Automated Design of Realistic Contingencies for Big Data Generation

Tetiana Bogodorova, *Member, IEEE,* Denis Osipov, *Member, IEEE,* and Luigi Vanfretti, *Senior Member, IEEE*

***Abstract*—The letter proposes an algorithm for big data generation based on realistic selection of a set of contingencies for power systems described by undirected graphs. Every contingency is created by eliminating a certain number of elements in the system represented by graph edges. The number of elements as well as the distance between elements of the contingency is randomly selected according to a geometric probability distributions based on historical data. The duration of a fault that starts the contingency as well as the time intervals between elements of the contingency are chosen by sampling from a gamma distribution. In addition, the absence of islands in the system is assessed by analyzing the connectedness of the graph with deleted edges, which is quantified by computing the number of zero eigenvalues of the Laplacian matrix of the resulting graphs. The algorithm is validated on the Nordic 44-bus power system.**

***Index Terms*—Big data, contingency, gamma distribution, geometric distribution, graph connectedness, Laplacian matrix.**

I. Introduction

APPLICATION of machine learning (ML) methods are becoming more widespread in different fields of science and engineering, including power systems. The effectiveness of ML methods depends on quality and amount of data used for training and testing. These data can be accumulated from measurements and/or synthetically generated using physics-based models that have been validated with respect to measurements of a real system. Power system data generation using such models usually includes the time-domain simulation of a limited set of contingencies. The set of contingencies tend to consist of $n-1$ contingencies, excluding generator outages [1], [2]. In rare cases, typical $n-k$ contingencies are added to the set [3].

In this letter we propose a systematic approach to automatically design realistic single and multi-event contingencies based on the probability of a number of events in a contingency, as well as the probability of how far the next event is located with respect to the previous events in a multi-event contingency. To guarantee that the result of applying contingencies do not degenerate the power network into separate islands, the network connectedness is tested using graph theory. In addition, graph theory is used to verify that at least one generator remains connected after a contingency is applied. To make contingencies realistic, the duration of the short-circuit, as well as, time between the following events in a contingency is sampled from a probability distribution.

T. Bogodorova, D. Osipov, and L. Vanfretti are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA (e-mail: bogodt2@rpi.edu; osipod@rpi.edu; vanfrl@rpi.edu). This research is supported by NYPA/NYSERDA DeepGrid project (Grant No. A50626).

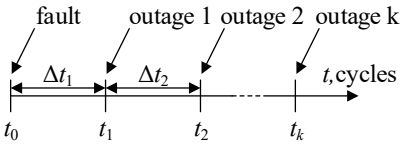

Fig. 1. Chronological structure of a contingency.

II. Proposed Method

Chronologically a contingency consists of a fault, the fault clearing time, the outage of faulted element 1, time interval to the outage of element 2, and further elements of a $n-k$ contingency (see Fig. 1).

A. Fault sampling

Fault sampling consists of fault clearing time sampling, fault type sampling, and fault location sampling.

A typical value of a fault clearing time is 5 cycles of a 60-Hz sine wave, while the shortest fault clearing time is about 3 cycles [4]. The maximum time that a circuit-breaker can remain closed under the short-time withstand current is 30 cycles or 0.5 seconds [4]. Using this information, a gamma distribution that models a probability of the fault clearing time deviations from the shortest fault clearing time of 3 cycles is parameterized. The idea is to design the spread of the final probability density function, so that the confidence interval bounds of at least 95% correspond to minimum and maximum fault clearing time. This interval corresponds to the 3σ bound according to the Vysochanskij-Petunin inequality for a unimodal distribution. Thus, the final distribution function with the aforementioned properties is the gamma distribution $f(x) = \beta^{\alpha}x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$, where $\alpha = 1.36$ is the shape parameter, $\beta = 0.18$ is the rate parameter, $\Gamma(\alpha)$ is the gamma function. The obtained probability density is shifted by 3 cycles that correspond the minimum fault clearance time (Fig. 2a). For this probability density function the range of fault clearing time from 3 to 30 cycles covers 98.4% of the distribution.

The type of a fault is sampled based on the probability of a single-phase fault of 70%, the probability of a two-phase fault of 20%, and the probability of a three-phase fault of 10% [5].

For a generator the fault location is set to be at the terminal bus. For a transformer the fault is uniformly selected among its terminal buses. For transmission lines the location of the fault within the length of a line is sampled using a uniform distribution.

B. Sampling the number of elements in a contingency

According to [6] the reliability criteria for a facility in the Western Interconnection, the frequency of n-1 contingency

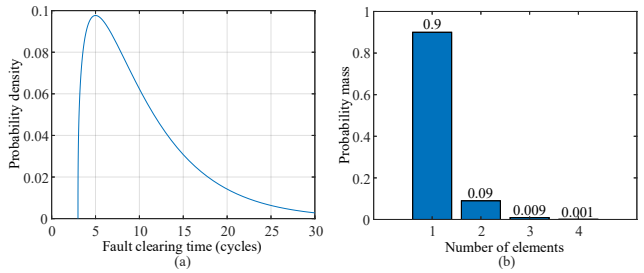

Fig. 2. Probability distributions: a) probability density function of the fault clearing time; b) probability mass function of the number elements involved in a contingency.

is 0.33 per year, the frequency of n-2 contingency is 0.033 per year, and the frequency of n-k contingency is 0.0033 per year. It is evident that probability of a contingency with an additional tripped element is 10 times smaller. Therefore, the probability of a contingency being $n-k$ is derived in the form of $p_k = (m-1)/m^k$, where $m = 10$. This probability distribution is identified as the geometric distribution with infinite support. To limit the support by the total number of elements n, the probability mass function is defined (1):

$$p(k) = \begin{cases} \frac{m-1}{m^k} & \text{if } k < n \\ 1 - \sum_{i=1}^{n-1} \frac{m-1}{m^i} & \text{if } k = n \end{cases} \quad (1)$$

An example of the probability mass function (1) for a system with 4 elements is shown in Fig. 2b.

C. Presence of generation

After a realistic contingency at least one generator has to be connected to the system. To verify this condition using graph theory, the power system is described by a multigraph (a graph with parallel edges and loops). Specifically, in the multigraph buses are represented by nodes, transformers and lines are represented by edges, and generators are represented by loops. An example of a multigraph with 4 nodes and 7 edges is shown in Fig. 3a. To verify that a multigraph has at least one loop, the trace (the sum of elements on the main diagonal) of an adjacency matrix A can be used. It shows how many loops a multigraph has. Element A_{ij} of the adjacency matrix shows how many edges are between node i and node j. The adjacency matrix (2) represents the multigraph from Fig. 3. For example, the trace of matrix in (2) is equal to 2.

$$A = \begin{pmatrix} 1 & 2 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \quad (2)$$

Thus, if generators are present in a system, the trace of the adjacency matrix of the corresponding multigraph will be positive.

D. Integrity of the system

In order for ML algorithms to learn the behaviour of a power system as a whole, it is necessary to include only those scenarios in which integrity of the system is not violated. To assess that no islanding has happened after a contingency, the connectedness of a multigraph can be verified. The property

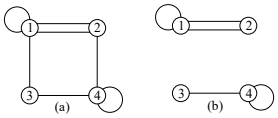

Fig. 3. An example of a multigraph.

of connectedness can be examined by analysing the Laplacian matrix of a multigraph. The Laplacian matrix is defined as the difference between the degree matrix and the adjacency matrix $L = D - A$,where the degree matrix D is a diagonal matrix whose elements D_{ii} show the number of edges connected to a node i. For example, the following degree and Laplacian matrices represent the multigraph from Fig. 3a:

$$D = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \quad L = \begin{pmatrix} 3 & -2 & -1 & 0 \\ -2 & 3 & 0 & -1 \\ -1 & 0 & 2 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix} \quad (3)$$

The number of zero eigenvalues of the Laplacian matrix shows the number of subgraphs of a multigraph [7]. To illustrate this fact, the eigenvalues of the Laplacian matrix in (3) are 0.00, 2.00, 2.59, and 5.41. If edges 1-3 and 2-4 are deleted, so that the multigraph has two subgraphs as shown in Fig. 3b, the eigenvalues change to 0, 0, 2, and 4. Thus, if the integrity of the system is not violated, the list of eigenvalues of the Laplacian matrix has only one zero element.

E. Sampling the location of outages

When the number of elements of a contingency is identified (Section II-B), the location of the first disconnected element is sampled using uniform distribution. If a contingency includes the outage of more than one element, the location of the next disconnected element is defined based on the location of the previous disconnected elements. When a branch of a power system is tripped, the change in power flow is larger in branches that are closer to the tripped branch. Therefore, based on power flow change ratio as a function of distance after a contingency [8], we propose to use the probability distribution similar to (1) with $m = 6$ to sample the distance between the location of the next disconnected element and the location of previous disconnected elements. The distance is described by the number of edges of a multigraph in the shortest path between the node representing a terminal bus of the next disconnected element and the nodes representing terminal buses of the previous disconnected elements.

F. Algorithm of the proposed approach

To summarize, the algorithm consists of steps that have been introduced above. First, sampling of fault type (Section II-A), fault location (Section II-A), fault duration (Section II-A) have to be performed from the distributions that have been identified above. After the fault characteristics are defined, the number of outages has to be selected by sampling from the geometric distribution (Section II-B). Next, the selection of which outage occurs first is done uniformly. Then, the condition of number of outages is checked. If the number of the elements of the contingency is equal to one, the connectedness of the graph which represents integrity of power system (Section II-D) and presence of generation (Section II-C) are assessed. The

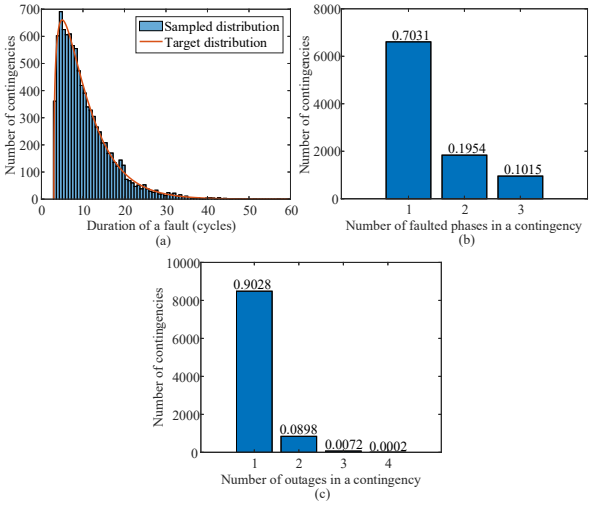

Fig. 4. Analysis of 9,399 contingencies in the Nordic 44-bus system.

contingencies that do not fulfill this constraint are eliminated as invalid. If the number of elements of the contingency is more than one, the algorithm continues to select the next element of the contingency with respect to the previously selected ones. The next element is identified by sampling the distance between the previous and the next elements of the contingency using the geometric distribution (Section II-E) defined from a uniformly selected terminal bus of the previously selected elements. When the distance is sampled, one can retrieve a list of the elements that are located at this distance from the previously tripped elements. The next step of the algorithm that follows is a uniform selection of the next element to be tripped from a list of equidistant elements. The process of selection of the next tripped element is repeated in the loop until the number of tripped elements equals the sampled number of elements that is done at the first step of the algorithm.

III. Case Study and Analysis

A. Verification of the proposed approach

To verify the proposed algorithm, a set of 10,000 contingencies is sampled using the Nordic 44-bus system [9]. The system has 80 generators, 12 transformers, 67 transmission lines. Among 10,000 contingencies 9,399 contingencies are valid (see Section II-C and II-D) and for which statistics is presented in Fig. 4. Figure 4a shows the histogram of the fault duration with the scaled-up distribution from Fig. 2a. Thus, the sampled contingencies closely follow the continuous distribution. In Fig. 4b a distribution of the type of fault with normalized probability mass values matches the data in Section II-A. In addition, a distribution of the number of outages in a contingency with normalized probability mass values (Fig. 4c) matches the probability mass values in Fig. 2b for $n-1$, $n-2$ and $n-3$ contingencies. The smaller number for $n-4$ contingencies is caused by the fact that some $n-4$ contingencies cause islanding and are moved to the invalid set.

B. Analysis of the proposed approach with respect to existing approaches

To demonstrate its advantages, the proposed approach for contingency design is compared with the existing approaches in terms of realistic characteristics of contingency sampling and contingency uniqueness. Approaches in [1], and [3] use a predefined set of n-1 and n-k contingencies correspondingly. In the data that is generated using these approaches, the same contingency is repeated multiple times, which can cause overfitting of machine learning algorithms. Another issue that can be caused by including same contingencies into both training and testing data, is an inaccurate assessment of machine learning method performance. Indeed, if the algorithm have seen the data in training data set, it will perform better seeing the same data in the testing data set. However, there is no guarantee that such performance can be generalized when testing on unseen data. In contrast, the proposed approach avoids these issues. The approach in [2] ensures contingency uniqueness by randomly sampling the fault clearing time for each contingency. Therefore, the comparison of the approach in [2] and the proposed approach is shown in Table I. The proposed approach provides more realistic sampling parameters for each category expanding the coverage of possible scenarios. If some scenarios are not present in the data, the algorithms will perform poorly on such scenarios. Avoiding this issue is one of the major strengths of the proposed approach.

TABLE I
Comparison of Contingency Design Approaches

Approach	In [2]	Proposed
Number of outages	only n-1	n-k, discrete distribution
Fault type	only 3-phase	all types
Fault location within a line	20, 40, 60, 80 %, discrete distribution	0 - 100 %, continuous distribution
Fault clearing time	0.1 - 0.4 seconds, uniform distribution	from 0.05 seconds, gamma distribution

IV. Conclusion

The proposed systematic approach for contingency design allows to generate a large number of unique realistic contingencies, which can be used to train machine learning models for power system security assessment.

References

[1] B. Wang *et al.*, "Power system transient stability assessment based on big data and the core vector machine," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2561–2570, 2016.

[2] J. James *et al.*, "Intelligent time-adaptive transient stability assessment system," *IEEE Trans. Power Syst.*, vol. 33, no. 1, pp. 1049–1058, 2017.

[3] C. Liu *et al.*, "A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 717–730, 2013.

[4] C. J. Nochumson, "Application of new technologies in power circuit breakers with higher interrupting capacity and short time ratings," in *IEEE Pulp and Paper Industry Technical Conference*. IEEE, 1999, pp. 27–41.

[5] M. Samotyj, "T & D system design and construction for enhanced reliability and power quality," EPRI, Tech. Rep., March 2006.

[6] "Phase I probabilistic based reliability criteria evaluation of exceptions list facilities," WECC, Tech. Rep., February 2001.

[7] A. E. Brouwer and W. H. Haemers, *Spectra of graphs*. Springer, 2011.

[8] S. Soltan *et al.*, "Cascading failures in power grids: Analysis and algorithms," in *Proc. e-Energy*, 2014, pp. 195–206.

[9] L. Vanfretti *et al.*, "An open data repository and a data processing software toolset of an equivalent Nordic grid model matched to historical electricity market data," *Data in brief*, vol. 11, pp. 349–357, 2017.