

2018

# Finding a Zipf distribution and cascading propagation metric in utility line outage data

Ian Dobson

*Iowa State University*, [dobson@iastate.edu](mailto:dobson@iastate.edu)

Follow this and additional works at: [https://lib.dr.iastate.edu/ece\\_pubs](https://lib.dr.iastate.edu/ece_pubs)



Part of the [Power and Energy Commons](#), and the [Systems and Communications Commons](#)

The complete bibliographic information for this item can be found at [https://lib.dr.iastate.edu/ece\\_pubs/199](https://lib.dr.iastate.edu/ece_pubs/199). For information on how to cite this item, please visit <http://lib.dr.iastate.edu/howtocite.html>.

---

This Article is brought to you for free and open access by the Electrical and Computer Engineering at Iowa State University Digital Repository. It has been accepted for inclusion in Electrical and Computer Engineering Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact [digirep@iastate.edu](mailto:digirep@iastate.edu).

---

# Finding a Zipf distribution and cascading propagation metric in utility line outage data

## **Abstract**

Observed transmission line outage data is grouped into successive generations of events. The empirical distribution of the number of generations in the cascades follows a Zipf distribution that implies the increasing propagation as cascades progress. The slope of the Zipf distribution gives a System Event Propagation Slope Index (SEPSI). This new metric quantifies the cascade propagation, varies as expected, and determines the probabilities of small, medium, and large cascades.

## **Disciplines**

Electrical and Computer Engineering | Power and Energy | Systems and Communications

## **Comments**

This is a pre-print of the article Dobson, Ian. "Finding a Zipf distribution and cascading propagation metric in utility line outage data." *arXiv preprint arXiv:1808.08434* (2018). Posted with permission.

# Finding a Zipf distribution and cascading propagation metric in utility line outage data

Ian Dobson, Iowa State University, dobson@iastate.edu, August 2018

**Abstract**—Observed transmission line outage data is grouped into successive generations of events. The empirical distribution of the number of generations in the cascades follows a Zipf distribution that implies the increasing propagation as cascades progress. The slope of the Zipf distribution gives a System Event Propagation Slope Index (SEPSI). This new metric quantifies the cascade propagation, varies as expected, and determines the probabilities of small, medium, and large cascades.

## I. INTRODUCTION

Long sequences of cascading outages occasionally cause large blackouts of power transmission systems. The cascade sequence starts with initial outages and is followed by propagating outages [1]. Cascading risk mitigation should address both the initiation and propagation of outages, but there has been no single scalar metric that quantifies the propagation. We find in historical utility data that the distribution of the number of cascading generations characterizes the propagation, follows a Zipf distribution, and gives a new scalar metric of cascading propagation.

## II. PROCESSING UTILITY DATA INTO GENERATIONS

The transmission line outage data consists of 10942 automatic line outages recorded over 14 years by a North American utility [2], [3]. The data includes the line outage start time to the nearest minute. This data is standard and routinely collected by utilities worldwide, such as in the Transmission Availability Data System (TADS) in North America [4], [5].

The historical outage data is grouped into cascades and generations based on the outage start time using the simple method described in [6]. 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 of outage events, 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 generational event. This processing produces 6687 cascades. The power system is usually resilient, so that most of these cascades are a single generation of outages that does not propagate further.

## III. PROPAGATION OF GENERATIONS AND THE ZIPF DISTRIBUTION OF THE NUMBER OF GENERATIONS

Instead of focussing on line outages as in previous work [6], here we analyze the propagation of generations of line outages.

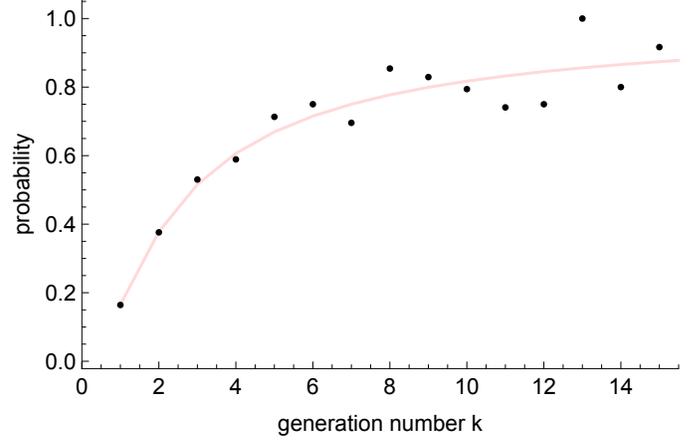


Fig. 1. Generation  $k$  propagation  $\rho_k$  as a function of  $k$ . Dots are utility data. Solid curve is calculated from the Zipf distribution solid line in Fig. 2.

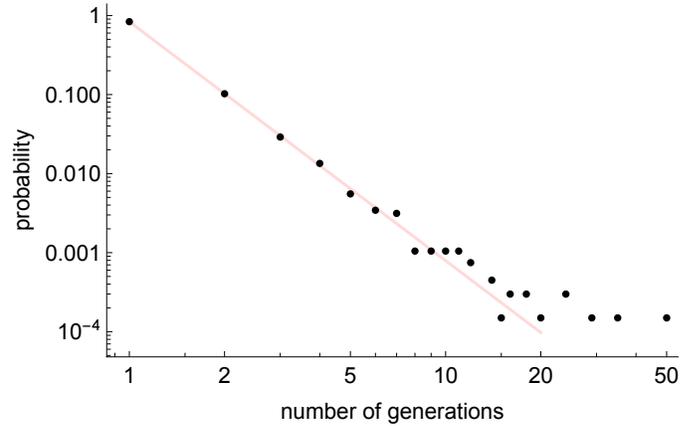


Fig. 2. Probability distribution of the number of generations on log-log plot. Dots are utility data. Solid line indicates Zipf distribution best fitting data.

Suppose a cascade has at least  $k$  generations of outages and let  $\rho_k$  be the probability that generation  $k$  propagates to produce a further generation  $k + 1$ . The dots in Fig. 1 show how the propagation  $\rho_k$  increases with  $k$  in the data.

It is hard to characterize the increasing propagation in Fig. 1 with a single number. Let us look at the propagation of generations in a different way. In Fig. 2 the dots show the empirical distribution of the number of generations  $G$ . The distribution of  $G$  is linear on this log-log plot and is a Zipf distribution (or zeta distribution) of the form

$$P[G = k] = \frac{1}{\zeta(s)} \frac{1}{k^s}, \quad k = 1, 2, 3, \dots \quad (1)$$

where the slope of the line is  $-s$  and  $\zeta$  is the Riemann zeta function. Indeed, the maximum likelihood method of [7] fits the data with the line of slope  $-3.02$  shown in Fig. 2 and a Pearson  $\chi^2$  test of goodness of fit shows consistency with the Zipf distribution with p-value 0.88.

Figs. 1 and 2 are different descriptions of the same propagation information, because the hazard function of  $G$  (the probability of the cascade stopping at generation  $k$ , given that generation  $k$  has been reached) is  $1 - \rho_k$ . Indeed, the propagation  $\rho_k$  implied by the fitted Zipf distribution of  $G$  is indicated by the solid curve in Fig. 1. Note how this solid curve interpolates the more erratic estimates from the sparse data for the higher generations.

#### IV. SYSTEM EVENT PROPAGATION SLOPE INDEX

We propose using the negative of the slope of the line fitted to the distribution of the number of generations on the log-log plot as a cascading metric called the System Event Propagation Slope Index (SEPSI). For example, the slope of the fitted line in Fig. 2 is  $-3.02$ , so that  $\text{SEPSI} = 3.02$ . A *lower* value of SEPSI indicates a shallower slope and an *increased* probability of large cascades.

The data for all automatic outages can be divided into two parts according to whether the cascade occurs during a NOAA storm condition in the same weather zone, or during the summer months June to September, or during peak load hours 3 to 8 pm [8]. Table I shows SEPSI calculated for each condition. As expected, SEPSI is smaller and cascading is more severe for the stressed cases, with the lowest value of  $\text{SEPSI} = 2.2$  achieved when there are storms. Moreover, as the condition varies, the distribution of the number of generations remains linear on the log-log plot, as shown by the Pearson  $\chi^2$  p-values in Table I for the fit of the Zipf distribution.

We can use the Zipf distribution to deduce the probabilities of small, medium, and large cascades from SEPSI. These probabilities are conditional on a cascade starting. Define a small cascade as 3 or less generations, a medium cascade as 4 to 9 generations, and a large cascade as 10 or more generations (different cut-offs can be chosen). Then, substituting SEPSI for  $s$  in (1), we compute the probabilities  $p_{\text{small}} = \sum_{k=1}^3 P[G=k]$ ,  $p_{\text{medium}} = \sum_{k=4}^9 P[G=k]$ , and  $p_{\text{large}} = 1 - p_{\text{medium}} - p_{\text{small}}$  as shown in Table I. As expected,  $p_{\text{large}}$  varies with SEPSI by the largest factor.

To determine how many cascades need to be observed for a given accuracy of SEPSI, we bootstrapped samples from the Zipf distribution to find the approximate variation of the estimates of SEPSI. Suppose that there are  $n$  cascades and  $\text{SEPSI} = 3.0$  so that  $p_{\text{large}} = 0.005$ . Then we find that the 95% confidence interval of SEPSI is approximately  $\pm 5/\sqrt{n}$ . For example, with 95% confidence, 1000 cascades determine SEPSI within  $\pm 0.16$  and  $p_{\text{large}}$  within a factor of 1.5. 1000 cascades are needed for this accuracy because the larger cascades are rare. There are 782 cascades per year in our data, so that accumulating 1000 cascades takes 1.3 years.

#### V. CONCLUSIONS

The number of generations of events in a cascade is an indication of cascade size and severity. We discover in utility

line outage historical data that the distribution of the number of generations of events closely follows a Zipf distribution. This intriguing pattern in the data suggests using the negative of the slope of the Zipf distribution as a metric of cascading propagation, the System Event Propagation Slope Index or SEPSI. For our data the overall SEPSI is 3.0 and this reduces to 2.2 in the presence of storms. The probabilities of small, medium and large cascades can be computed from SEPSI. Observing roughly 1000 cascades seems to determine SEPSI and the probability of large cascades to a useful accuracy.

The results in this paper rely on line outages recorded by one utility. But there is a clear and testable possibility of generalization to cascading events of a variety of components in other power systems and other infrastructures. More general cascading events should be grouped into generations of events, and the empirical distribution of the number of the generations of cascading events should be obtained. If the distribution of the number of generations can be satisfactorily approximated by a straight line on a log-log plot, then the cascading propagation and the probabilities of small, medium, and large cascades can be quantified with SEPSI.

TABLE I  
SEPSI VARIES WITH CONDITIONS AND GIVES CASCADE PROBABILITIES

condition	SEPSI	p-value	$p_{\text{small}}$	$p_{\text{medium}}$	$p_{\text{large}}$
all	3.0	0.88	0.967	0.029	0.005
storms	2.2	0.29	0.877	0.086	0.037
no storms	3.1	0.93	0.972	0.025	0.004
summer	2.9	0.98	0.961	0.033	0.006
not summer	3.2	0.91	0.976	0.021	0.003
peak hours	2.7	0.93	0.946	0.044	0.010
non-peak hours	3.1	0.94	0.972	0.025	0.004

p-value indicates fit of data to Zipf distribution

We gratefully acknowledge funding from NSF grants 1609080, 1735354 and thank BPA for making outage data publicly available. The analysis and conclusions are strictly the author's and not BPA's.

#### REFERENCES

- [1] I. Dobson, D.E. Newman, Cascading blackout overall structure and some implications for sampling and mitigation, Intl. J. Elect. Power & Energy Systems, vol. 86, 2017, pp. 29-32.
- [2] BPA transmission services operations & reliability website <http://transmission.bpa.gov/Business/Operations/Outages>
- [3] I. Dobson et al., Obtaining statistics of cascading line outages spreading in an electric transmission network from standard utility data, IEEE Trans. Power Systems, vol. 31, no. 6, 2016, pp. 4831-4841.
- [4] NERC, Transmission Availability Data System (TADS) Data Reporting Instruction Manual, Aug. 2014.
- [5] J.J. Bian, S. Ekisheva, A. Slone, Top risks to transmission outages, IEEE PES General Meeting, National Harbor, MD USA, July 2014.
- [6] 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, 2012.
- [7] M.L. Goldstein, S.A. Morris, G.G. Yen, Problems with fitting to the power-law distribution, European Phys. Journal B, vol. 41, pp. 255-258, 2004.
- [8] I. Dobson et al., Exploring cascading outages and weather via processing historic data, Hawaii Intl. Conf. System Sciences, Big Island, HI, Jan. 2018.