Estimating the propagation and extent of cascading line outages from utility data with a branching process

Ian Dobson, *Fellow IEEE*

***Abstract*—Large blackouts typically involve the cascading outage of transmission lines. We estimate from observed utility data how much transmission line outages propagate, and obtain parameters of a probabilistic branching process model of the cascading. The branching process model is then used to predict the distribution of total number of outages for a given number of initial outages. We study how the total number of lines outaged depends on the propagation as the cascade proceeds. The analysis gives a new way to quantify the effect of cascading failure from standard utility line outage data.**

***Index Terms*—Power transmission reliability, reliability modeling, reliability theory, risk analysis, failure analysis**

I. Introduction

Cascading failure is a series of dependent failures that progressively weakens the system. Large electric power transmission systems occasionally have cascading failures that cause widespread blackouts, with up to tens of millions of people affected [1], [2], [3]. The many mechanisms involved in cascading outages are complicated and varied, but all cascades include transmission line outages. In this paper we describe and quantify the bulk statistical behavior of cascading transmission line outages from standard utility data that records the times of line outages. Working with observed data to quantify cascading failure is complementary to simulation of cascading failure and has the advantage of not requiring the approximation of a selected subset of cascading mechanisms.

II. Previous work

Branching process models are bulk statistical models of cascading failure that are analytically and computationally tractable. Branching processes have long been used to study cascading processes in many other subjects, including genealogy, cosmic rays, and epidemics [4], but their application to cascading failure is much more recent, first appearing in [5], [6]. Evidence is accumulating that branching process models can represent probability distributions of blackout size. Qualitatively, observed [7], [8], [9], [10] and simulated [8], [11], [12], [13], [14], [15] blackout statistics show features such as probability distributions of blackout sizes with power law regions that can be shown by branching processes [5]. Quantitatively, branching processes have approximately reproduced the distribution of blackout sizes obtained from simulations of mechanisms of cascading failure in blackouts [16], [17]. Branching processes are used to analyze observed blackout data in [6], [18]. Moreover, branching processes can approximate other high-level models of cascading failure [5], [19], [20], [21].

Ren and Dobson [18] analyze propagation in a transmission line outage data set that is smaller and has a different source than the data set considered in this paper. They estimate an average value of propagation and use a branching process to predict the distribution of the number of line outages from the estimated propagation and the distribution of the initial line outages. One key difference between this paper and [18] is that this paper accounts for the varying propagation as the cascade progresses whereas [18] assumes a constant value of propagation throughout the cascade.

Chen and McCalley describe an accelerated propagation model for the number of transmission line outages in [22]. For parameters based on combined data for North American transmission line outages from [23], the accelerated propagation model applies to up to 7 outages. They examine the fit of the accelerated propagation model, a generalized Poisson distribution, and a negative binomial distribution to the North American transmission line outage data. Both the accelerated propagation model and the generalized Poisson distribution are consistent with the data. The generalized Poisson distribution is the distribution of the total number of outages produced by a Galton-Watson branching process with Poisson offspring distribution, whose mean number of offspring is constant except in the first generation.

Simulations of cascading failure blackouts in power transmission systems are reviewed in [8], [24]. These simulations approximate some selection of cascading mechanisms and compute some possible cascading sequences. Greig [25] represents cascading failure in more general flow networks.

There is a large literature on cascading in graphs and its relation to graph topology [26], [27] that is largely motivated by cascading failures in the internet. Roy [28] considers Markov reliability models on abstract influence graphs. Work on network vulnerability that accounts for network loading includes Watts [29], Motter [30], Crucitti [31], and Lesieutre [32]. Lindley [33] and Swift [34] represent cascading failure in systems with a few components by increasing the failure rate of remaining components when a component fails. Sun [35] applies accelerating failure to gradual degradation of a mechanical system.

Some initial results for this paper are in [36]. This paper rewrites [36] and has more data and new statistical analysis.

Ian Dobson is with ECpE department, Iowa State University, Ames IA 50011 USA dobson@iastate.edu. We gratefully acknowledge funding in part by California Energy Commission, Public Interest Energy Research Program, Power Systems Engineering Research Center (PSERC), and US DOE project "The Future Grid to Enable Sustainable Energy Systems," an initiative of PSERC. This paper does not necessarily represent the views of the Energy Commission, its employees or the State of California. It has not been approved or disapproved by the Energy Commission nor has the Energy Commission passed upon the accuracy or adequacy of the information.

III. Transmission line outage data

Transmission line outages are useful diagnostics in monitoring the progress and extent of blackouts. One common feature of large blackouts is the successive outage of transmission lines, and the number of lines outaged is a measure of the blackout extent. The number of transmission lines outaged is not a measure directly impacting society such as energy unserved or customers disconnected, but it is a measure of blackout size that is useful to utilities, and for which data is available. The line outages that do not lead to load shed can be regarded as precursor data for the line outages that do lead to load shed and blackout.

Transmission owners in the USA are required to report transmission line outage data to NERC for the Transmission Availability Data System (TADS). The transmission line outage data used in this paper is 10 512 outages in TADS data recorded by a North American utility over a period of 12.4 years [37]. The TADS data for each transmission line outage includes the outage time (to the nearest minute) as well as other data. All the line outages are automatic trips. More than 99% of the outages are of lines rated 69 kV or above and 97% of the outages are of lines rated 115 kV or above. There are several types of line outages in the data and a variety of reasons for the outages. In processing the data, both voltage levels and all types of line outages are regarded as the same and the reasons for the line outages are neglected. For this initial bulk statistical analysis, neglecting these distinctions is a useful first step as we proceed. It is best to start new methods of analysis in the simplest way first.

The power system is carefully designed and operated so that most transmission line outages have only one or a few outages occurring together. Most of these short cascades do not cause blackouts (no load is shed), but the longer cascades can lead to blackouts. In this paper we statistically analyze the propagation of line outages in observed cascades of automatic line outages, regardless of whether load is shed.

IV. Grouping outages into cascades, generations

For our analysis it is necessary to group the line outages first into different cascades, and then into different generations within each cascade. Here we use a simple method based on outages' timing [18], [6]. Since operator actions are usually completed within one hour, we assume that successive outages separated in time by more than one hour belong to different cascades. Since fast transients or auto-recloser actions are completed within one minute, we assume that successive outages in a given cascade separated in time by more than one minute are in different generations within that cascade.

The result of this grouping of the outages into cascades and generations is that there are $J = 6316$ cascades. The data can be tabulated as follows, where $Z_k^{(j)}$ is the number of outages in generation k of cascade j:

	generation number 0	1	2	3 ⋯	sum over generations
cascade 1	$Z_0^{(1)}$	$Z_1^{(1)}$	$Z_2^{(1)}$	$Z_3^{(1)} \cdots$	$Z^{(1)}$
cascade 2	$Z_0^{(2)}$	$Z_1^{(2)}$	$Z_2^{(2)}$	$Z_3^{(2)} \cdots$	$Z^{(2)}$
cascade 3	$Z_0^{(3)}$	$Z_1^{(3)}$	$Z_2^{(3)}$	$Z_3^{(3)} \cdots$	$Z^{(3)}$
⋮	⋮	⋮	⋮	⋮	⋮
cascade J	$Z_0^{(J)}$	$Z_1^{(J)}$	$Z_2^{(J)}$	$Z_3^{(J)} \cdots$	$Z^{(J)}$
sum over cascades	Z_0	Z_1	Z_2	Z_3 ⋯	

Summing over the generations gives $Z^{(j)}$, the total number of outages in cascade number j. Summing over the cascades gives Z_k, the total number of outages in generation number k. Table I shows $Z_0, Z_1, \cdots, Z_{10}$.

The probability distribution of the total number of outages in a cascade is given by

$$p_r = \text{probability of } r \text{ outages in a cascade}, \quad r = 1, 2, 3, \cdots$$

and is shown in Fig. 1 and Table III. p_r is estimated by counting the number of cascades with a given number r of outages and dividing by the number of cascades J:

$$\hat{p}_r = \frac{1}{J} \sum_{j=1}^{J} I\left[Z^{(j)} = r\right], \qquad r = 1, 2, 3, \cdots \tag{1}$$

(The notation $I[\text{event}]$ is the indicator function that evaluates to one when the event happens and evaluates to zero when the event does not happen.) Similarly, the probability distribution of the number of initial outages is given by

$$p_{0r} = \text{probability of } r \text{ initial outages in a cascade}, \quad r = 1, 2, 3, \cdots$$

and is shown in Fig. 1. p_{0r} is estimated as

$$\hat{p}_{0r} = \frac{1}{J} \sum_{j=1}^{J} I\left[Z_0^{(j)} = r\right], \qquad r = 1, 2, 3, \cdots \tag{2}$$

Since there are $J = 6316$ cascades, an event that occurs for only one of these cascades has an estimated probability 1/6316 = 0.00016. The cascades with a large number of outages may only occur one or a few times. The probability estimates for these rarer events are not reliable because of their extremely large variance and it is better to bin the data for the larger number of outages as shown in Fig. 2.[1]

TABLE I
Number of line outages in generations 0 to 10

Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
7539	1328	499	266	172	107	85	59	49	34	37

The method makes no claims about the causes of the outages or how they are related. It should be noted that there can be a

[1]For the total number of outages, the bins widths 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 4, 31 are chosen so that each bin has at least 10 outages. The estimated probability of the number of outages at the midpoint of the bin is then the proportion of outages in the bin divided by the bin width. For the initial outages the bins widths are 1, 1, 1, 1, 1, 3.

variety of causes for the outages propagating in generations. In addition to outages for which previous outages causing that outage can be identified, there may be outages with no clear antecedents other than a general weakening of the system and independently occurring or exogenously forced outages. For example, a moving storm front could exogenously force a sequence of outages that would be processed in the same way as a dependent cascade. One option is to adjust the processing based on additional information to account for this, with the simplest possibility being to neglect weather-related data. One caution is that it is unknown to what extent exogenous forcing from weather is augmented by additional dependent cascading effects.

V. Propagation in the Cascades

In the branching process model of cascading, each outage in each generation (a "parent" outage) independently produces a random number 0,1,2,3,... of outages ("child" outages) in the next generation according to an offspring distribution that is a Poisson distribution. The child outages then become parents to produce the next generation, and so on. If the number of outages in a generation is zero, the cascade stops. The mean number of child outages in generation k for each parent in generation $k-1$ (the average family size in generation k) is the propagation λ_k. λ_k quantifies the average tendency for the cascade to propagate from generation $k-1$ to generation k.

The propagation λ_k is estimated by dividing the number of outages in generation k by the number of outages in generation $k-1$:

$$\hat{\lambda}_k = \frac{Z_k}{Z_{k-1}}, \tag{3}$$

assuming that $Z_{k-1} \neq 0$. For example, from Table I, generation 0 has 7539 outages and these parents produce 1328 child outages in generation 1. Therefore the average number of children in generation 1 per parent in generation 0 is $\lambda_1 = 1328/7539 = 0.18$. Generation 1 has 1328 outages and these outages, as parents, produce 499 child outages in generation 2. Therefore $\lambda_2 = 499/1328 = 0.38$. Continuing these calculations leads to Fig. 3. In Fig. 3, as the cascade progresses, the propagation λ_k increases from 0.18 and then, although the results for higher generations become noisy, the propagation appears to level off. To quantify the noise, Fig. 4 shows error bars for λ_k computed using the methods explained in Appendix B. The higher generations have too few outages to accurately estimate λ_k. Therefore we estimate the propagation $\lambda_{5+} = 0.76$ for generations 4 and above by dividing the total number of children in generations 5 through 20 by the total number of parents in generations 4 through 19:

$$\hat{\lambda}_{5+} = \frac{\sum_{j=1}^{J} \left(Z_5^{(j)} + ... + Z_{20}^{(j)} \right)}{\sum_{j=1}^{J} \left(Z_4^{(j)} + ... + Z_{19}^{(j)} \right)} \tag{4}$$

(Equation (4) is the standard Harris estimator [4], [18], [16] applied to generations 4 through 20.) In summary, the estimates for propagation in the utility data are:

$$\begin{aligned} \lambda_1 &= 0.18 \\ \lambda_2 &= 0.38 \\ \lambda_3 &= 0.53 \\ \lambda_4 &= 0.65 \\ \lambda_{5+} &= 0.76 \end{aligned} \tag{5}$$

The increase in propagation as the cascade proceeds is a significant feature of the data. It confirms and quantifies the explanation and modeling in [22] for bulk statistics up to seven outages that conditional probabilities of further outages increase as the cascade proceeds.

Doubling the time intervals that are assumed when defining the generations and cascades shows that the propagation estimates (5) are not very sensitive to these assumptions.[2]

VI. Estimating Distribution of Number of Outages

We estimate the distribution of the total number of outages from a distribution of initial outages and the propagation using

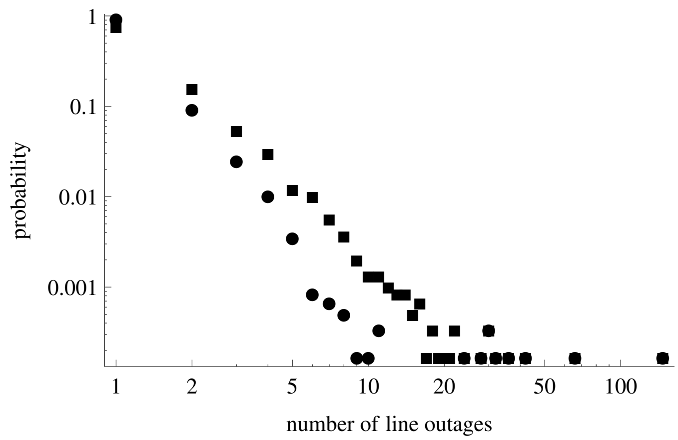

Fig. 1. Probability distribution of initial (circles) and and total (squares) line outages. Raw data shown with no binning.

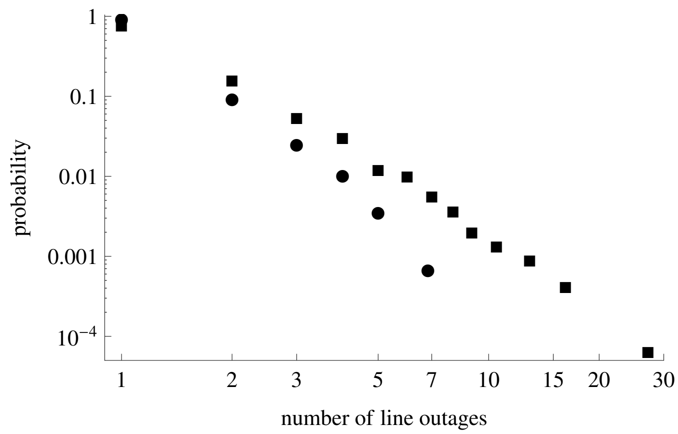

Fig. 2. Probability distribution of initial (circles) and and total (squares) line outages. Data is binned to have at least 10 outages per bin.

[2]If successive outages separated by more than 2 minutes are in different generations, then $\lambda_1 = 0.17$, $\lambda_2 = 0.37$, $\lambda_3 = 0.56$, $\lambda_4 = 0.55$, $\lambda_{5+} = 0.78$. If successive outages separated by more than 120 minutes are in different cascades, then $\lambda_1 = 0.23$, $\lambda_2 = 0.41$, $\lambda_3 = 0.59$, $\lambda_4 = 0.65$, $\lambda_{5+} = 0.79$. We are unable to decrease the interval of one minute that separates different generations because the line outage timings are specified in minutes.

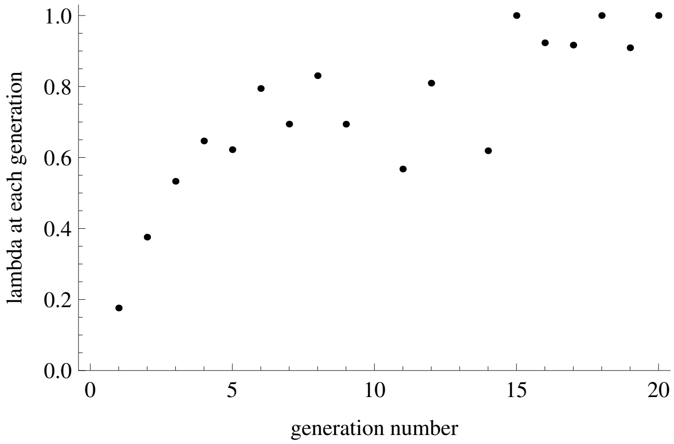

Fig. 3. Propagation λ_k estimated from the outage data for generations $k = 1, 2, ..., 20$.

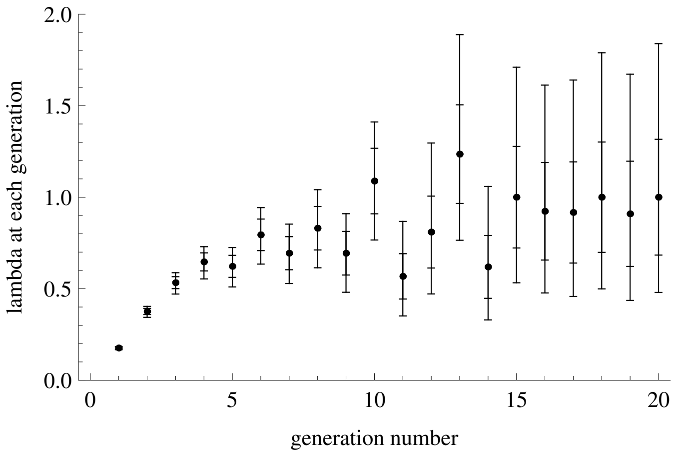

Fig. 4. Error bars on generation propagation λ_k. Inner bars are one standard deviation; outer bars are 95% confidence intervals.

a branching process model. This section shows the results; the details of the branching process calculations are given in appendix A.

A. Consistency with the empirical distribution of outages

We assume the initial distribution of outages p_{0r} estimated from the data using (2) and propagation at each generation given by (5). Then the branching process calculations estimate the distribution b_r of the total number of outages.

$$b_r = \text{probability of } r \text{ outages in a cascade}, \;\; r = 1, 2, 3, \cdots$$

Fig. 5 compares the estimated distribution b_r with the empirical distribution p_r directly estimated from the data. The satisfactory match between the estimated and empirical distributions shows that the branching process calculation of the distribution of the total number of outages gives a result consistent with the data.[3]

[3]The branching process describes the cascading averaged over the time period of the data. Note that in order to have sufficient empirical data for comparison, we are testing the branching process model averaged over 12.4 years of data, but propose applying it to describe the cascading averaged over one year of data.

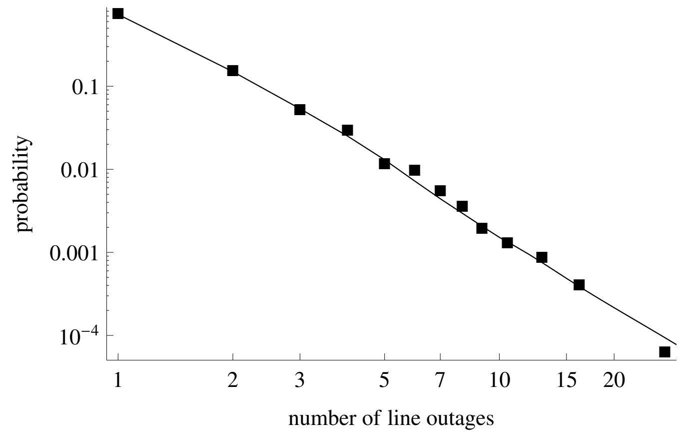

Fig. 5. Distribution of total number of outages from data (squares) and estimated using branching process (line joining the points of the discrete distribution).

A reasonable objection to the comparison in Fig. 5 is that the same data is used both to estimate the distribution b_r and to obtain the empirical distribution p_r. To address this objection, we estimate the distribution b_r from the odd numbered cascades and compare with the empirical distribution p_r for the even numbered cascades, and vice versa. The resulting matches, which are shown in Appendix E, are also satisfactory.

B. Predicting the distribution of outages

We can use the branching process model with the estimated propagation to predict the distribution of the total number of line outages from other assumptions of initial line outages. Either a particular number or a distribution of initial line outages is assumed and then branching process calculations are applied with the propagation (5) to predict the distribution b_r of the total number of outages.

For example, if there are 5 initial line outages, then the distribution of the total number of outages is shown in Fig. 6. This new capability is significant because traditional risk analysis methods can give good estimates of initial outages, and the method presented here can, based on observed data, quantify the number of additional outages caused by cascading. Note that the total number of outages is predicted, but there is no information about which outages occur.

Another application is to estimate the probabilities of extreme events that are difficult to determine empirically due to their rarity. For example, redoing the calculation of Fig. 5 up to 100 lines outaged predicts the total number of outages shown in Fig. 7. The statistical accuracy degrades for a large number of outages. Section VII calculates that for the full 12.4 years of data and a confidence level of 95%, the probabilities estimated in Fig. 7 are accurate within a factor of 2 up to 50 outages. However, the estimation for up to 100 outages assumes that the propagation for generations 4 to 100 remains constant at the value $\lambda_{5+} = 0.76$ estimated from the data for generations 4 to 20. Thus this assumption extrapolates the propagation from generations 4 to 20 to generations 21 to 100.

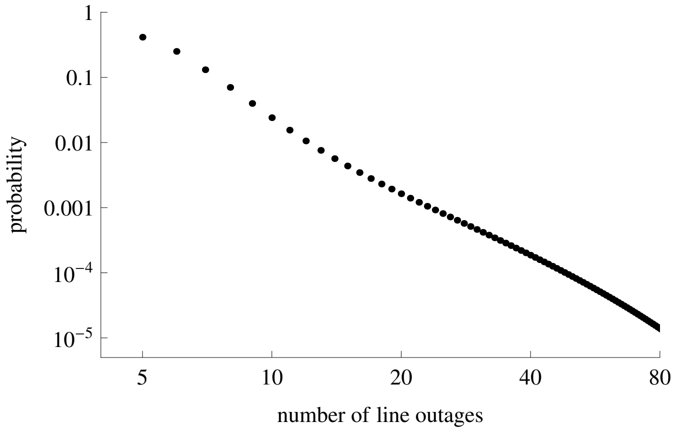

Fig. 6. Probability distribution of total number of line outages predicted by the branching process with estimated propagation (5) and assuming 5 initial line outages.

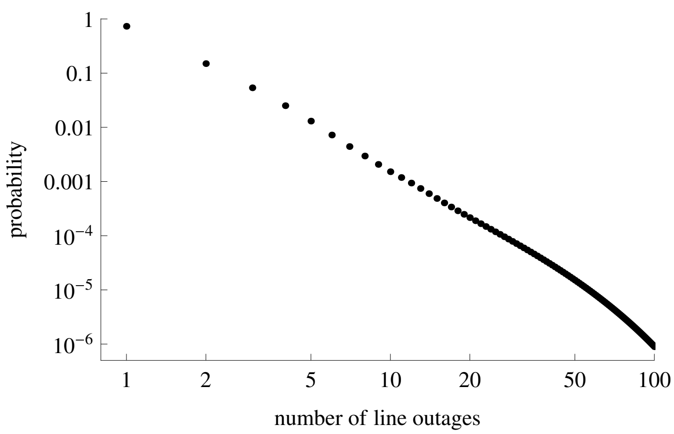

Fig. 7. Distribution of total number of line outages predicted using the branching process up to 100 outages. This figure extends and extrapolates the predicted line outage distribution in Fig. 5 to 100 outages.

C. Sensitivity of the distribution of outages to propagation

We consider how the predicted distributions depend on the propagation. Formulas for the sensitivity of the distribution of the number of outages with respect to the propagations λ_1, λ_2, λ_3, λ_4, λ_{5+} are derived in appendix C. For the data set of the paper, these formulas are evaluated in Table II.

For example, consider $D_{\lambda_3} b_r$, the derivative of b_r with respect to λ_3. Increasing λ_3 has no effect on the probability of 1 or 2 lines outaged, decreases the probability of 3 lines outaged (since it is then more likely that 3 lines outaged increases to more lines outaged), and increases the probability of 5 or more lines outaged. These general effects are expected, but they are quantified in Table II. The positive values of $D_{\lambda_k} b_r$ for the larger line numbers implies that large cascades can be mitigated by reducing λ_k, but it can be seen that the effectiveness of this mitigation reduces as k increases. Thus, if measures can be taken to reduce propagation λ_k at a single generation k, then it is more effective to do this for the early generations. However, it can be expected that measures to reduce propagation will often affect propagation in multiple generations, and then the effect on the distribution of the total number of lines outaged can be estimated by suitably combining $D_{\lambda_k} b_r$ and the changes in λ_k for multiple k. The positive and negative signs in the entries of Table II suggest that mitigation measures could involve tradeoffs between shorter and longer cascades.

VII. Statistical error in the distribution of line outages

We study the variability of the estimated probability $\hat{b}_r$ of r total line outages calculated by the branching process. Estimating probabilities of numbers of line outages within a factor of 2 seems reasonable, especially in the context of risk calculations (risk = probability times cost) in which direct and indirect blackout costs are substantially uncertain. Thus, given a particular observation b_r, we seek to quantify the accuracy of the estimate $\hat{b}_r$ with the probability

$$q(\hat{b}_r) = P\{b_r/2 \le \hat{b}_r \le 2b_r\}. \quad (6)$$

We evaluate (6) numerically as explained in Appendix D for 12.4 years of data, 10 years of data, and one year of data.

For 12.4 years of data, b_r is estimated up to $r = 50$ outages to within a factor of two at a confidence level of 95%. In detail, $q(\hat{b}_1) = 1.00$ and $q(\hat{b}_r)$ decreases as r increases, and $q(\hat{b}_{50}) = 0.95$ and $q(\hat{b}_{51}) = 0.94$.

For 10 years of data, b_r is estimated up to $r = 48$ outages to within a factor of two at a confidence level of 95%.

For one year of data, the results for $r = 1, 2, 3, ..., 20$ outages are shown in the third column of Table III. According to Table III, one year of data estimates b_r up to $r = 13$ outages[4] to within a factor of two at a confidence level of 95% and estimates b_r up to $r = 17$ outages to within a factor of two at a confidence level of 90%. This illustrates the performance of the branching process method for estimating the distribution of total number of outages from one year of data in one utility. The statistical accuracy can be improved by gathering data over a wider area or for a longer time.

The empirical probability estimator $\hat{p}_r$ of the probability of r total line outages is estimated using (1). The distribution of $J\hat{p}_r$ is Binomial(J, p_r), and hence the standard deviation of $\hat{p}_r$ is

$$\sigma(\hat{p}_r) = \sqrt{\frac{p_r(1-p_r)}{J}}. \quad (7)$$

Moreover, knowing the binomial distribution of $J\hat{p}_r$ allows us to directly evaluate

$$q(\hat{p}_r) = P\{p_r/2 \le \hat{p}_r \le 2p_r\} = P\{Jp_r/2 \le J\hat{p}_r \le 2Jp_r\}.$$

The fifth and sixth columns of Table III show the empirical probability p_r and its standard deviation $\sigma(\hat{p}_r)$) obtained from 12.4 years of data. The last column of Table III shows the $q(\hat{p}_r)$ that corresponds to $\hat{p}_r$ calculated from 10 years of data.

It can be seen in Table III that $q(\hat{b}_r)$ for one year of data is greater than or equal to $q(\hat{p}_r)$ for ten years of data. This shows an order of magnitude reduction in the amount of data needed for a given statistical accuracy to estimate the probability of the larger cascades with the branching process.

[4] $q(\hat{b}_{14})$ is calculated to be 0.949 with standard deviation 0.0005, so it is judged to be less than 0.95 but rounds to 0.95 in Table III.

It is not surprising that an approach estimating the parameters of a branching process model of cascading outperforms an empirical approach.

The preceding calculations analyze only the statistical error due to the cascading propagation and neglect any statistical uncertainty in the distribution of the initial outages. This can either be irrelevant or a significant part of the total statistical error. The statistical uncertainty in the distribution of the initial outages is irrelevant when the number or distribution of initial outages is assumed. The statistical uncertainty in the distribution of the initial outages can be a significant part of the total statistical error if the distribution of initial outages is also estimated from data. However, the distribution of initial outages can be obtained by conventional risk analysis, and there is always more data on the initial outages than on the subsequent cascading outages.

VIII. Conclusions

We analyze 12.4 years of transmission line outage data recorded by a North American utility. The key information used is the timing of each outage, and this is included in the standard Transmission Availability Data System (TADS) data that is reported to NERC by all American transmission owners. The outages are grouped into cascades and generations within each cascade based on the outage timings. Then the propagation of outages at each generation is estimated. The propagations are the parameters of a branching process model of the cascading. Given some observed or assumed initial outages, the branching process model can estimate the distribution of the total number of line outages. The number of line outages is a measure of the blackout extent. Thus the method quantifies the effect of cascading failure on the blackout extent based on standard utility data. The sensitivity of the results to the propagation and the confidence intervals of the results are computed. This new method of cascading failure analysis is practical and the computations are easy to implement with computer algebra.

For this utility data set we conclude that:

1) The distribution of the total number of line outages predicted with the branching process matches well the empirical distribution of the total number of outages. This validates the branching process model for predicting the distribution of the total number of outages in the sense that it is consistent with this data set. That is, a branching process model that accounts for the varying propagation as the cascade progresses can give a good

TABLE II
Derivative $D_{\lambda_k} b_r$ of the distribution of total number of outages with respect to λ_k

	number of lines outaged r										
	1	2	3	4	5	6	7	8	9	10	11
$D_{\lambda_1} b_r$	-0.73	0.29	0.16	0.087	0.054	0.036	0.024	0.016	0.012	0.0086	0.0062
$D_{\lambda_2} b_r$	0	-0.089	0.0069	0.016	0.013	0.010	0.0079	0.0060	0.0045	0.0035	0.0028
$D_{\lambda_3} b_r$	0	0	-0.020	-0.0052	0.00048	0.0021	0.0025	0.0024	0.0021	0.0018	0.0016
$D_{\lambda_4} b_r$	0	0	0	-0.0055	-0.0034	-0.0014	-0.00029	0.00030	0.00059	0.00071	0.00072
$D_{\lambda_{5+}} b_r$	0	0	0	0	-0.0016	-0.0022	-0.0021	-0.0018	-0.0014	-0.0010	-0.00075

TABLE III
Distributions of line outages and their standard deviations and probabilities of estimating within a factor of two

	predicted distribution			empirical distribution		
r	b_r	$\sigma(\hat{b}_r)$	$q(\hat{b}_r)$ (1 year)	p_r	$\sigma(\hat{p}_r)$	$q(\hat{p}_r)$ (10 year)
1	0.733	0.00354	1.00	0.730	0.00559	1.00
2	0.150	0.00205	1.00	0.150	0.00449	1.00
3	0.0536	0.00101	1.00	0.051	0.00277	1.00
4	0.0251	0.000592	1.00	0.0287	0.00210	1.00
5	0.0130	0.000386	1.00	0.0114	0.00134	1.00
6	0.00722	0.000272	1.00	0.0095	0.00122	1.00
7	0.00442	0.000203	1.00	0.00538	0.000921	1.00
8	0.00296	0.000157	1.00	0.00348	0.000741	0.99
9	0.00207	0.000127	0.99	0.00190	0.000548	0.96
10	0.00152	0.000104	0.99	0.00127	0.000448	0.87
11	0.00119	0.0000858	0.98	0.00127	0.000448	0.87
12	0.000937	0.0000721	0.97	0.00095	0.000388	0.84
13	0.000744	0.0000615	0.96	0.000792	0.000354	0.74
14	0.000597	0.0000530	0.95	0.000792	0.000354	0.74
15	0.000487	0.0000461	0.93	0.000475	0.000274	0.60
16	0.000403	0.0000404	0.92	0.000633	0.000317	0.79
17	0.000339	0.0000357	0.90	0.000158	0.000158	0.36
18	0.000288	0.0000318	0.89	0.000317	0.000224	0.72
19	0.000248	0.0000285	0.88	0.000158	0.000158	0.36
20	0.000216	0.0000257	0.87	0.000158	0.000158	0.36

prediction of the distribution of the total number of line outages from the initial number of line outages.

2) The prediction of the distribution of total number of line outages with the branching process requires significantly fewer observed cascades than empirical estimation, and useful results can be obtained from standard annual data from a large utility. The requirement of fewer observed cascades suggests that the monitoring of line outage propagation and cascading failure risk is practical.
3) Propagation of line outages increases as the cascade progresses and then appears to level out.
4) To mitigate long cascades it appears to be more effective to reduce the amount of propagation at the early generations of cascading. (This assumes that the mitigation only adjusts the early propagations; it is not yet known how typical mitigation methods affect the propagation at different generations.)

Our statistical analysis is significant in determining over what area and what length of time line outage data must be gathered in order to make valid conclusions about the propagation and extent of cascading in that area. This is necessary to determine whether valid assessments of cascading risk can be made in more localized areas of the system to indicate whether the localized areas require additional cascading risk mitigation. The statistical analysis also enables a more rational risk-based assessment that is better than the extremes of overreacting to occasional large events or neglecting cascading risk when there is a period of time with no large events.

This paper quantifies propagation and cascading extent from data observed over a time period. This can indicate areas of high propagation that are candidates for mitigation of cascading risk, and can confirm after implementation of a mitigation how the propagation was changed. There remains a need to verify the performance of a proposed mitigation method before it is implemented, and this can be done by applying similar methods [16], [17] to simulations of cascading processes on power system models with and without the mitigation. Moreover, quantifying the cascading propagation opens the possibility of directly assessing the effect of mitigation on the propagation, which complements the more traditional reliability focus on mitigating the initiating outages of the cascading.

There are several directions for future work. In this paper we analyze one publicly available data set. If access to other data sets can be obtained, they should be analyzed to better determine the general characteristics of propagation of cascading line outages and provide further validation and experience with the method. In this paper, all the line outages are regarded as the same, and only the timing of line outages is considered. Also we statistically analyze the propagation of line outages in all the cascades, regardless of whether load is shed or whether the cascade is short or long. The value of making these sort of distinctions and possibly elaborating the data processing could be examined if larger data sets become available. It would be valuable to extend the prediction of the number of lines outaged due to cascading to other measures of blackout size such as load shed [17]. We also expect the quantitative description of cascading line propagation based on observed data to be useful for validating cascading failure simulations.

ACKNOWLEDGEMENT

We gratefully thank Bonneville Power Administration for making publicly available the outage data that made this paper possible. The analysis and any conclusions are strictly those of the author and not of Bonneville Power Administration. We gratefully thank anonymous reviewers for advice improving the paper.

REFERENCES

[1] U.S.-Canada Power System Outage Task Force, Final Report on the August 14th blackout in the United States and Canada. United States Department of Energy and National Resources Canada, April 2004.
[2] NERC (North American Electric Reliability Council). 1996 system disturbances. (Available from NERC, Princeton Forrestal Village, 116-390 Village Boulevard, Princeton, New Jersey 08540-5731), 2002.
[3] Union for the Co-ordination of Transmission of Electricity (UCTE). Final report, system disturbance on 4 November 2006. www.ucte.org.
[4] T.E. Harris, Theory of branching processes, Dover NY 1989.
[5] I. Dobson, B.A. Carreras, D.E. Newman, A branching process approximation to cascading load-dependent system failure, 37th Hawaii Intl. Conference on System Sciences, Hawaii, HI Jan. 2004.
[6] I. Dobson, B.A. Carreras, D.E. Newman, Branching process models for the exponentially increasing portions of cascading failure blackouts, 38th Hawaii Intl. Conference on System Sciences, Hawaii, HI, Jan. 2005.
[7] B.A. Carreras, D.E. Newman, I. Dobson, A.B. Poole, Evidence for self organized criticality in a time series of electric power system blackouts, IEEE Trans. Circuits & Systems Part I, vol. 51, no. 9, Sept. 2004, pp. 1733-1740.
[8] I. Dobson, B.A. Carreras, V.E. Lynch, D.E. Newman, Complex systems analysis of series of blackouts: cascading failure, critical points, and self-organization, Chaos, vol. 17, no. 2, 026103 June 2007.
[9] D. Cornforth, Long tails from the distribution of 23 years of electrical disturbance data, IEEE Power Systems Conference and Exposition, Seattle WA, March 2009.
[10] P. Hines, J. Apt, S. Talukdar, Large blackouts in North America: historical trends and policy implications, Energy Policy, vol. 37, no. 12, December 2009, pp. 5249-5259.
[11] B.A. Carreras, V.E. Lynch, I. Dobson, D.E. Newman, Critical points and transitions in an electric power transmission model for cascading failure blackouts, Chaos, vol. 12, no. 4, 2002, pp. 985-994.
[12] B.A. Carreras, V.E. Lynch, I. Dobson, D.E. Newman, Complex dynamics of blackouts in power transmission systems, Chaos, vol. 14, no. 3, September 2004, pp. 643-652.
[13] J. Chen, J.S. Thorp, I. Dobson, Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model, Intl. Journal Electrical Power & Energy Systems, vol. 27, no. 4, May 2005, pp. 318-326.
[14] D.P. Nedic, I. Dobson, D.S. Kirschen, B.A. Carreras, V.E. Lynch, Criticality in a cascading failure blackout model, Intl. Journal Electrical Power and Energy Systems, vol. 28, 2006, pp. 627-633.
[15] S. Mei, X. Zhang, M. Cao, Power grid complexity, Springer 2011.
[16] I. Dobson, J. Kim, K. R. Wierzbicki, Testing branching process estimators of cascading failure with data from a simulation of transmission line outages, Risk Analysis, vol. 30, no. 4, 2010, pp. 650 - 662.
[17] J. Kim, K.R. Wierzbicki, I. Dobson, R.C. Hardiman, Estimating propagation and distribution of load shed in simulations of cascading blackouts, to appear in *IEEE Systems Journal*, special issue on complex systems.
[18] H. Ren, I. Dobson, Using transmission line outage data to estimate cascading failure propagation in an electric power system. IEEE Trans. Circuits and Systems Part II, 2008, vol. 55, no. 9, pp. 927-931.
[19] I. Dobson, B.A. Carreras, D.E. Newman, A loading-dependent model of probabilistic cascading failure, Probability in the Engineering and Informational Sciences, vol. 19, no. 1, 2005, pp. 15-32.
[20] J. Kim, I. Dobson, Approximating a loading-dependent cascading failure model with a branching process, IEEE Trans. Reliability, vol. 59, no. 4, December 2010, pp. 691-699.
[21] M. Gathy, C. Lefèvre, From damage models to SIR epidemics and cascading failures, Advances in Applied Probability, vol. 41, 2009, pp. 247-269.
[22] Q. Chen, C. Jiang, W. Qiu, J.D. McCalley, Probability models for estimating the probabilities of cascading outages in high-voltage transmission network, IEEE Trans. Power Systems, vol. 21, no. 3, Aug. 2006, pp. 1423-1431.

[23] R. Adler, S. Daniel, C. Heising, M. Lauby, R. Ludorf, T. White, An IEEE survey of US and Canadian overhead transmission outages at 230 kV and above, IEEE Trans. Power Delivery, vol. 9, no. 1, Jan. 1994, pp. 21-39.

[24] IEEE PES CAMS Task Force on Cascading Failure, Initial review of methods for cascading failure analysis in electric power transmission systems, IEEE Power Engineering Society General Meeting, Pittsburgh PA USA, July 2008.

[25] G.L. Greig, Second moment reliability analysis of redundant systems with dependent failures, Reliability Engineering and System Safety, vol. 41, 1993, pp. 57-70.

[26] M.E.J. Newman, The structure and function of complex networks, SIAM Review, vol. 45, no. 2, 2003, pp. 167-256.

[27] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics, Physics Reports, vol. 424, no. 4-5, February 2006, pp. 175-308.

[28] S. Roy, C. Asavathiratham, B. C. Lesieutre, G. C. Verghese, Network models: growth, dynamics, and failure, 34th Hawaii International Conference on System Sciences, Maui, Hawaii, Jan. 2001.

[29] D.J. Watts, A simple model of global cascades on random networks, Proceedings of the National Academy of Sciences USA, vol. 99, no. 9, 2002, pp. 5766-5771.

[30] A.E. Motter, Y-C. Lai, Cascade-based attacks on complex networks, Physical Review E, vol. 66, no. 6, 065102, 2002.

[31] P. Crucitti, V. Latora, M. Marchiori, Model for cascading failures in complex networks, Physical Review E, vol. 69, 045104(R), 2004.

[32] B.C. Lesieutre, S. Roy, V. Donde, A. Pinar, Power system extreme event screening using graph partitioning, 38th North American Power Symposium, Southern Illinois University, Carbondale IL, Sept. 2006.

[33] D.V. Lindley, N.D. Singpurwalla, On exchangeable, causal and cascading failures, Statistical Science, 2002, vol. 17, no. 2, pp. 209-219.

[34] A.W. Swift, Stochastic models of cascading failure, Journal of Applied Probability, vol. 45, 2008, pp. 907-921.

[35] Y. Sun, L. Ma, J. Matthew, S. Zheng, An analytical model for interactive failures, Reliability Engineering and System Safety, vol. 91, 2006, pp. 495-503.

[36] I. Dobson, B.A. Carreras, Number and propagation of line outages in cascading events in electric power transmission systems, 48th Annual Allerton Conference on Communication, Control and Computing, Monticello IL USA, September 2010.

[37] Bonneville Power Administration Transmission Services Operations & Reliability website http://transmission.bpa.gov/Business/Operations/Outages

[38] K.B. Athreya, P.E. Ney, Branching Processes, Dover NY 2004 (reprint of Springer-verlag Berlin 1972).

[39] K. Ulm, A simple method to calculate the confidence interval of a standardized mortality ratio (SMR), American Journal of Epidemiology, vol. 131, no. 2, 1990, pp. 373-375.

[40] N.L. Johnson, A.W. Kemp, S. Kotz, Univariate Discrete Distributions, Third Edition, Wiley Hoboken NJ, 2005.

[41] S. Wolfram, The Mathematica Book, 5th edition, Wolfram Media, 2003.

[42] N.M. Yanev, On the statistics of branching processes, Theory of Probability and its Applications, vol. 20, 1975, pp. 612-622.

APPENDIX
BRANCHING PROCESS CALCULATIONS

The branching process used in this paper is a Galton-Watson branching process with Poisson offspring distribution whose mean varies with generation. There is an arbitrary distribution of initial outages in generation zero. Some familiarity with branching processes is desirable and the initial chapters of classic references [4], [38] are recommended.

A. Calculating the distribution of the total number of outages

The offspring distribution is the probability distribution of the number of outages in a generation assuming that there is a single parent outage in the previous generation. Let the generating function of the offspring distribution producing generation k outages from generation $k-1$ outages be $f_k(s)$. Generating functions are formal power series in the variable s and the coefficients of the powers of s are the probabilities of the offspring distribution.

For the initial generation, $f_0(s)$ is the generating function of the distribution of initial outages p_{0r}.

$$f_0(s) = \sum_{r=1}^{\infty} p_{0r} s^r \tag{8}$$

In (8) the probability p_{0r} of r initial outages is the coefficient of s^r. The distribution of initial outages can be estimated empirically from (2). Alternatively, a particular distribution of outages can be assumed. For example, if there are 5 initial line outages, then $p_{05} = 1$ and $f_0(s) = s^5$.

For generation $k \geq 1$,

$$f_k(s) = e^{\lambda_k(s-1)} = \sum_{r=0}^{\infty} e^{-\lambda_k} \frac{\lambda_k^r}{r!} s^r \tag{9}$$

is the generating function of the Poisson offspring distribution with mean λ_k. That is, assuming a single outage in generation $k-1$, the probability of r outages in generation k is $e^{-\lambda_k}\lambda_k^r/(r!)$, the coefficient of s^r in (9).

The offspring distributions determine the statistics of the branching process via functional composition of the generating functions. For example, if there are 5 initial outages so that $f_0(s) = s^5$, the distribution of the number of outages in generation 3 is given by the generating function

$$f_3(f_2(f_1(s^5))) = e^{-\lambda_1+\lambda_1 e^{-\lambda_2+\lambda_2 e^{-\lambda_3+\lambda_3 s^5}}} \tag{10}$$

The coefficients of the power series (10) give the probability distribution of the number of outages in generation 3. For example, the probability of 4 outages in generation 3 is the coefficient of s^4 in (10). Generating function (10) illustrates the most common calculation in branching processes, but here we are interested not in the number of outages in a given generation, but in the total number of outages in all generations, and we now present this calculation.

Consider a single line outage that occurs in generation k and let the total number of outages that are descendants of this outage in any subsequent generation (children plus grandchildren plus great grandchildren and so on) be Y_k. Let the generating function of Y_k be $F_k(s) = Es^{Y_k}$. The number of descendants of the single line outage plus the single line outage itself is $Y_k + 1$ and $Y_k + 1$ has generating function $sF_k(s)$. Then the basic recursion for computing all the descendants of an outage at a given generation is

$$F_{k-1}(s) = f_k(sF_k(s)) \tag{11}$$

Since $\lambda_k = \lambda_{5+}$ for $k \geq 5$, the total number of outages that are descendants of an outage in generation 4 plus the outage itself is given by a Borel distribution with parameter λ_{5+}. We write $f_B(s)$ for the generating function of the Borel distribution with parameter λ_{5+}:

$$f_B(s) = \sum_{r=0}^{\infty} (r\lambda_{5+})^{r-1} \frac{e^{-r\lambda_{5+}}}{r!} s^r \tag{12}$$

Write $F(s)$ for the generating function of the total number of outages. We wish to compute $F(s)$ to obtain the distribution of the total number of outages. Applying recursion (11) successively,

$$\begin{aligned} sF_4(s) &= f_B(s) \\ F_3(s) &= f_4(sF_4(s)) = f_4(f_B(s)) \\ F_2(s) &= f_3(sf_4(f_B(s))) \\ F_1(s) &= f_2(sf_3(sf_4(f_B(s)))) \\ F_0(s) &= f_1(sf_2(sf_3(sf_4(f_B(s))))) \\ F(s) &= f_0(sf_1(sf_2(sf_3(sf_4(f_B(s)))))) \end{aligned} \quad (13)$$

Equation (13) shows that $F(s)$ is a complicated power series in s, but it can be evaluated by computer algebra [41] for as many terms as needed. The recursive structure of (13) mirrors the generations of the cascade. One way to illustrate this for the case of one initial failure (so that $f_0(s) = s$) is by writing in full formula (13):

$$F(s) = se^{-\lambda_1 + \lambda_1 se^{-\lambda_2 + \lambda_2 se^{-\lambda_3 + \lambda_3 se^{-\lambda_4 + \lambda_4 f_B(s)}}}}$$

B. Error bars for propagation

Consider the estimation of propagation λ_k and its confidence interval. Because of the assumption of a branching process model, Z_k is the sum of Z_{k-1} independent Poisson(λ_k) random variables and is distributed as Poisson($Z_{k-1}\lambda_k$). The estimated mean $\hat{\lambda}_k$ is then Z_k/Z_{k-1}. Since the mean and variance of a Poisson distribution are equal, the variance of Z_k is also $Z_{k-1}\lambda_k$. Therefore the variance of $\hat{\lambda}_k$ is λ_k/Z_{k-1} and

$$\sigma(\hat{\lambda}_k) = \sqrt{\frac{\lambda_k}{Z_{k-1}}}, \quad k = 1, 2, 3, 4. \quad (14)$$

According to [39], using the correspondence between Poisson and chi-squared random variables, if there are Z_k observations of a Poisson random variable, then $(1-\alpha)$ confidence limits λ_L, λ_U on the estimate of the mean of the Poisson distribution are solutions of

$$\sum_{i=Z_k}^{\infty} e^{-\lambda_L} \frac{\lambda_L^i}{i!} = P[\chi^2_{2Z_k} \le 2\lambda_L] = \alpha/2 \quad (15)$$

$$\sum_{i=0}^{Z_k} e^{-\lambda_U} \frac{\lambda_U^i}{i!} = 1 - P[\chi^2_{2(Z_k+1)} \le 2\lambda_U] = \alpha/2 \quad (16)$$

Then $(1-\alpha)$ confidence limits on $\hat{\lambda}_k$ are $[\lambda_L/Z_{k-1}, \lambda_U/Z_{k-1}]$, except that λ_U/Z_{k-1} is calculated differently for $Z_k \ge 30$.

For $Z_k \ge 30$, solving (16) numerically is difficult, and the following normal approximation to the Poisson distribution is used [40]. Let $u_{\alpha/2}$ satisfy

$$\frac{1}{2\pi} \int_{u_{\alpha/2}}^{\infty} e^{-u^2/2} du = \alpha/2. \quad (17)$$

Then the upper confidence limit λ_U is

$$\lambda_U = Z_k + \tfrac{1}{2}u^2_{\alpha/2} + u_{\alpha/2}\sqrt{Z_k + \tfrac{1}{4}u^2_{\alpha/2}}\ , \quad (18)$$

and the upper confidence limit of $\hat{\lambda}_k$ is λ_U/Z_{k-1}.

We now consider estimate (4) for λ_{5+} and its uncertainty. Using the approach of [42], $\hat{\lambda}_{5+}$ has an asymptotically normal distribution with approximate variance

$$\sigma^2(\hat{\lambda}_{5+}) = \frac{(1-\lambda_{5+})(1-e^{-\lambda_{5+}})}{J_4(1-\lambda_{5+}^{16})}, \quad (19)$$

where J_4 is the number of cascades j with $Z_4^{(j)} \ne 0$.

C. Sensitivity to propagation

We compute the sensitivity of the distribution of the total number of outages to the propagation λ_k. It is convenient to rewrite (13) using functional composition notation

$$F = f_0 \circ S \circ f_1 \circ S \circ f_2 \circ S \circ f_3 \circ S \circ f_4 \circ f_B \quad (20)$$

Here the function S is multiplication by s. Write D for differentiation with respect to s and D_{λ_k} for differentiation with respect to λ_k, and note that

$$\begin{aligned} Df_k(s) &= \lambda_k f_k(s), \qquad k \ge 1 \\ D_{\lambda_k} f_k(s) &= (s-1) f_k(s), \qquad k \ge 1. \end{aligned}$$

Then (20) can be differentiated with respect to λ_k for $k = 1, 2, 3, 4, 5$ to give

$$\begin{aligned} D_{\lambda_1}F &= sDf_0(sF_0)F_0(sF_1 - 1) \\ D_{\lambda_2}F &= s^2\lambda_1 Df_0(sF_0)F_0F_1(sF_2 - 1) \\ D_{\lambda_3}F &= s^3\lambda_1\lambda_2 Df_0(sF_0)F_0F_1F_2(sF_3 - 1) \\ D_{\lambda_4}F &= s^4\lambda_1\lambda_2\lambda_3 Df_0(sF_0)F_0F_1F_2F_3(sF_4 - 1) \\ D_{\lambda_{5+}}F &= s^4\lambda_1\lambda_2\lambda_3\lambda_4 Df_0(sF_0)F_0F_1F_2F_3F_4 D_{\lambda_{5+}} f_B(s) \end{aligned}$$

To derive in more detail one of these cases, differentiating (20) with respect to λ_3 gives

$$\begin{aligned} D_{\lambda_3}F &= Df_0(sF_0)sDf_1(sF_1)sDf_2(sF_2)sD_{\lambda_3}f_3(sF_3) \\ &= s^3 Df_0(sF_0)Df_1(sF_1)Df_2(sF_2)D_{\lambda_3}f_3(sF_3) \\ &= s^3 Df_0(sF_0)\lambda_1 F_0 \lambda_2 F_1\,(sF_3 - 1)f_3(sF_3) \\ &= s^3\lambda_1\lambda_2 Df_0(sF_0)F_0F_1F_2(sF_3 - 1) \end{aligned}$$

D. Calculating $q(\hat{b}_r)$

$q(\hat{b}_r)$ is the probability that the estimated probability $\hat{b}_r$ lies within a factor of 2 of b_r. $q(\hat{b}_r)$ depends on the amount of data or, equivalently, the observation time, so the the first step is to fix the amount of data used. Then $q(\hat{b}_r)$ is calculated by repeated simulation of a branching process with initial outages distributed according to the empirical initial distribution p_0 and propagation given by (5).

The number of cascades simulated corresponds to the amount of data used. For example, if one year of data is assumed, then the amount of data is 6316/12.4=509 cascades. Then in principle one simulates 509 cascades to compute each sample of $\hat{b}_r$ (in practice, for each sample of $\hat{b}_r$ it is equivalent and easier to simulate one cascade with initial outages combined from 509 samples of the initial distribution p_0). To compute $q(\hat{b}_r)$, we simulate many samples of $\hat{b}_r$, count the number of samples of $\hat{b}_r$ within a factor of 2 of b_r, and then divide by the number of samples to obtain $q(\hat{b}_r)$. Our caculations used 200 000 or 400 000 samples of $\hat{b}_r$.

An alternative approach would assume that $\hat{b}_r$ is normally distributed with mean b_r and variance approximately given by linearizing the dependence of b_r on propagation:

$$\sigma^2(\hat{b}_r) = \sum_{k=1}^{5+} (D_{\lambda_k} b_r)^2 \sigma^2(\hat{\lambda}_k) \tag{21}$$

where $\sigma^2(\hat{\lambda}_k)$ would be obtained from (14). This alternative approach gives similar results to the simulation of $\hat{b}_r$, but we could not find a convincing justification of the assumption of normally distributed $\hat{b}_r$.

E. Comparing estimated and empirical distributions obtained from separate data

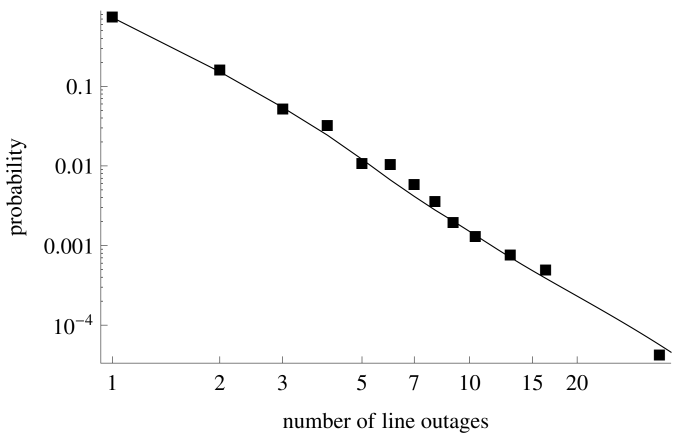

Fig. 8. Distribution of total number of outages from data in even numbered cascades (squares) and estimated using branching process obtained from odd numbered cascades (line joining the points of the discrete distribution).

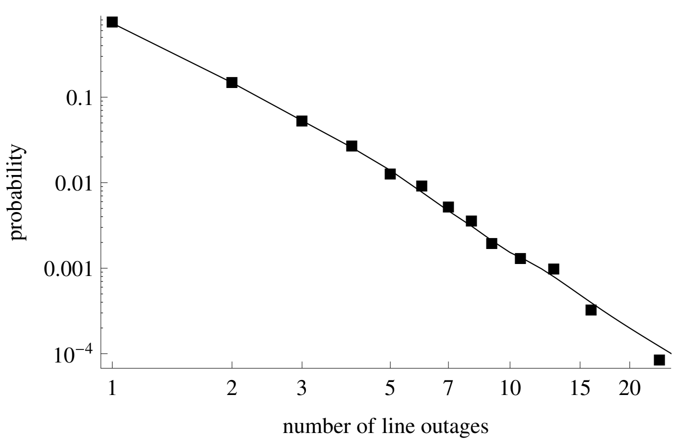

Fig. 9. Distribution of total number of outages from data in odd numbered cascades (squares) and estimated using branching process obtained from even numbered cascades (line joining the points of the discrete distribution).

Ian Dobson (F 06) received the BA in Mathematics from Cambridge University and the PhD in Electrical Engineering from Cornell University. He previously worked for the University of Wisconsin-Madison and is currently Sandbulte professor of electrical and computer engineering at Iowa State University. His interests are cascading failure risk, complex systems, blackouts, electric power system stability, synchrophasors, and nonlinear dynamics.