. Y. R. Sohouenou, L. A. C. Neves, A. Christodoulou, P. Christidis, and D. Lo Presti, "Using a hazard-independent approach to understand road-network robustness to multiple disruption scenarios," *Transportation Research Part D: Transport and Environment*, vol. 93, p. 102672, Apr. 2021, doi: 10.1016/j.trd.2020.102672.

[14] M. Guo, T. Zhao, J. Gao, X. Meng, and Z. Gao, "Quantifying extreme failure scenarios in transportation systems with graph learning," *Patterns*, vol. 6, no. 4, 2025.

[15] Y. Gu, A. Chen, and X. Xu, "Measurement and ranking of important link combinations in the analysis of transportation network vulnerability envelope buffers under multiple-link disruptions," *Transportation Research Part B: Methodological*, vol. 167, pp. 118–144, Jan. 2023, doi: 10.1016/j.trb.2022.11.013.

[16] Y. Gu, S. Ryu, Y. Xu, A. Chen, H.-Y. Chan, and X. Xu, "A random-key genetic algorithm-based method for transportation network vulnerability envelope analysis under simultaneous multi-link disruptions," *Expert Systems with Applications*, vol. 248, p. 123401, Aug. 2024, doi: 10.1016/j.eswa.2024.123401.
[17] B. K. Bhavathrathan and G. R. Patil, "A Critical State of Multiple Simultaneous Link Disruptions," presented at the Transportation Research Board 93rd Annual MeetingTransportation Research Board, 2014. Accessed: Aug. 17, 2026. [Online]. Available: https://trid.trb.org/View/1289332
[18] Q. Tu, Y. Zhang, J. Feng, W. Wang, and Z. Zheng, "Link criticality identification and analysis in road networks using path flow weighted betweenness centrality," *Physica A: Statistical Mechanics and its Applications*, p. 130911, 2025.
[19] H. Barati, A. Yazici, and M. Bargahi, "Identification of Critical Links in Transportation Networks: An Integrated Neural Networks Approach that Incorporates Network Topology," *Available at SSRN 4914618*.
[20] D. Jana, S. Malama, S. Narasimhan, and E. Taciroglu, "Edge-based graph neural network for ranking critical road segments in a network," *Plos one*, vol. 18, no. 12, p. e0296045, 2023.
[21] Z. Chen, C. Yang, J.-H. Qian, D. Han, and Y.-G. Ma, "Recursive traffic percolation on urban transportation systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 33, no. 3, 2023.
[22] M. Owais and I. Ramadan, "Integrated deep learning and global sensitivity analysis framework for transportation link criticality evaluation," *Transportation Research Record*, vol. 2680, no. 8, pp. 355–378, 2026.
[23] A. Nagurney and Q. Qiang, "A transportation network efficiency measure that captures flows, behavior, and costs with applications to network component importance identification and vulnerability," in *Proceedings of the POMS 18th Annual Conference, Dallas, Texas, USA, MAY*, 2007.
[24] H. Barati, A. Yazici, and A. Almotahari, "A methodology for ranking of critical links in transportation networks based on criticality score distributions," *Reliability Engineering & System Safety*, vol. 251, p. 110332, 2024.