preprint of paper presented at North American Power Symposium, College Station TX USA, November 2021; Best paper award.

Transmission grid outage statistics extracted from a web page logging outages in Northeast America

Nichelle'Le K. Carrington Ian Dobson Zhaoyu Wang
Electrical and Computer Engineering Department
Iowa State University, Ames IA USA
emails nkcarrin,dobson,wzy@iastate.edu

***Abstract*—Detailed outage data is foundational for the study of power transmission grid reliability and resilience, and particularly for dependent outages and rarer events, but there are very few such data sets that are published and freely accessible to all engineers and researchers. There are voluminous logs of scheduled and actual outages in a region of Northeast America available on the web. We show how to compress and process these logged data to obtain bulk statistics describing the outages, such as event size, propagation, and spread on the network. These statistics are very useful for calibrating and validating models of resilience to ensure realism, and in developing data-driven approaches.**

I. Introduction

Outage data is the foundation of engineering practice, models, and simulations for reliability and resilience. Outage rates averaged over classes of equipment and time periods are published by some national organizations such as [1], and there are many books and papers with realistic annual outage rates for specific test systems, and sometimes for broad categories of weather conditions. All these averaged and typical data are useful, especially for steady state Markov modeling of reliability and detecting trends in reliability. However, for many problems involving dependencies between outages and rare events, including common cause outages, cascading, and resilience, averaged and typical single component data do not suffice, and the timings and details of many specific outages are required to advance the field.

While it is sometimes feasible for engineers and researchers to gain access to detailed industry outage data with non-disclosure agreements and publish some suitably non-identifying overall results, there is a special role for public data in advancing the field, since methods based on public data can be reproduced and improved on by other investigators. Moreover, the developed methods can subsequently be applied across the industry, since transmission utilities and system operators in North America and worldwide routinely collect their own detailed outage data.

There are few detailed outage records freely available to engineers and researchers; indeed, to the authors' knowledge there has been only one such source for transmission line outage data, namely the Bonneville Power Administration (BPA) website [2]. In this paper we show how to obtain detailed outage data from a second public website.

The BPA published data has been processed in various ways in [3]–[5] and used to validate and calibrate models in [4], [6]–[10]. There are many potential applications for detailed outage data, and we seek to generally facilitate applications by showing how to extract the new data. One application studies the size, propagation and spread of outages that bunch together in cascading or weather induced events, and we show the bulk statistics that can be obtained from the detailed outage data. These bulk statistics are useful in calibrating and validating cascading models and simulations [11], [12], or can be sampled to directly drive resilience quantification [13]. There are also parallel advances in methods driven by detailed outage data in distribution systems such as [14]–[16].

II. Transmission Utility Data

The New York Independent System Operator (NYISO) is the organization that manages New York State's electric grid and wholesale electric marketplace [17]. Detailed power grid outage data can be publicly accessed from the NYISO website [17]. The outage data on the website span from July 2002 to the present. For this paper, we use twelve years of these outage data from November 2008 to November 2020.

NYISO uses data collection methods that check the current status of the system every 5 minutes and record the status in a database. This 5 minute granularity of recording results in about 35 000 records per day and 12.6 million records per year for each data type. The two data types that we are interested in using are the real-time actual outages and the real-time scheduled outages.

The real-time actual outage data records the current status of all outages present in the system at the time of checkpoint, including the timestamp, part identification (PTID), equipment name and the outage date/time as shown in Table I. The timestamp is the checkpoint of the date and time at which the system recorded the information. Part identification (PTID) is a unique numerical tag identifying each system component. The equipment name for a transmission line identifies the names of the sending and receiving buses and the rated voltage; for example, N.SIMONE-COLTRANE_138_361. The equipment name for a transformer identifies the substation. Table I also shows outages of filter capacitors and circuit breakers. Note that the outage date/time is the date and time at which the component went out, which is different from the timestamp. Although the data is public, we follow good practice in anonymizing the substation names in Tables I and II.

TABLE I
REAL-TIME ACTUAL OUTAGE DATA

Timestamp	PTID	Equipment Name	Outage Date/Time
4/5/2021 2:22	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15
4/5/2021 2:22	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15
4/5/2021 2:22	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15
4/5/2021 2:22	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29
4/5/2021 2:22	25908	BRADFORD345KV_8________CB	1/26/2018 10:15
4/5/2021 2:22	25116	ELLIS_DC_GC1	3/12/2018 00:59
4/5/2021 2:22	25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09
4/5/2021 2:22	25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15
4/5/2021 2:22	25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15
4/5/2021 2:22	25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13
4/5/2021 2:22	25937	WYNTON_120KV_CAP_GC1 FILTER	3/20/2018 00:15
4/5/2021 2:22	25921	MARSALIS 345KV_1500-A________CB	4/23/2019 1:25
4/5/2021 2:22	25927	MARSALIS 345KV_77-2X________CB	4/23/2019 10:13
4/5/2021 2:27	25912	DELFAYO120KV_120-101________CB	1/15/2018 10:15
4/5/2021 2:27	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15
4/5/2021 2:27	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15
4/5/2021 2:27	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15
4/5/2021 2:27	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29
4/5/2021 2:27	25908	BRADFORD345KV_8________CB	1/26/2018 10:15
4/5/2021 2:27	25116	ELLIS_DC_GC1	3/12/2018 00:59
⋮	⋮	⋮	⋮

The real-time scheduled data record the outages that are scheduled to occur for operational or maintenance reasons. The real-time scheduled data include the timestamp, PTID, equipment name, scheduled out date/time, and scheduled in date/time as shown in Table II. The definition of the timestamp, PTID, and equipment name are the same as in the real-time actual data. The in date/time is the date and time that the component is scheduled to be re-energized. It is very common that scheduled outages are rescheduled.

TABLE II
REAL-TIME SCHEDULED OUTAGE DATA

Timestamp	PTID	Equipment Name	Out Date/Time	In Date/Time
4/5/2021 2:22	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15	12/6/2021 10:59
4/5/2021 2:22	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15	5/12/2022 10:15
4/5/2021 2:22	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15	4/23/2021 2:22
4/5/2021 2:22	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59
4/5/2021 2:22	25908	BRADFORD345KV_8________CB	1/26/2018 10:15	4/13/2023 4:59
4/5/2021 2:22	25116	ELLIS_DC_GC1	3/12/2018 00:59	1/13/2025 00:59
4/5/2021 2:22	25916	E.FITZGERALD-DAVIS_345_31	10/20/2020 17:09	1/25/2025 1:59
4/5/2021 2:22	25917	GILLESPIE-ELLINGTON_345_30	1/25/2018 10:15	2/1/2025 0:59
4/5/2021 2:22	25904	N.SIMONE-HOLIDAY_138_465	1/16/2020 10:15	2/1/2025 23:00
4/5/2021 2:22	25905	SIMONE-N.SIMONE_C_115_3-VI	1/15/2018 4:13	12/6/2021 3:45
4/5/2021 2:22	25937	WYNTON_120KV_CAP_GC1 FILTER	3/20/2018 00:15	10/13/2023 0:59
4/5/2021 2:22	25921	MARSALIS 345KV_1500-A________CB	4/23/2019 1:25	3/28/2023 7:45
4/5/2021 2:22	25927	MARSALIS 345KV_77-2X________CB	4/23/2019 10:13	3/20/2021 1:15
4/5/2021 2:27	25912	DELFAYO120KV_120-101________CB	1/15/2018 10:15	5/26/2021 10:30
4/5/2021 2:27	25312	NRTHSIMONE_138N_138E_PAR 1	1/15/2018 10:15	12/6/2021 10:59
4/5/2021 2:27	25126	WYNTON_120_SVC_CLC1	1/26/2018 10:15	5/12/2022 10:15
4/5/2021 2:27	25913	DELFAYO_120KV_CAP_GC2 FILTER	1/15/2018 10:15	4/23/2021 2:22
4/5/2021 2:27	25909	N.SIMONE-COLTRANE_138_361	3/25/2021 12:29	10/20/2021 12:59
4/5/2021 2:27	25908	BRADFORD345KV_8________CB	1/26/2018 10:15	4/13/2023 4:59
4/5/2021 2:27	25116	ELLIS_DC_GC1	3/12/2018 00:59	1/13/2025 00:59
⋮	⋮	⋮	⋮	⋮

III. DATA PROCESSING

This section describes the details of the processing that compresses and combines the real-time actual outages and the real-time scheduled outages into a single dataset. The processing compresses the data to make it manageable, identifies the automatic outages, and removes repeated data. It is inherent in converting from data recording the outage status every 5 minutes to a list of outages described once that large amounts of repetitive data must be deleted. All the processing is done using Mathematica to help mitigate the difficulties of handling mixed alphanumeric and date and time data.

A. Compression

The objective of the compression is to discard most of the real-time data that is not needed in order to make the file sizes more manageable. The source files are for each day from November 2008 to November 2020 (except that there are some days missing in April and October 2010 and September 2016). Each month of daily source files is read and compressed as follows.

The real-time actual outage data is very repetitive as the outage is recorded every 5 minutes until it is restored. The real-time actual outage data is sorted according to PTID, then Equipment Name, then Outage Date/Time, and then Timestamp. Then the outages are grouped according to the same successive PTID, Equipment Name, and Outage Date/Time and only the first and last of each group (with the minimum and maximum timestamp respectively) are retained. This removes most of the repeated records for the same outage and compresses the real-time actual outage data.

The real-time scheduled outage data is very repetitive as the scheduled outage is recorded for every 5 minutes until it happens and the scheduled outages are frequently postponed to a later time. The timestamp is removed, and then duplicate records are discarded. Then only the outages that are not postponed are retained: the successive pairs of scheduled outages that either have different PTID or have Out Date/Times differing by more than 16 minutes are determined to be not postponed, and are retained. This leaves a record of only the last time the outage was scheduled in a month and compresses the real-time scheduled outage data.

Finally, for each of the actual and scheduled real-time data, the monthly compressed data is combined into a single dataset and sorted according to Out Date/Time.

B. Identifying automatic outages

One objective of the data processing is to identify the automatic outages. This is done by noting the outages that actually occurred but were not scheduled. That is, the automatic outages are those outages that are in the real-time actual outages but not in the real-time scheduled outages.

In detail, for each actual outage, the scheduled outage with the same PTID with scheduled Out Date/Time closest in time to the actual Out Date/Times is searched for. If there is no such scheduled outage, or the closest scheduled Out Date/Time is more than one hour different than the actual Out Date/Time, then the actual outage is identified as automatic. If the closest scheduled Out Date/Time is less than one hour different than the actual Out Date/Time, then the actual outage is identified as scheduled. The processing can now neglect the scheduled outage data and proceed with the actual outages identified as automatic or scheduled.

We were unable to deduce usable component repair times from the data.

The final step is to remove any remaining repeated records of the same outage. Any successive duplicated records of an outage with the same PTID, Equipment Name, and Out Date/Time are removed.

C. *Extracting transmission line outages*

The transmission lines in the outage data have a standard format in their Equipment Name of two 8 character sending and receiving bus names separated by a hyphen, followed by the rated voltage and other information. It is straightforward to extract the transmission line outages by detecting this format (select Equipment Names with the 9th character a hyphen).

From the twelve years of data, the processing results in 45 178 transmission line outages, comprising 9600 automatic line outages and 35 578 scheduled line outages.

D. *Forming the network*

The first step in forming the network is to clean the bus names. There can be slight variations in spaces, punctuation or abbreviation that prevent the bus being uniquely identified by its bus name that need to be resolved.

After the bus names are cleaned, since almost all transmission lines have a planned or automatic outage in 12 years of observation, it is feasible to form the network from outage data simply by adding a link between the sending and receiving buses of each line that was outaged in the data, as explained in detail in [5]. A key feature of the resulting network is that it is completely compatible with the outage data in that, by construction, all the outaged lines can be located on the network. (Note that it can be difficult in practice to precisely relate the outages with other descriptions of the network.)

There are 1192 buses in the cleaned bus data. Forming the network directly from the outage data yields a large connected component as shown in Figure 1 of 1139 buses. The majority of the 53 buses not in the large connected component are in portions of the grid outside New York state that are represented less comprehensively. 95.5% of the lines have voltage ratings ranging from 69 kV to 500 kV.

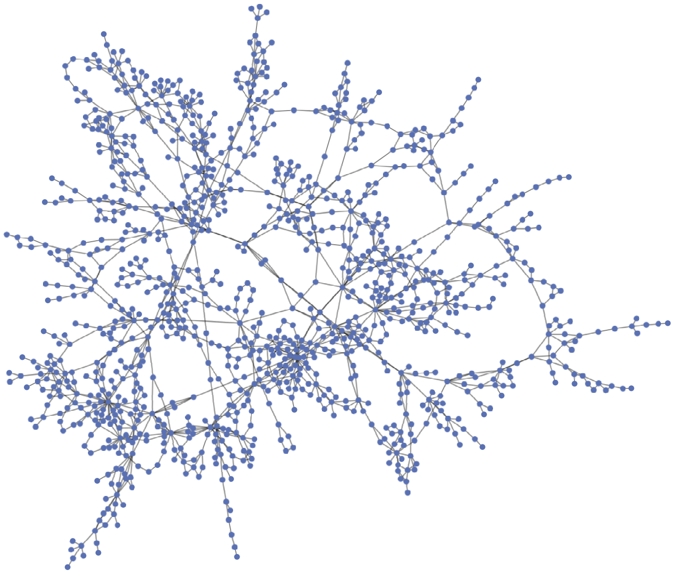

Fig. 1. Network formed from the outage data.

IV. Outage statistics

This section shows some bulk statistics derived from the automatic transmission line outage data that describe how the line outages cascade on the network. The methods used to derive the statistics are the same as in [4], [5], [18], where they were used to process the detailed outage data from BPA. The numerical values of the plotted data are given in the appendix to facilitate researchers making qualitative comparisons of the results with other simulated or real data.

The outage data is grouped into cascades and generations based on the outage start time using the simple method described in [4].[1] An outage occurring more than one hour after the preceding outage is assumed to start a new cascade, and within each cascade a series of outages less than one minute apart are grouped into the same generation. Thus each cascade consists of a series of generations, with each generation containing one or more line outages that occur closely spaced in time. For example, outages caused by protection within one minute are grouped together in the same generation. This processing produces 6687 cascades. Since the power system is generally resilient, 66% of cascades have only one outage, and 84% of cascades have only a single generation of outages that does not propagate further.

The initiating line outages are those in the first generation of outages. The probability distributions of the number of initiating line outages and the number of line outages in each cascade are shown in Figure 2, and the corresponding survival functions are shown in Figure 3. Figures 2 and 3 show how cascading increases the number of line outages beyond the initiating outages. Note the heavy tailed nature of the distributions, which is also seen in the analysis of BPA data in [4, Figure 1].

The propagation from generation k to generation $k+1$ in terms of the number of lines is defined as

$$\lambda(k) = \frac{\text{\# lines out in generation } k{+}1}{\text{\# lines out in generation } k}.$$

The line propagation in each generation is shown in Figure 4. The line propagation increases from a low value and then becomes more noisy for the higher generations due to the sparse data for the longer cascades. This general behavior is also seen in the analysis of BPA data in [4, Figure 3].

A better way to measure propagation [18] uses the probability distribution of the number of generations in cascades or events as shown in Figure 5. The absolute value of the slope of the fitted red line in Figure 5 is the System Event Propagation Slope Index, or SEPSI, that is a single number describing the propagation of the generations [18], [19]. In this case SEPSI = 3.17. Figure 5 can be compared with the analysis of BPA data in [18, Figure 2] and with NERC data in [19, Figure 4]. However, a strict quantitative comparison with [19, Figure 4] is not appropriate because [19] uses a different grouping of outages into events or cascades than this paper or [18].

[1] Note that alternative ways of grouping outages into events are being developed [19].

The network distance between two lines can be measured as the number of "hops" on the network between the lines [5].[2] For example, the distance of line to itself is zero and the distance of a line to a neighboring line with at least one bus in common is one.

One use of locating the outages on the network is to obtain the bulk statistics of how the outages spread on the network in cascades. For example, Figure 6 shows the network distances between random pairs of distinct lines in the same cascade. Successive sampling from this distribution gives an approximate high-level statistical model of cascade spread that can be used in modeling the cascading phase of resilience, as is done in [13, Figure 2] with the corresponding BPA data in order to give a data-driven quantification of overall transmission grid resilience.

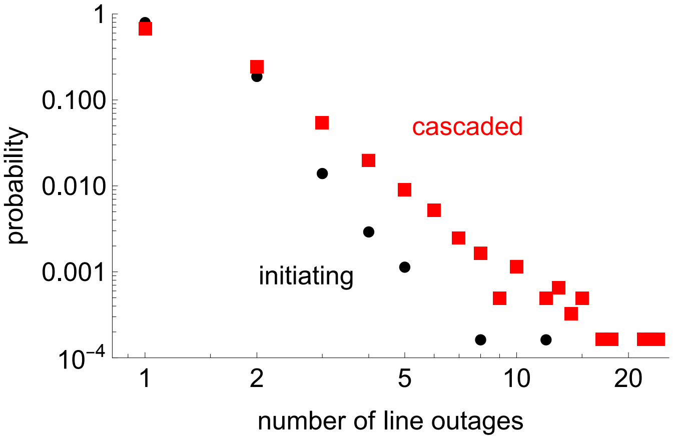

Fig. 2. Probability distributions of the number of line outages in initiating and cascaded outages.

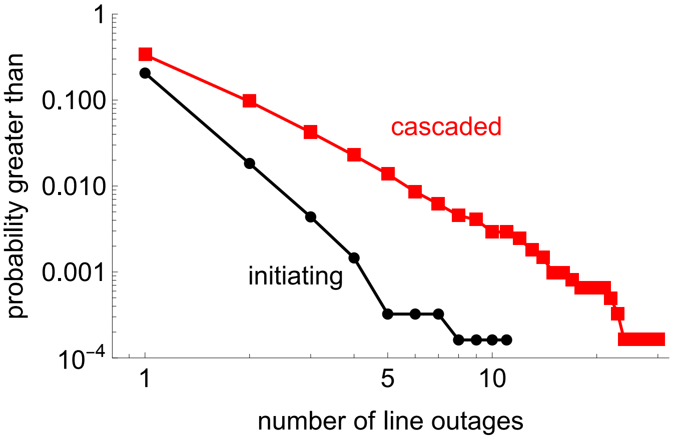

Fig. 3. Survival functions of the number of line outages in initiating and cascaded outages.

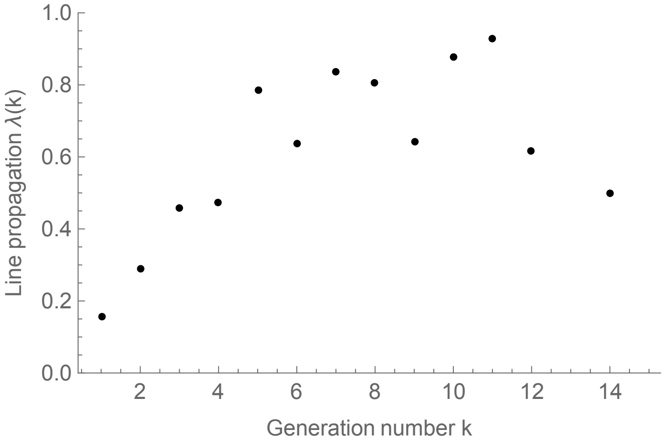

Fig. 4. Line propagation $\lambda(k)$ as a function of generation number k.

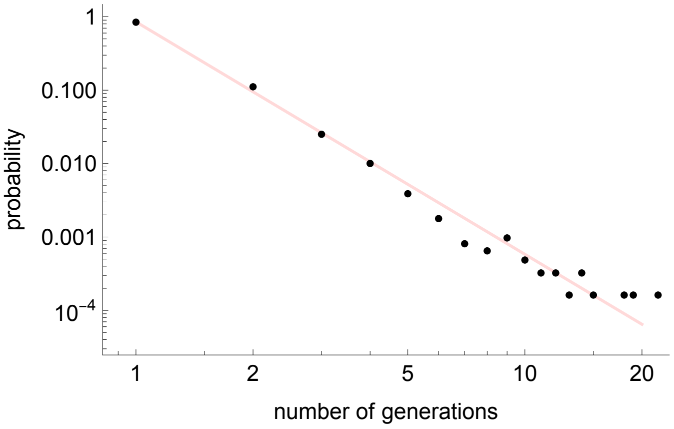

Fig. 5. Distribution of number of generations in cascades.

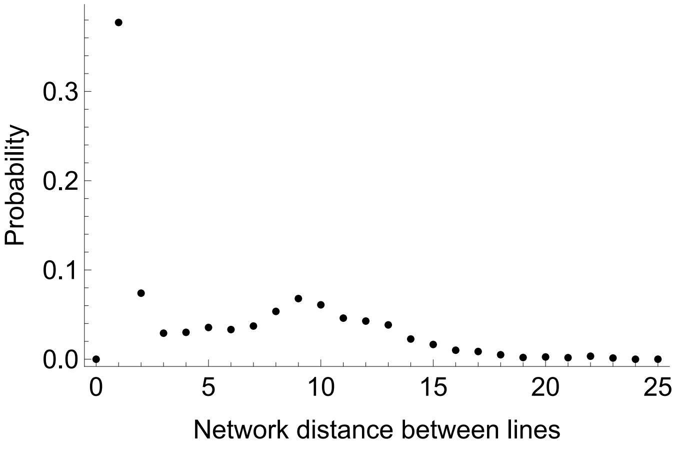

Fig. 6. Distribution of network distances between random pairs of distinct lines in the same cascade.

[2] More precisely, the network distance between lines L_i and L_j is defined as the minimum number of buses in a network path joining L_i to L_j.

V. Conclusions

In this paper we show how to process public website data for a region of Northeast America logging transmission grid outages every 5 minutes to obtain a detailed list of component outages that occurred, with the outage times to the nearest minute, the component details, and whether the outage is scheduled or automatic identified. To the authors' knowledge, this is only the second such public source of detailed outage data for transmission grids. We were unable to extract repair times from the data. The 12 years of processed data is sufficient to form a network on which the outages can be located.

The detailed outage data is valuable to engineers and researchers, especially in studying the rarer dependencies between outages that occur in cascading and resilience events. The outage data is rich with possibilities, including the study of outages of a variety of equipment. To illustrate one of the uses of the data, we show how bulk statistics obtained from the automatic transmission line outages in the data can be used to quantify how cascading or resilience events propagate and spread. These statistics have forms similar to those in the other publicly available source of detailed outage data, showing that the main features of the previous work with this other source of data are reproduced in another region of North America. The statistics from this and the other public source are foundational in ensuring realism and validation of simulations and models of cascading and resilience.

Acknowledgements

The authors gratefully thank the New York Independent System Operator for making publicly available the outage data that made this paper possible. The analysis and any conclusions are strictly those of the authors and not of the New York Independent System Operator.

Support in part by NSF under Grants 1609080 and 1735354 is gratefully acknowledged.

References

[1] North American Electric Reliability Corporation, [Online] Available: https://www.nerc.com

[2] Bonneville Power Administration Transmission Services Operations & Reliability website [Online] Available: https://transmission.bpa.gov/Business/Operations/Outages

[3] I. Dobson, N. K. Carrington, K. Zhou, Z. Wang, B.A. Carreras, J.M. Reynolds-Barredo, "Exploring cascading outages and weather via processing historic data," *Fifty-first Hawaii Intl. Conf. on System Sciences*, Jan. 2018, Big Island, Hawaii.

[4] I. Dobson, "Estimating the propagation and extent of cascading line outages from utility data with a branching process," *IEEE Trans. Power Systems*, vol. 27, no. 4, pp. 2146–2155, Nov. 2012.

[5] I. Dobson, B. A. Carreras, D. E. Newman, J. M. Reynolds-Barredo, "Obtaining statistics of cascading line outages spreading in an electric transmission network from standard utility data", *IEEE Trans. Power Systems*, vol. 31, no. 6, pp. 4831–4841, Nov. 2016 .

[6] B. A. Carreras, D. E. Newman, I. Dobson, N. S. Degala, "Validating OPA with WECC data," 46th Hawaii Intl. Conf. System Sciences, Maui, HI USA, Jan. 2013.

[7] B. A. Carreras, J. M. Reynolds Barredo, I. Dobson, D. E. Newman, "Validating the OPA cascading blackout model on a 19402 bus transmission network with both mesh and tree structures," 52nd Hawaii Intl. Conf. System Sciences, Maui, HI USA, Jan. 2019.

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

[9] B. Gjorgiev, B. Li, and G. Sansavini, "Calibration of cascading failure simulation models for power system risk assessment," *28th Intl. European Safety and Reliability Conf.*, Sep. 2019.

[10] M. Noebels, R Preece, M. Panteli, "AC Cascading Failure Model for Resilience Analysis in Power Networks," early access in *IEEE Systems Journal*, Dec. 2020.

[11] J. Bialek et al., "Benchmarking and validation of cascading failure analysis tools," *IEEE Trans. Power Systems*, vol. 31, no. 6, pp. 4887–4900, Nov. 2016.

[12] P. Henneaux et al., "Benchmarking quasi-steady state cascading outage analysis methodologies," Probability Methods Applied to Power Systems, Boise, Idaho, USA, Jun. 2018.

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

[14] N. K. Carrington, I. Dobson, and Z. Wang, "Extracting resilience metrics from distribution utility data using outage and restore process statistics," *early access in IEEE Trans. Power Systems*, 2021.

[15] A. Jaech, B. Zhang, M. Ostendorf, and D. S. Kirschen, "Real-time prediction of the duration of distribution system outages," *IEEE Trans. Power Systems*, vol. 34, pp. 773–781, Jan. 2019.

[16] C. Ji, Y. Wei et al., "Large-scale data analysis of power grid resilience across multiple US service regions," *Nature Energy*, vol. 1, pp. 1–8, Apr. 2016.

[17] New York Independent System Operator website,
[Online] Available: https://www.nyiso.com/

[18] I. Dobson, Finding a Zipf distribution and cascading propagation metric in utility line outage data. arXiv preprint arXiv:1808.08434. Aug. 2018.

[19] S. Ekisheva, R. Rieder, J. Norris, M. Lauby, I. Dobson, "Impact of extreme weather on North American transmission system outages," IEEE PES General Meeting, Washington DC USA, Jul. 2021.

Appendix
Numeric values of plotted results

This appendix gives the numbers used to obtain the plots to facilitate comparison with other results.

The probability distribution of the number of initiating line outages in Figure 2 is {0.7941, 0.1876, 0.01392, 0.002914, 0.001133, 0., 0., 0.0001619, 0., 0., 0., 0.0001619, 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.}.

The probability distribution of the number of cascaded line outages in Figure 2 is {0.6643, 0.2397, 0.0539, 0.01942, 0.008903, 0.00518, 0.002428, 0.001619, 0.0004856, 0.001133, 0., 0 .0004856, 0.0006475, 0.0003237, 0.0004856, 0., 0.0001619, 0.0001619, 0., 0., 0., 0.0001619, 0.0001619, 0.0001619, 0., 0., 0., 0., 0., 0.}.

The survival functions in Figure 3 can easily be calculated from the corresponding PDFs.

The number of outages in each generation used to obtain Figure 4 are {7609, 1183, 341, 156, 74, 58, 37, 31, 25, 16, 14, 13, 8, 10, 5, 8, 4, 3, 2, 1, 1, 1}.

The counts of the number of generations in all the cascades used to obtain Figure 5 are {1, 5210}, {2, 687}, {3, 155}, {4, 62}, {5, 24}, {6, 11}, {7, 5}, {8, 4}, {9, 6}, {10, 3}, {11, 2}, {12, 2}, {13, 1}, {14, 2}, {15, 1}, {18, 1}, {19, 1}, {22, 1}.

The probabilities in Figure 6 are {0., 0.3772, 0.07398, 0.02917, 0.03013, 0.03556, 0.03323, 0.03718, 0.05356, 0.06793, 0.06093, 0.04605, 0.04275, 0.03843, 0.02264, 0.01652, 0.01008, 0.008638, 0.005005, 0.001932, 0.002527, 0.001718, 0.003386, 0.001354, 0., 0.00004955}.