Abstract—We use observed transmission line outage data to make a Markov influence graph that describes the probabilities of transitions between generations of cascading line outages, where each generation of a cascade consists of a single line outage or multiple line outages. The new influence graph defines a Markov chain and generalizes previous influence graphs by including multiple line outages as Markov chain states. The generalized influence graph can reproduce the distribution of cascade size in the utility data. In particular, it can estimate the probabilities of small, medium and large cascades. The influence graph has the key advantage of allowing the effect of mitigations to be analyzed and readily tested, which is not available from the observed data. We exploit the asymptotic properties of the Markov chain to find the lines most involved in large cascades and show how upgrades to these critical lines can reduce the probability of large cascades.

Index Terms—cascading failures, power system reliability, mitigation, Markov, influence graph.

I. INTRODUCTION

Cascading outages in power transmission systems can cause widespread blackouts. These large blackouts are infrequent, but are high-impact events that occur often enough to pose a substantial risk to society [1], [2]. The power industry has always analyzed specific outages and taken steps to mitigate cascading. However, especially for the largest blackouts of highest risk, the challenges of evaluating and mitigating cascading risk in a quantitative way remain.

There are two main approaches to evaluating cascading risk: simulation and analyzing historical utility data. Cascading simulations can predict some likely and plausible cascading sequences [3], [4]. However, only a subset of cascading mechanisms can be approximated, and simulations are only starting to be benchmarked and validated for estimating blackout risk [5], [6]. Historical outage data can be used to estimate blackout risk [2] and detailed outage data can be used to identify critical lines [7]. However it is clear that proposed mitigation cannot be tested and evaluated with historical data. In this paper, we process historical line outage data to form a Markovian influence graph that statistically describes the interactions between the observed outages. The Markovian influence graph can quantify the probability of different sizes of cascades, identify critical lines and interactions, and assess the impact of mitigation on the probability of different sizes of cascades.

Influence graphs describing the interactions between successive cascading outages were developed using simulated data in [8]–[11] and suggested for processing historical data in our initial conference paper [12]. These influence graphs describe the statistics of cascading data with networks whose nodes represent outages of single transmission lines and whose directed edges represent probabilistic interactions between successive line outages in cascades. The more probable edges correspond to the interactions between line outages that appear more frequently in the data. Cascades in the influence graph start with initial line outages at the nodes and spread probabilistically along the directed graph edges. Once the influence graph is formed from the simulated cascading data, it can be used to identify critical components and test mitigation of blackouts by upgrading the most critical components [9], [11]. Even for simulated cascade data, there remain challenges in extracting good statistics for the influence graph from limited data.

The probabilistic transitions in an influence graph define the transitions of a Markov chain. However, previous influence graphs made approximations in formulating this Markov chain. As well as outages of single lines, cascading data typically includes multiple line outages that occur nearly simultaneously. These multiple simultaneous outages cause problems in clearly defining Markov chain states and transitions because it is not clear how to obtain the transitions between single outages from transitions involving multiple outages. For example, if the outage of two lines causes an outage in the next generation, it is hard to tell which line caused the subsequent outage or whether it is actually the two lines together that caused the subsequent outage. To address this, [11] makes some approximations that are hard to clarify. And Qi [9] assumes that the subsequent outage is caused by the most frequent line outage, since there is no information about the causal relationship among the outages. Whereas Qi [10] considers the causal relationships among successive outages as hidden variables and uses an expectation maximization algorithm to estimate the interactions underlying the multiple outage data.

The Markovian influence graph of this paper generalizes and improves the previous influence graph work in several ways. In particular, the paper

- obtains a clearly defined Markovian influence graph that solves the problem of multiple simultaneous outages by using additional states with multiple outages. This generalized influence graph rigorously defines transitions of a transient Markov chain, creating opportunities to exploit the theory of transient Markov chains.
- computes the probabilities of small, medium and large cascades, better matching the historical data statistics than the initial work in [12].
II. FORMING THE MARKOVIAN INFLUENCE GRAPH FROM HISTORICAL OUTAGE DATA

We use detailed line outage data consisting of automatic transmission line outages recorded by a large North American utility over 14 years starting in 1999 that is publicly available [13]. The data specifies the line outaged and the outage time to the nearest minute. Similar detailed line outage data is routinely recorded by utilities worldwide, for example in the North American Transmission Availability Data System.

The first step in building an influence graph is to take many cascading sequences of transmission line outages and divide each cascade into generations of outages as detailed in [14]. Each cascade starts with initial line outages in generation 0, and continues with subsequent generations of line outages 1,2,3,... until the cascade stops. Each generation of line outages is a set of line outages that occur together on a fast time scale of less than one minute. Often there is only one line outage in a generation, but protection actions can act quickly to cause several line outages in the same generation. (Sometimes in a cascading sequence an outaged line recloses and outages in a subsequent generation. In contrast to [12], [14], here we neglect these repeated outages.) For our data, we group the 9,741 automatic transmission line outages into 6,687 cascades. Most of the cascades (87%) have one generation because initial outages often do not propagate further.

The influence graph represents cascading as a Markov chain \( X_0, X_1, \ldots \), in which \( X_k \) is the set of line outages in generation \( k \) of the cascade. We first illustrate the formation of the influence graph from cascading data with the simple example of four observed cascades involving three lines shown in Fig. 1. The first cascade has line 1 outaged in generation 0, line 3 outaged in generation 1, line 2 outaged in generation 2, and then the cascade stops with no lines (indicated by the empty set {}) outaged in generation 3. All cascades eventually stop by transitioning to and remaining in the state {} for all future generations. The second cascade has lines 1 and 3 outaging together in generation 1. The five states observed in the data are {}, {line 1}, {line 2}, {line 3}, {line 1, line 3}. We can estimate the probabilities of transitioning from state \( i \) to state \( j \) in the next generation by counting the number of those transitions in all the cascades and dividing by the number of occurrences of state \( i \). For example, the probability of transitioning from state \{line 1\} to state \{line 3\} is 1/3 and the probability of transitioning from state \{line 2\} to state \{line 1, line 3\} is 1/2. By working out all the transition probabilities from the data, we can make the network graph of the Markov process as shown in Fig. 1. The transitions between states with higher probability are shown with thicker lines. In this generalized influence graph, the nodes are sets of line outages and the edges indicate transitions or interactions between sets of line outages in successive generations of cascading. The influence graph is different than the physical grid network and cascades are generated in the influence graph by moving along successive edges, selecting them according to their transition probabilities.

As discussed in the introduction, introducing the state \{line 1, line 3\} in addition to the states with only one line outage avoids the problems in previous work in accounting for transitions to and from the simultaneous outages of line 1 and line 3 in terms of states with only one line outage.

Returning to our real utility data, there are 614 lines and the observed cascades have 1094 subsets of these lines that form the 1094 states \( s_0, s_1, \ldots, s_{1093} \). Among these 1094 states, 50% have multi-line outages. And among these multi-line outage states, about 20% are comprised of lines sharing no common buses. While in theory there are \( 2^{614} \) subsets of 614 lines, giving an impractically large number of states, we find in practice with our data that the number of states is less than twice the number of lines.

Let \( P_k \) be the Markov chain transition matrix for generation \( k \) whose entry \( P_k[i,j] \) is the conditional probability that the set of outaged lines is \( s_j \) in generation \( k + 1 \), given that the set of outaged lines is \( s_i \) in generation \( k \); that is,

\[
P_k[i,j] = P[X_{k+1} = s_j \mid X_k = s_i]
\]

(1)

The key task of forming the Markov chain is to estimate the transition probabilities in the matrix \( P_k \) from the cascading data. If one supposed that \( P_k \) does not depend on \( k \), a straightforward way to do this would first construct a counting matrix \( N \) whose entry \( N[i,j] \) is the number of transitions from \( s_i \) to \( s_j \) among all generations in all the cascades. Then \( P_k \) can be estimated as

\[
P_k[i,j] = \frac{N[i,j]}{\sum_j N[i,j]}
\]

(2)

However, we find that \( P_k \) must depend on \( k \) in order to reproduce the increasing propagation of outages observed in the data [14]. On the other hand, there is not enough
data to accurately estimate $P_k$ individually for each $k > 0$. Our solution to this problem involves both grouping together data for higher generations and having $P_k$ varying with $k$, as well as using empirical Bayesian methods to improve the required estimates of cascade stopping probabilities. The detailed explanation of this solution is postponed to section V, and until section V we assume that $P_k$ has already been estimated for each generation $k$ from the utility data. Forming the Markov chain transition matrix from the data in this way makes the Markovian assumption that the statistics of the lines outaged in a generation only depend on the lines outaged in the previous generation. This is a pragmatic assumption that yields a tractable data-driven probabilistic model of cascading.

One way to visualize the influence graph interactions between line outages in $P_k$ is to restrict attention to the interactions between single line states, and show these as the red network in Fig. 2. The gray network is the actual grid topology, and the gray transmission lines are joined by a red line of the thickness proportional to the probability of being in successive generations, if that probability is sufficiently large.

Let the row vector $\pi_k$ be the probability distribution of states in generation $k$, whose entry $\pi_k[i]$ is the probability that the set of outaged lines is $s_i$ in generation $k$; that is,

$$\pi_k[i] = P[X_k = s_i]$$  \hspace{1cm} (3)

Then the propagation of line outages from generation $k$ to generation $k + 1$ is given by

$$\pi_{k+1} = \pi_k P_k$$  \hspace{1cm} (4)

and, using (4), the distribution of states in generation $k$ depends on the initial distribution of states $\pi_0$ according to

$$\pi_k = \pi_0 P_0 P_1 P_2 ... P_{k-2} P_{k-1}$$  \hspace{1cm} (5)

III. COMPUTING THE DISTRIBUTION OF CASCADE SIZES AND ITS CONFIDENCE INTERVAL

We compute the distribution of cascade sizes from the Markov chain and check that it reproduces the empirical distribution of cascade sizes, and estimate its confidence interval with bootstrap methods.

We can measure the cascade size using its number of generations. Define the survival function of the number of generations in a cascade as

$$S(k) = P[\text{number of cascade generations} > k]$$  \hspace{1cm} (6)

Since $\pi_k[0]$ is the probability that a cascade is in state $s_0 = \{\}$ in generation $k$ and the probability that the cascade stops at or before generation $k$, we can compute $S(k)$ as follows, writing 1 for the column vector with all entries one, and $e_0$ for the column vector with one for the first entry (indexed by 0) and the remaining entries zero:

$$S(k) = 1 - \pi_k[0] = \pi_k(1 - e_0)$$  
$$= \pi_0 P_0 P_1 P_2 ... P_{k-2} P_{k-1}(1 - e_0)$$  \hspace{1cm} (7)

The initial state distribution $\pi_0$ can be estimated directly from the cascading data.

Then we can confirm that the influence graph reproduces the statistics of cascade size in the cascading data by comparing the survival function of the number of generations $S(k)$ computed from (7) with the empirical survival function computed directly from the cascading data as shown in Fig. 3. The Markov chain reproduces the statistics of cascade size closely, with a Pearson $\chi^2$ goodness-of-fit test $p$-value of 0.99.

We use bootstrap resampling [15] to estimate the variance of our cascade size results. A bootstrap sample resamples the observed cascades with replacement, reconstructs the Markov chain, and recomputes the probabilities of cascade sizes. The variance of the probabilities of cascade sizes is then obtained as the empirical variance of 500 bootstrap samples. When the probability $p$ of a given blackout size is used, it is multiplied by cost of that blackout size to obtain the risk of that blackout size. Therefore it is appropriate to use a multiplicative form of confidence interval specified by a parameter $\kappa$ so that a 95% confidence interval for an estimated probability $\hat{p}$ means
that $P[p/\kappa \leq \hat{p} \leq p\kappa] = 0.95$. The confidence interval for the estimated survival function is shown in Fig. 4. Since larger cascades are rarer than small cascades, the variation increases as the number of generations increases.

To apply and communicate the probability distribution of cascade size, it is convenient to combine sizes together to get the probabilities of small, medium, and large cascades, where a small cascade has 1 or 2 generations, a medium cascade has 3 to 9 generations, and a large cascade has 10 or more generations. The respective probabilities are calculated as $1 - S(2)$, $S(2) - S(9)$, and $S(9))$. The 95% confidence intervals of the estimated probabilities of small, medium, and large cascades are shown in Table I. The probability of large cascades is estimated within a factor of 1.5, which is adequate for the purposes of estimating large cascade risk, since the cost of large cascades is so poorly known: estimates of the direct costs of cascading blackouts vary by more than a factor of 2.

**TABLE I**

<table>
<thead>
<tr>
<th>cascade size</th>
<th>probability</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>small (1 or 2 generations)</td>
<td>0.3666</td>
<td>1.005</td>
</tr>
<tr>
<td>medium (3 to 9 generations)</td>
<td>0.0372</td>
<td>1.132</td>
</tr>
<tr>
<td>large (10 or more generations)</td>
<td>0.0022</td>
<td>1.440</td>
</tr>
</tbody>
</table>

IV. CRITICAL LINES AND CASCADE MITIGATION

A. The transmission lines involved in large cascades

The lines eventually most involved in large cascades can be calculated from the asymptotic properties of the Markov chain. While all cascades eventually stop, we can consider at each generation those propagating cascades that are not stopped at that generation. And the probability distribution of states involved in these propagating cascades converges to a probability distribution $d_\infty$, which is called the quasi-stationary distribution. $d_\infty$ can be computed directly from the transition matrices (as explained in Appendix A, $d_\infty$ is the left eigenvector corresponding to the dominant eigenvalue of the transition submatrix $Q_{1+9}$). That is, except for a transient that dies out after some initial generations, the participation of states in the cascading that continues past these initial generations is well approximated by $d_\infty$. Thus the high probability states corresponding to the highest probability entries in $d_\infty$ are the critical states most involved in the latter portion of large cascades. Since $d_\infty$ does not depend on the initial outages, the Markov chain is supplying information about the eventual cascading for all initial outages.

We now find the critical lines corresponding to these critical states by projecting the states onto the lines in those states. Let $\ell_k$ be the row vector whose entry $\ell_k[j]$ is the probability that line $j$ outages in generation $k$. Then

$$\ell_k[j] = \sum_{i, j \in s_i} \pi_k[i]$$

where the matrix $R$ projects states to lines according to

$$R[i, j] = \begin{cases} 1; & j \in s_i \\ 0; & j \notin s_i \end{cases}$$

Then the probability distribution of lines eventually involved in the propagating cascades that are not stopped is $c_\infty = d_\infty R$ and the critical lines most involved in the latter portion of large cascades correspond to the highest probability entries in $c_\infty$. Fig. 5 shows the probabilities in $c_\infty$ in order of decreasing probability. We identify the top ten lines as critical and as candidates for upgrading to decrease the probability of large cascades.

B. Modeling and testing mitigation in the Markov chain

A transmission line is less likely to fail due to other line outages after the line is upgraded, its protection is improved, or its operating limit is reduced. These mitigations have the effect of decreasing the probability of transition to states containing the upgraded line, and are an adjustment of the columns of the
transition matrix corresponding to these states. The mitigation is represented in the Markov chain by reducing the probability of transition to the state \( s \) containing the upgraded line by \((r/|s|)\%\), where \(|s|\) is the number of lines in the state. Thus the reduction is \( r\% \) if the state contains only the upgraded line, and the reduction is less if the state contains multiple lines.

We demonstrate using the Markov chain to quantify the impact of mitigation by upgrading the ten lines critical for large cascades identified in section IV-A with \( r = 80\% \). The effect of this mitigation on cascade probabilities is shown in Fig. 6. It shows that upgrading the critical lines reduces the probability of large cascades by 45\%, while the probability of medium cascades is slightly decreased and the probability of small cascades is slightly increased.

![Fig. 6. Cascade size distribution before (red) and after (light green) mitigating lines critical in propagating large cascades.](image)

To show the effectiveness of the method of identifying critical lines, we compare the mitigation effect of upgrading critical lines and upgrading ten random lines. Randomly upgrading ten lines only decreases the probability of large cascades by 11\% on average.

So far we have only considered upgrading the lines critical for propagating large cascades. Now, in order to discuss this mitigation of large cascades in a larger context, we briefly consider and contrast a different mitigation tactic of upgrading lines that are critical for initial outages. Since initial outages are caused by external causes such as storm, lightning, or misoperation, they often have different mechanisms and different mitigations than for propagating outages. A straightforward method to identify lines critical for initial outages selects the ten lines in the data with the highest frequencies of initial outage [12]. Upgrading these ten lines will reduce their initial outage frequencies and hence reduce the overall cascade frequency. This reduction of initial outage frequencies is represented by reducing the frequencies of states corresponding to these lines. (The corresponding state frequencies are reduced by \( r = 80\% \) \((r/|s|)\%\) for multi-line state \(|s|\)). The main effect is that by reducing the initial outage frequencies of the critical lines by 80\%, we reduce the frequency of all cascades by 19\%. In addition, this mitigation will change the probabilities of states \( \pi_0 \) after renormalizing the frequencies of states. It turns out for our case that there is no overlap between critical lines for initial outages and for propagation.

Changing the initial state distribution \( \pi_0 \) has no effect on the distribution of cascade sizes in the long-term. However, it directly reduces the frequency of all cascades. In contrast, mitigating the lines critical for propagating large cascades reduces the probability of large cascades relative to all cascades but has no effect on the frequency of all cascades. (Note that Fig. 6 shows the distribution of cascade sizes assuming that there is a cascade, but gives no information about the frequency of all cascades.)

In practice, a given mitigation measure can affect both the initial outages and the propagation of outages into large cascades. The combined mitigation effects can also be represented in the influence graph by changing both the initial state distribution and the transition matrix, but here it is convenient to discuss them separately.

V. ESTIMATING THE TRANSITION MATRIX

The Markov chain has an absorbing first state \( s_0 = \{\} \), indicating no lines outaged as the cascade stops and after the cascade stops. Therefore the transition matrix has the structure

\[
P_k = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
\bar{u}_k & Q_k
\end{bmatrix}
\]

(10)

where \( u_k \) is a column vector of stopping probabilities; that is, \( u_k[i] = P_k[i,0] \). \( Q_k \) is a submatrix of transition probabilities between transient states which contains the non-stopping probabilities. The first row of \( P_k \) is always \( e'_0 \), so the transition probabilities to be estimated are \( u_k \) and \( Q_k \) for each generation \( k \). The rows and columns of \( P_k \) are indexed from 0 to \( |S| - 1 \) and the rows and columns of \( Q_k \) are indexed from 1 to \( |S| - 1 \), where \( |S| \) is the number of states.

As summarized in section II after (1), we need to both group together multiple generations to get sufficient data and account for variation with generation \( k \). The statistics of the transition from generation 0 to generation 1 are different than the statistics of the transitions between the subsequent generations. For example, stopping probabilities for generation 0 are usually larger than stopping probabilities for subsequent generations [12]. Also, the data for the subsequent generations is sparser. Therefore, we construct from counts of the number of transitions from generation 0 to generation 1 a probability transition matrix \( \tilde{P}_0 \), and construct from the total counts of the number of transitions from all the subsequent generations a probability transition matrix \( \tilde{P}_1 \). Specifically, we first use the right-hand side of (2) to construct two corresponding empirical transition matrices, and then we update stopping probabilities by the empirical Bayes method and adjust non-stopping probabilities to obtain \( P_0 \) and \( P_1 \). Finally, we adjust \( \tilde{P}_0 \) and \( \tilde{P}_1 \) to match the observed propagation rates to obtain \( P_k \) for each generation \( k \).

A. Bayesian update of stopping probabilities

The empirical stopping probabilities are improved by an empirical Bayes method [16], [17] to help mitigate the sparse data for some of these probabilities. Since the method is applied to both \( \tilde{P}_0 \) and \( \tilde{P}_1 \), we simplify notation by writing \( P \) for either \( \tilde{P}_0 \) or \( \tilde{P}_1 \).
The matrix of empirical probabilities obtained from the transition counts $N[i,j]$ is

$$\hat{P}^{\text{counts}}[i,j] = \frac{N[i,j]}{\sum_j N[i,j]}$$ (11)

We construct $\hat{P}$ from $\hat{P}^{\text{counts}}$ in two steps. First, Bayesian updating is used to better estimate stopping probabilities and form a matrix $\hat{P}^{\text{bayes}}$. Second, the non-stopping probabilities in $\hat{P}^{\text{bayes}}$ are adjusted to form the matrix $\hat{P}$ to account for the fact that some independent outages are grouped into cascading outages when we group outage data into cascades.

We need to estimate the probability of the cascade stopping at the next generation for each state encountered in the cascade. For some of the states, the stopping counts are low, and cannot give good estimates of the stopping probability. However, by pooling the data for all the states we can get a good estimate of the mean probability of stopping over all the states. We use this mean probability to adjust the sparse counts in a conservative way. In particular, we form a prior that maximizes its entropy subject to the mean of the prior being the mean of the pooled data. This maximum entropy prior can be interpreted as the prior distribution that makes the least possible further assumptions about the data [18] [19].

a) Finding a maximum entropy prior: Assuming the stopping counts are independent with a common probability, the stopping counts follow a binomial distribution. Its conjugate prior distribution is the beta distribution, whose parameters are estimated using the maximum entropy method.

Let stopping counts $C_i$ be the observed number of transitions from state $s_i$ to $s_0$ ($i = 1, ..., |S| - 1$). Then $C_i = N[i,0]$. Let $n_i = \sum_{j=0}^{j=|S|-1} N[i,j]$ be the row sum of the counting matrix $N$. The stopping counts $C_i$ follow a binomial distribution with parameter $U_i$, with probability mass function

$$f_{C_i|U_i}(c_i|u_i) = \frac{n_i!}{c_i!(n_i - c_i)!} u_i^{c_i} (1 - u_i)^{n_i - c_i}$$ (12)

The conjugate prior distribution for the binomial distribution is the beta distribution. Accordingly, we use the beta distribution with hyperparameters $\beta_1, \beta_2$ for the stopping probability $U_i$:

$$f_{U_i}(u_i) = B(\beta_1, \beta_2) u_i^{\beta_1 - 1} (1 - u_i)^{\beta_2 - 1}$$ (13)

where $B(\beta_1, \beta_2) = \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1) \Gamma(\beta_2)}$. Alternative parameters for the beta distribution are its precision $m = \beta_1 + \beta_2$ and its mean $\mu = \frac{\beta_1}{\beta_1 + \beta_2}$. The entropy of the beta distribution is

$$\text{Ent}(m, \mu) = \log B(m, \mu) - (m(1 - \mu) - (m - 1)\psi(m)) - (m(1 - \mu) - 1)\psi(m - 1) + (m - 2)\psi(m)$$ (14)

where $\psi(x) = \frac{d\Gamma(x)}{dx}$ is the digamma function.

We want to estimate hyperparameters $\beta_1, \beta_2$ to make the beta distribution have maximum entropy subject to the mean being the average stopping probability of the pooled data. Let $\bar{u}$ be the average stopping probability of the pooled data:

$$\bar{u} = \frac{\sum_{i=1}^{i=|S|-1} c_i}{\sum_{i=1}^{i=|S|-1} n_i}$$ (15)

Then we can obtain hyperparameters $\beta_1, \beta_2$ by finding the $m > 0$ that maximizes $\text{Ent}(m, \bar{u})$ and evaluating $\beta_1 = m\bar{u}$ and $\beta_2 = m(1 - \bar{u})$. The hyperparameters used for $\hat{P}_0^{\text{bayes}}$ are $(\beta_1, \beta_2) = (2.18, 0.32)$, and the hyperparameters for $\hat{P}_1^{\text{bayes}}$ are $(\beta_1, \beta_2) = (1.10, 0.93)$.

b) Updating the observed data using the prior: The posterior distribution of the stopping probability $U_i$ is a beta distribution with parameters $c_i + \beta_1, n_i - c_i + \beta_2$. We use the mean of the posterior distribution as a point estimate of the stopping probability:

$$\hat{P}^{\text{bayes}}[i,0] = E(U_i|C_i = c_i) = \frac{c_i + \beta_1}{n_i + \beta_1 + \beta_2}$$ (16)

Fig. 7 shows a comparison between the empirical stopping probabilities and the updated stopping probabilities. Black dots are the empirical probabilities sorted in ascending order (if two probabilities are equal, they are sorted according to the total counts observed). Red dots are the updated stopping probabilities. As expected, the empirical probabilities with the fewest counts move towards the mean the most when updated. As the counts increase, the effect of the prior decreases and the updated probabilities tend to the empirical probabilities.

Equation (16) forms the first column of $\hat{P}^{\text{bayes}}$. Then the non-stopping probabilities in the rest of the columns of the $\hat{P}^{\text{counts}}$ matrix are scaled so that they sum to one minus the stopping probabilities of (16) to complete the matrix $\hat{P}^{\text{bayes}}$:

$$\hat{P}^{\text{bayes}}[i,j] = \frac{1 - \hat{P}^{\text{bayes}}[i,0]}{\sum_{r=1}^{r=|S|-1} \hat{P}^{\text{counts}}[i,r]} \hat{P}^{\text{counts}}[i,j], \ j > 0$$ (17)

This Bayesian updating is applied to form $\hat{P}_0^{\text{bayes}}$ for the first transition and $\hat{P}_1^{\text{bayes}}$ for the subsequent transitions.
B. Adjusting nonstopping probabilities for independent outages

The method explained in section II that groups outages into cascades has an estimated 6% chance that it groups independent outages into cascading outages [20]. These 6% of outages occur independently while the cascading of other outages proceeds and do not arise from interactions with other outages. The empirical data for the nonstopping probabilities includes these 6% of outages, and we want to correct this. Therefore, the non-stopping probabilities are modified by shrinking the probabilities in transition matrix by 6%, and sharing this equally among all the states. That is,

$$\bar{P}(i,j) = 0.94 \hat{P}^\text{bayes}(i,j) + \frac{0.06}{|S| - 1} (1 - \hat{P}^\text{bayes}(i,0)) \ (18)$$

where $\hat{P}^\text{bayes}$ indicates the transition matrices after the Bayesian update of section V-A. Notice that $\bar{P}$ is a probability matrix since $\sum_j \bar{P}(i,j) = 1$ for each $i$. A benefit is that this adjustment makes the submatrix $Q_k$ have non-zero off-diagonal entries, making $\bar{P}$ irreducible.

C. Adjustments to match propagation

The average propagation $\rho_k$ for generation $k$ [14] is estimated from the data using

$$\hat{\rho}_k = \frac{\text{Number of cascades with } > k + 1 \text{ generations}}{\text{Number of cascades with } > k \text{ generations}} = \frac{S(k+1)}{S(k)} = \frac{\pi_{k+1}(1-e_0)}{\pi_k(1-e_0)} \ (19)$$

An important feature of the cascading data is that average propagation $\rho_k$ increases with generation $k$ as shown in Table II. To do this, we need to form transition matrices for each of these generations that reproduce this propagation. We define a matrix $A_k$ to adjust $\bar{P}_0$ and $\bar{P}_{1+}$ so that the

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>PROPACTIONS OF GENERATIONS k = 0 TO 17</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\rho_k$</td>
<td>$0.13$</td>
</tr>
<tr>
<td>$\hat{\rho}_k$</td>
<td>$0.73$</td>
</tr>
</tbody>
</table>

propagation in $P_k$ matches the empirical propagation for each generation up to generation 8. For generation 9 and above, the empirical propagation for each generation is too noisy to use individually and we combine those generations to obtain a constant transition matrix. That is, $P_0 = P_0A_0$, $P_1 = P_1A_1$, ..., $P_8 = P_8A_8$, $P_{9+} = P_{1+}A_{9+}$. Then the transition matrices for all the generations are $P_0, P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_{9+}, P_{1+}, P_{9+}, ...$

The matrix $A_k$ has the effect of transferring a fraction of probability from the transient to stopping transitions and has the following form:

$$A_k = \begin{pmatrix}
1 & 0 & ... & 0 \\
0 & 1 - a_k & ... & 0 \\
... & ... & ... & ... \\
a_k & 0 & ... & 1 - a_k
\end{pmatrix} \ (20)$$

$a_k$ is determined from the estimated propagation rate $\hat{\rho}_k$ as follows. Using (19), we have

$$\hat{\rho}_k = \frac{\pi_k P_{Ak}(1-e_0)}{\pi_k(1-e_0)} = (1 - a_k) \frac{1 - \pi_k \bar{P}e_0}{1 - \pi_k e_0} \ (21)$$

and we solve (21) to obtain $a_k$ for each generation $k$.

VI. DISCUSSION AND CONCLUSION

We process observed transmission line outage utility data to form a generalized influence graph and the associated Markov chain that statistically describe cascading outages in the data. Successive line outages, or, more precisely, successive sets of near simultaneous line outages in the cascading data correspond to transitions between nodes of the influence graph and transitions in the Markov chain. The more frequently occurring successive line outages in the cascading data give a stronger influence between nodes and higher transition probabilities. The generalized influence graph introduces additional states corresponding to multiple line outages that occur nearly simultaneously. This innovation adds a manageable number of additional states and solves some problems with previous influence graphs, making the formation of the Markov chain clearer and more rigorous.

One of the inherent challenges of cascading is the sparse data for large cascades. We have used several methods to partially alleviate this when estimating the Markov chain transition matrices, including combining data for several generations, conservatively improving estimates of stopping probabilities with an empirical Bayes method, accounting for independent outages during the cascade, and matching the observed propagation for each generation. The combined effect of these methods is to improve estimates of the Markov chain transition matrices. Although some individual elements of these transition matrices are nevertheless still poorly estimated, what matters is the variability of the results from the Markov chain, which are the probabilities of small, medium and large cascades. We assess the variability of these estimated probabilities with a bootstrap and find them to be estimated to a useful accuracy.

The Markov chain only models the statistics of successive transitions in the observed data. Also, there is an inherent limitation of not being able to account for outages and transitions not present in the observed data. However, the Markov chain can produce, in addition to the observed cascades, combinations of the observed transitions that are different than the observed cascades. That is, the Markov chain approximates the statistics of cascading rather than reproducing only the observed cascades. The asymptotic properties of the Markov chain can be exploited to calculate the transmission lines most involved in the propagation of larger cascades, and we show that upgrading these lines can significantly reduce the probability of large cascades.

We show how to estimate the Markov chain from observed utility data. Being driven by observed data has some advantages of realism. In particular, and in contrast with simulation approaches, no assumptions about the detailed mechanisms of cascading need to be made. Since the Markov chain driven by utility data has different assumptions than simulation,
we regard the Markov chain and simulation approaches as complementary. The Markov chain driven by observed data offers another way to test proposed mitigations of cascading that can predict the effect of the mitigation on the probabilities of small, medium, and large cascades.

APPENDIX A
DERIVING THE QUASI-STATIONARY DISTRIBUTION \( d_\infty \)

The quasi-stationary distribution can be derived in a standard way [21], [22]. Let \( d_k \) be a vector with entry \( d_k[i] \) which is the probability that a cascade is in nonempty state \( s_i \) at generation \( k \) given that the cascade is propagating, that is

\[
d_k[i] = \frac{P[X_k = s_i]}{P[X_k \neq s_0]} = \frac{\pi_k[i]}{1 - \pi_k[0]} , \quad i = 1, \ldots, |S|
\]

Then the quasi-stationary distribution is \( d_\infty = \lim_{k \to \infty} d_k \).

Diagonal entries of \( Q_{1+} \) corresponding to \( P_{1+} \) are all zero and all other entries are positive. According to the Perron–Frobenius theorem [23], \( Q_{1+} \) has a unique maximum modulus eigenvalue \( \lambda \), which is real, positive and simple with left eigenvector \( \nu' \). By normalizing \( \nu' \), we make \( \nu' \) a probability vector. We write \( \nu \) for the corresponding right eigenvector. Moreover, \( 0 < \mu < 1 \) and \( \mu \) is strictly greater than the modulus of the other eigenvalues of \( Q_{1+} \). Suppose the cascade starts with probability distribution \( \pi_0 \) (note that \( \pi_0[0] = 0 \)). According to (5), the probability of being in state \( i \) at generation \( k \) is \( \pi_k[i] = (\pi_0 P_1 P_2 \ldots P_{k-2} P_{k-1})[i] = (\pi_0 P^{(k-1)})[i] \).

In particular, the probability that the cascade terminates by generation \( k \) is \( \mu_0 i \) and \( \mu_0 = \frac{\nu_0}{\pi_0} \) for the corresponding right eigenvector. Then for \( i = 1, \ldots, |S| \),

\[
d_{k+1}[i] = \frac{\pi_{k+1}[i]}{1 - \pi_{k+1}[0]} = \frac{(\pi_0 P^{(k)}[i])}{1 - \pi_0 P^{(k)} \nu_0} = \frac{(\pi_0 P^{(k)})[i]}{\pi_0 P^{(k)}(1 - \nu_0)}
\]

The first row of \( P_k \) is always \([1 \ 0 \ldots \ 0]\). Since \( \pi_0[0] = 0 \), let \( \pi_0 = [0 \ \pi_0] \). Then \( \pi_0 P^{(1)}[1 - \nu_0] = \pi_0 Q^{(1)}[1] \) and \( (\pi_0 P^{(k)})[i] = (\pi_0 Q^{(k)})[i] \) for \( i = 1, \ldots, |S| \). And \( Q^{(k)} = Q_0 Q_{1+} \sum_{m=0}^{k} \mu^{(1 - \alpha_m)} \), so that

\[
d_\infty = \lim_{k \to \infty} d_{k+1} = \lim_{k \to \infty} \frac{\bar{p}_0 Q^{(k)}}{\bar{p}_0 Q^{(k)} Q_1 + 1}
\]

\[
\lim_{k \to \infty} \frac{\bar{p}_0 Q_0 Q_1 \sum_{m=0}^{k} \mu^{(1 - \alpha_m)}}{\bar{p}_0 Q_0 \mu^{(1 - \alpha_m)} Q_1 + 1} = \nu'
\]

where \( Q^{(k-1)} \to \mu^{(1 - \alpha_m)} \nu' \) as \( k \to \infty \). Therefore, the dominant left eigenvector of \( Q_{1+} \) is \( d_\infty \).

For our data, the top three eigenvalues in modulus are \( \mu = 0.502 \) and \( -0.136 \pm 0.122i \) with corresponding moduli \( 0.502 \) and \( 0.381 \).

ACKNOWLEDGEMENT

We gratefully thank BPA for making the outage data public. The analysis and any conclusions are strictly the author’s and not BPA’s. We gratefully acknowledge support in part from NSF grants 1609080 and 1735354.

REFERENCES