REACTIVE POINT PROCESSES:
A NEW APPROACH TO PREDICTING POWER FAILURES IN UNDERGROUND ELECTRICAL SYSTEMS

BY ŞEYDA ERTEKIN, CYNTHIA RUDIN AND TYLER MCCORMICK

Massachusetts Institute of Technology and University of Washington

Reactive point processes (RPP’s) are a new statistical model designed for predicting discrete events, incorporating self-exciting, self-regulating, and saturating components. The self-excitement occurs as a result of a past event, which causes a temporary rise in vulnerability to future events. The self-regulation occurs as a result of an external “inspection” which temporarily lowers vulnerability to future events. RPP’s can saturate when too many events or inspections occur close together, which ensures that the probability of an event stays within a realistic range. RPP’s were developed to handle an important problem within the domain of electrical grid reliability: short term prediction of electrical grid failures (“manhole events”), including outages, fires, explosions, and smoking manholes, which can cause threats to public safety and reliability of electrical service in cities. For the self-exciting, self-regulating, and saturating elements of the model, we develop a conditional frequency estimator and also introduce a class of flexible parametric functions reflecting how the influence of past events and inspections on vulnerability levels gradually fades over time. We use the model to predict power grid failures in Manhattan over a short term horizon.

1. Introduction. We present a new statistical model for predicting discrete events over time, called Reactive Point Processes (RPP’s). RPP’s are a natural fit for many different domains, and their development was motivated by the problem of predicting serious events (fires, explosions, power failures) in the underground electrical grid of New York City (NYC). RPP’s capture several important properties of power failures on the grid:

Keywords and phrases: Point processes, Self-exciting processes, Energy grid reliability, Bayesian analysis, Time series
• There is an instantaneous rise in vulnerability to future serious events immediately following an occurrence of a past serious event, and the vulnerability gradually fades back to the baseline level. This is a type of *self-exciting* property.

• There is an instantaneous decrease in vulnerability due to an “inspection,” repair, or other action taken. The effect of this inspection fades gradually over time. This is a *self-regulating* property.

• The cumulative effect of events or inspections can saturate, ensuring that vulnerability levels never stray too far beyond their baseline level. This captures *diminishing returns* of many events or inspections in a row.

• The baseline level can be altered if there is at least one past event.

• Vulnerability between similar entities should be similar. RPP’s can be incorporated into a Bayesian framework that shares information across observably similar entities.

RPP’s extend self-exciting point processes (SEPP’s), which have only the self-exciting property mentioned above. Self-exciting processes date back at least to the 1960’s (Bartlett, 1963; Kerstan, 1964). The applicability of self-exciting point processes for modeling and analyzing time-series data has stimulated interest in diverse disciplines, including seismology (Ogata, 1988, 1998), criminology (Mohler et al., 2011; Egesdal et al., 2010; Lewis et al., 2010; Louie, Masaki and Allenby, 2010), finance (Chehrazi and Weber, 2011; Aït-Sahalia, Cacho-Diaz and Laeven, 2010; Bacry et al., 2013; Filimonov and Sornette, 2012; Embrechts, Liniger and Lin, 2011; Hardiman, Bercot and Bouchaud, 2013), computational neuroscience (Johnson, 1996; Krumin, Reutsky and Shoham, 2010), genome sequencing (Reynaud-Bouret and Schbath, 2010), and social networks (Crane and Sornette, 2008; Mitchell and Cates, 2009; Simma and Jordan, 2010; Masuda et al., 2012; Du et al., 2013). These models appear in so many different domains because they are a natural fit for time series data where one would like to predict discrete events in time, and where the occurrence of a past event gives a temporary boost to the probability of an event in the future.

RPP’s can be used in an even broader array of settings, as they can handle more than self-excitation. For instance, as mentioned above, RPP’s can take into account external influences (“inspections”) that cause vulnerability levels to drop, and where effects can saturate. Inspections are made according to a predetermined policy of an external source, which may be deterministic or random. In the application that self-exciting point processes are the most well known for, namely
earthquake modeling, it is not possible to take an action to preemptively reduce the risk of an earthquake; however, in other applications it is clearly possible to do so. In our power failure application, power companies can perform preemptive inspections and repairs in order to decrease electrical grid vulnerability. In neuroscience, it is possible to take an action to temporarily reduce the firing rate of a neuron. There are many actions that police can take to temporarily reduce crime in an area (e.g., temporary increased patrolling or monitoring). In medical applications, doses of medicine can be preemptively applied to reduce the probability of a cardiac arrest or other event. Or, for instance, the self-regulation can come as a result of the patient’s lab tests or visits to a physician.

Another way that RPP’s expand upon SEPP’s is that they allow deviations from the baseline vulnerability level to saturate. Even if there are repeated events or inspections in a short period of time, the vulnerability level still stays within a realistic range. In the original self-exciting point process model, it is possible for the self-excitation to escalate to the point where the probability of an event gets very close to one, which is generally unrealistic. In RPP’s, the saturation function prevents this from happening. Also if many inspections are done in a row, the vulnerability level does not drop to zero, and there are diminishing returns for the later ones because of the saturation function.

In New York City, and in other major urban centers, power grid reliability is a major source of concern as demand for electrical power is expected to soon exceed the amount we are able to deliver with our current infrastructure (DOE, 2008; Rhodes, 2013; NYBC, 2010). Many American electrical grids are massive and have been built gradually since the time of Thomas Edison in the 1880’s; for instance, in Manhattan alone, there are over 21,216 miles of underground cable, which is almost enough cable to wrap once around the earth. Manhattan’s power distribution system is the oldest in the world, and NYC’s power utility company, Con Edison, has cable databases that started in the 1880’s. Within the last decade, in order to handle increasing demands on NYC’s power grid and increasing threats to public safety, Con Edison has developed and deployed various proactive programs and policies (So, 2004). In Manhattan, there are approximately 53 thousand access points to the underground electrical grid, which are called electrical service structures, or manholes. Problems in the underground distribution network are manifested as problems within manholes, such as underground burnouts or serious events. A multi-year, ongoing collaboration
to predict these events in advance was started in 2007 (Rudin et al., 2010, 2012), where diverse historical data were used to predict manhole events over a long term horizon, as the data were not originally processed enough to predict events in the short term. Being able to predict manhole events accurately in the short term could immediately lead to reduce risks to public safety and increased reliability of electrical service. The data from this collaboration have sufficiently matured due to iterations of the knowledge discovery process and maturation of the Con Edison inspections program, and in this paper, we show that it is indeed possible to predict manhole events to some extent within the short term. A short version of this paper appeared in late-breaking developments track in AAAI-13 (Ertekin, Rudin and McCormick, 2013).

**Related work.** A general linear self-exciting process is identified by the conditional intensity function

\[
\lambda(t) = \mu + k_0 \sum_{t_k < t} g(t - t_k)
\]

where \(\mu\) is the baseline hazard rate. The Hawkes Process (Hawkes, 1971a,b) is a particularly tractable self-exciting point process with an exponential kernel, developed for earthquake modeling. Historically, a lot of effort has gone into modeling earthquakes, e.g., works of Ogata (1988, 1998), which consider Epidemic-Type-Aftershock-Sequences (ETAS). The formulation and parametrization of the ETAS model divides the effect of earthquakes into two components, namely background and aftershock events. Background events, corresponding to \(\mu\) in the expression above, occur independently according to a stationary Poisson process. An earthquake elevates the risk of aftershocks, where the magnitude and temporal properties of this elevated risk are determined by the kernel \(g\). Subsequent works incorporated the spatial distance of past seismic events for modeling the seismic activity at a point of interest (Musmeci and Vere-Jones, 1992; Ogata, 1998) (note that spatial distance elements do not appear to be relevant for manhole event prediction).

In recent applications of SEPP’s, Mohler et al. (2011) described a SEPP model that is based on the space-time ETAS model for modeling the rate of crimes. In particular, they drew a parallel between the spread of seismic activity and crime behavior, and proposed a model for residential burglaries that captures the spatial-temporal clustering patterns observed in crime data. Egesdal et al. (2010) studied a similar problem of modeling gang rivalries, where an event involving rival gangs can lead to retaliatory (self-excitatory) acts of violence. The same problem was addressed by
Louie, Masaki and Allenby (2010), which extends the model of Egesdal et al. (2010) to incorporate
the spatial influence of past events (similar to the space-time ETAS model of Ogata, 1998). Lewis
et al. (2010) analyzed temporal patterns of civilian death reports in Iraq between from 2003 until
2007 and defined a SEPP model with an exponential kernel similar to the one defined by Egesdal
et al. (2010). The paper mainly focuses on modeling the background violence rate $\mu$ since exogenous
factors such as political decisions, changes in troop levels, etc., can have a significant impact on the
rate of events.

In the financial world, SEPP models have been used both at micro and macro scales. Engle and
Russell (1998) proposed a SEPP model for analyzing transaction data on individual stocks and
how they are spread over a given time interval. At the macro scale, contagion of financial crises
was modeled as a mutually exciting jump process by Aït-Sahalia, Cacho-Diaz and Laeven (2010).
They modeled the dynamics of asset returns and how they fluctuate based on financial events in the
same region (self-excitation) as well as in other regions (mutual excitation). Note that usually when
handling self-exciting processes, the “multivariate setting” (as described by Embrechts, Liniger and
Lin, 2011) has the marks (where excitation occurs) as multivariate. In our case there is only one
dimension: vulnerability level. In our model, we have multiple manholes, each with their own
multidimensional vector of features. These features influence all of the marks for that manhole.

Recently, self-exciting point processes have been shown to be effective for genome analysis
(Reynaud-Bouret and Schbath, 2010), preventative maintenance for water pipes (Yan et al., 2013),
as well as analyzing and capturing the dynamics and neural spike trains (Krumin, Reutsky and
Shoham, 2010). These domains have distinct properties that require unique models (for example,
the self-excitation component of the model for DNA sequences replaces geographic distance with a
distance between basepairs).

In the context of social networks, Crane and Sornette (2008) studied endogenous and exogenous
factors that influence the dynamics of viewing activity for YouTube videos. They argue that an
individual’s viewing behavior is influenced by their communication with others about what to
watch next. Based on this assumption, their SEPP model defines the baseline rate $\mu$ to capture
spontaneous views that are not triggered by epidemic effects on the network. The self-excitation
component is controlled by the number of potential viewers influenced directly by person $i$ who
viewed a video at time $t_i$, and a “memory” kernel that captures the waiting time distribution for
the influenced individuals (the time between finding out about a video and viewing the video). A similar model with a memory kernel has been used to model conversation event sequences (Masuda et al., 2012).

RPPs differ from all of these related models in several ways: the form of the model is different, where the excitation is a proportion of the baseline rate, the saturation functions are unique to RPP’s, and the external inspections are not present in other models.

There has also been much recent work on Bayesian modeling for dependent point processes (see Guttorp and Thorarinsdottir, 2012, for an overview). Paralleling the development of frequentist literature, many Bayesian approaches are motivated by data on natural events. Peruggia and Santner (1996), for example, develop a Bayesian framework for the ETAS model. Non-parametric Bayesian approaches for modeling data form non-homogeneous point pattern data have also been developed (see Taddy and Kottas, 2012, for example). Blundell, Beck and Heller (2012) present a non-parametric Bayesian approach that uses Hawkes models for relational data.

We know of no previous works using SEPP’s for power grid maintenance. We remark that for manhole event prediction, it is not clear whether the self-exciting property that we observe is due to a weakening in the infrastructure caused by previous events (in analogy with aftershocks of earthquakes) or whether it is due to a weakening in the infrastructure caused by another (hidden) source.

Outline of paper. We first introduce the general form of the RPP model in Section 3. We then develop a conditional frequency estimator (CF estimator) for the RPP, and a parametric RPP model. The CF estimator approach allows us to trace out the shape of the self-exciting and self-regulating functions. Those functions can be parameterized afterwards if desired. The parametric model can be fit using a classical likelihood-based approach or using a Bayesian framework. The Bayesian formulation, which we implement using Approximate Bayesian Computation (ABC), allows us to share information across observably similar entities (manholes in our case). The likelihood and Bayesian methods use self-exciting, self-regulating, and saturating functions that are all parameterized, and the parameters can be chosen adaptively, based on covariates. Sharing information across entities is especially important for this application, since serious events are relatively rare, and because manholes can differ substantially in their amount of cable, age of cable, and in other
ways. All three statistical modeling approaches for RPP’s are discussed in Section 4: the CF estimator method is in Section 4.1, the likelihood-based RPP is in Section 4.2, and the Bayesian RPP model is in Section 4.3. For each method we fit the model to NYC data and performed simulation studies. Section 5 contains a prediction experiment, demonstrating the RPP’s ability to predict future events in NYC. Once the RPP model is fit to data from the past, it can be used for simulation. In particular, we can simulate various inspection policies for the Manhattan grid and examine the costs associated with each of them in order to choose the best inspection policy. Section 6 shows this type of simulation using the RPP, illustrating how it is able to help choose between different inspection policies, and thus assist with broader policy decisions for the NYC inspections program.

The paper’s Supplementary Material includes a description of the inspection policy used in Section 6 and simulation studies for validating the fitting techniques for the models in the paper. It also includes a description of a publicly available simulated dataset that we generated, based on statistical properties of the Manhattan dataset.

2. Description of Data. The data used for the project includes records from the Emergency Control Systems (ECS) trouble ticket system of Con Edison, which includes records of responses to past events (total 213,504 records for 53,525 manholes from 1995 until 2010). Part of the trouble ticket for a manhole fire is in Figure 1. Events can include serious problems such as manhole fires or explosions, or non-serious events such as wire burnouts. These tickets are heavily processed into a structured table, where each record indicates the time, manhole type (“service box” or “manhole,” FDNY/250 REPORTS F/O 45536 E.51 ST & BEEKMAN PL...MANHOLE FIRE MALDONADO REPORTS F/O 45536 E.51 ST FOUND SB-9960012 SMOKING HEAVY...ACTIVE...SOLID...ROUND...NO STRAY VOLTAGE...29-L... SNOW...FLUSH REQUESTED...ORDERED #100103.
12/22/09 08:10 MALDONADO REPORTS 3 2WAY-2WAY CRABS COPPERED CUT OUT & REPLACED SAME. ALSO STATES 5 WIRE CROSSING COMES UP DEAD WILL INVESTIGATE IN SB-9960013.
FLUSH # 100116 ORDERED FOR SAME 12/22/09 14:00 REMARKS BELOW WERE ADDED BY 62355 12/22/09 01:45 MASON REPORTS F/O 4553 E.51ST CLEARED ALL B/O-S IN SB9960013 ALSO FOUND A MAIN MISSING FROM THE WEST IN 12/22/09 14:08 REMARKS BELOW WERE ADDED BY 62355 SB9960011 F/O 1440 BEEKMAN ..JMC

Fig 1: Part of the ECS Remarks from a manhole fire ticket in 2009. The ticket implies that the manhole was actively smoking upon the worker’s arrival. The worker located a crab connector that had melted (“coppered”) and a cable that was not carrying current (“dead”). Addresses and manhole numbers were changed for the purpose of anonymity.
and we refer to both types as manholes colloquially), the unique identifier of the manhole, and details about the event. The trouble tickets are classified automatically as to whether they represent events (the kind we would like to predict and prevent) or not (in which case the ticket is irrelevant and removed). The processing of tickets is based on a study where Con Edison engineers manually labeled tickets, and is discussed further by Passonneau et al. (2011).

We have more-or-less complete event data from 1999 until the present, and incomplete event data between 1995 and 1999. A plot of the total number of events per year (using our definition of what constitutes an event) is provided in Figure 2 (left).

We also have manhole location and cable record information, which contains information about the underground electrical infrastructure. These two large tables are joined together to determine which cables enter into which manholes. The inferential join between the two tables required substantial processing in order to correctly match cables with manholes. Main cables are cables that connect two manholes, as opposed to service or streetlight cables which connect to buildings or streetlights. In our studies on long term prediction of power failures, we have found that the number of main phase cables in a manhole is a relatively useful indicator of whether a manhole is likely to have an event. Figure 2 (center) contains a histogram of the number of main phase cables in a manhole.

The electrical grid was built gradually over the last $\sim 130$ years, and as a result, manholes often contain cables with a range of different ages. Figure 2 (right) contains a histogram of the age of the oldest main cables in each manhole, as recorded in the database. Cable age is also used a feature for our RPP model. Cable ages range from less than a year old to over 100 years old; Con Edison
started keeping records back in the 1880’s during the time of Thomas Edison. We remark that it is not necessarily true that the oldest cables are the ones most in need of replacement. Many cables have been functioning for a century and are still functioning reliably.

We also have data from Con Edison’s new inspections program. Inspections can be scheduled in advance, according to a schedule determined by a state mandate. This mandate currently requires an inspection for each structure at least once every 5 years. Con Edison also performs “ad hoc” inspections. These occur when a worker is inside a manhole for another purpose (for instance to connect a new service cable), and chooses to fill in an inspection form. The inspections are broken down into 5 distinct types, depending on whether repairs are urgent (Level I), or whether the inspector suggests major infrastructure repairs (Level IV) that are placed on a waiting list to be completed. Sometimes when continued work is being performed on a single manhole, this manhole will have many inspections performed within a relatively small amount of time - hence our need for “diminishing returns” on the influence of an inspection that motivates the saturation function of the RPP model.

3. The Reactive Point Process Model. We begin with a simpler version of RPP’s where there is only one time series, corresponding to a single entity (manhole). Our data consist of a series of \( N_E \) events with event times \( t_1, t_2, \ldots, t_{N_E} \) and a series of given inspection times denoted by \( \bar{t}_1, \bar{t}_2, \ldots, \bar{t}_{N_I} \). The inspection times are assumed to be under the control of the experimenter. RPP’s model events as being generated from a non-homogeneous Poisson process with intensity \( \lambda(t) \) where

\[
\lambda(t) = \lambda_0 \left[ 1 + g_1 \left( \sum_{t_e < t} g_2(t - t_e) \right) - g_3 \left( \sum_{\bar{t}_i < t} g_4(t - \bar{t}_i) \right) + C_1 \mathbf{1}_{[N_E \geq 1]} \right]
\]

where \( t_e \) are event times and \( \bar{t}_i \) are inspection times. The vulnerability level permanently goes up by \( C_1 \) if there is at least one past event, where \( C_1 \) is a constant that can be fitted. The \( C_1 \mathbf{1}_{[N_E \geq 1]} \) term is present to deal with “zero inflation,” where the case of zero events needs to be handled separately than one or more past events. Functions \( g_2 \) and \( g_4 \) are the self-excitation and self-regulation functions, which have initially large amplitudes and decay over time. Self-exciting point processes have only \( g_2 \), and not the other functions, which are novel to RPP’s. Functions \( g_1 \) and \( g_3 \) are the saturation functions, which start out as the identity function and then flatten farther from
the origin. If the total sum of the excitation terms is large, \( g_1 \) will prevent the vulnerability level from increasing too much. Similarly, \( g_4 \) controls the total possible amount of self-regulation.

The form of the \( g_1, g_2, g_3, g_4 \) functions is the key distinction between the CF estimator for the RPP and the likelihood-based or Bayesian RPP formulations. Our CF estimator approach to RPPs traces out these functions explicitly using data. In the likelihood-based and Bayesian settings, we use the family of functions below for fitting power grid data, where \( a_1, b_1, a_3, b_3, \beta, \) and \( \gamma \) are parameters that can be either modeled or fitted. The form of these functions were derived from the CF estimator approach, as we will discuss.

\[
\begin{align*}
g_1(\omega) &= a_1 \times \left( 1 - \frac{1}{\log(2)} \log \left( 1 + e^{b_1 \omega} \right) \right), \quad g_2(t) = \frac{1}{1 + e^{\beta t}} \\
g_3(\omega) &= a_3 \times \left( 1 - \frac{1}{\log(2)} \log \left( 1 + e^{b_3 \omega} \right) \right), \quad g_4(t) = -\frac{1}{1 + e^{\gamma t}}.
\end{align*}
\]

The factors of \( \log(2) \) ensure that the vulnerability level is not negative.

In the case that there are multiple entities, there are \( P \) time series, each corresponding to a unique entity \( p \). For medical applications, each \( p \) is a patient, for the electrical grid reliability application, \( p \) is a manhole. Our data consist of events \( \{ t(p)_e \}_{p,e} \), inspections \( \{ \bar{t}(p)_i \}_{p,i} \), and additionally, we may have covariate information \( M_{p,j} \) about every entity \( p \), with covariates indexed by \( j \). Covariates for the medical application might include a patient’s gender, age at the initial time, race, etc. For the manhole events application, covariates include the number of main phase cables within the manhole (current carrying cables connecting two manholes), and the age of the oldest cables.

Within the Bayesian framework, we can naturally incorporate the covariates to model functions \( \lambda_p \) for each \( p \) adaptively. Consider \( \beta \) in the expression for \( g_2 \) above. For the likelihood-based and Bayesian approaches, \( \beta \) depends on individual-level covariates. In notation:

\[
\begin{align*}
g_2^{(p)}(t) &= \frac{1}{1 + e^{\beta^{(p)} t}}, \quad g_4^{(p)}(t) = -\frac{1}{1 + e^{\gamma^{(p)} t}}.
\end{align*}
\]

The \( \beta^{(p)} \)'s are assumed to be generated via a hierarchical model of the form

\[
\beta = \log \left( 1 + e^{-M \nu} \right)
\]

where \( \nu \sim N(0, \sigma_\nu^2) \) are the regression coefficients and \( M \) is the matrix of observed covariates. The \( \gamma^{(p)} \)'s are modeled hierarchically in the same manner, \( \gamma = \log \left( 1 + e^{-M \omega} \right), \) with \( \omega \sim N(0, \sigma_\omega^2) \). This permits slower or faster decay of the self-exciting and self-regulating components based on
the characteristics of the individual. For the electrical reliability application, we have noticed that manholes with more cables and older cables tend to have faster decay of the self-exciting terms for instance.

3.1. Demonstrating the Saturation Function in the RPP Model. In order to show the need for the saturating components of the RPP model, we show how the standard linear self-exciting process can produce unrealistic results under ordinary conditions.

First we show that the self-excitation term can cause the rate of events $\lambda(t)$ to increase without bound. To show this, we considered a baseline vulnerability of $\lambda_0 = 0.01$, setting $C_1 = 0.1$, used $g_2(t) = \frac{1}{1 + e^{0.005t}}$, and omitted the other components of the model (no inspections, no saturation $g_1$). The self-excitation eventually causes the rate of events to escalate unrealistically as shown in Figure 3 (upper left). The embedded subfigure is a zoomed-in version of the first 1500 time steps.

When we include the saturation function $g_1$, the excitation is controlled, and the probability of an event no longer increases to unreasonable levels. We used $g_1(\omega) = 1 - \frac{1}{\log 2} \log(1 + e^{-\omega})$, so that the vulnerability $\lambda(t)$ can reach to a maximum value of 0.021. The result is in Figure 3 (upper right).

Now we show the effect of the saturation function $g_3$ for the effect of repeated inspections. If no manhole events occur and the manhole is repeatedly inspected, then using the linear SEPP model, its vulnerability levels can become arbitrarily close to 0. This is not difficult to show, and we do this in Figure 3 (lower left). Here we used $\lambda_0 = 0.2$, $g_4(t) = \frac{-0.25}{1 + e^{0.002t}}$, and omitted $g_3$. We ran the same experiment but with saturation, specifically, with $g_3(\omega) = 1 - \frac{1}{\log 2} \log(1 + e^\omega)$. The results in Figure 3 (lower right) show that the saturation function never lets the vulnerability drop unrealistically far below the baseline level.

4. Fitting RPP statistical models. In this section we propose three statistical modeling approaches for RPPs: a conditional frequency (CF) estimator, a likelihood-based approach, and a Bayesian approach. The RPP intensity in Equation 1 features prominently in each of these models, providing the structure to capture self-excitation, self-regulation, and saturation. We compare and discuss the fit from the three approaches in Section 4.4.

The long term vulnerability levels of the manholes vary, and for the purpose of the experiments conducted here, we worked with a large set of manholes that are not at either of the extremes of
being the most or least vulnerable.

The terminology we use below is specific to the energy grid reliability application, but the fitting methods are general and can be used in any of the domains discussed earlier.

We performed a set of “sanity check” simulation experiments, where the goal is to recover parameters of simulated data for which there is ground truth. Details of these additional experiments are in the Supplementary Material.

4.1. Conditional Frequency (CF) Estimator for the RPP. For this method, we locate event/inspection “trails” in the time-series data for each manhole. We define a trail as a sequence of time steps of fixed granularity (days, weeks, etc.) for a specific manhole where a predefined pattern of inspection/event records exists. The trail ends when we observe an inspection or event record. We estimate the model parameters in a case-controlled way, using only the applicable trails, in the following order:

1. Estimate \( \lambda_0 \). The baseline hazard rate \( \lambda_0 \) refers to the likelihood of observing an event under steady-state conditions, for a manhole that has no previous event record in the database. That is, at a given time \( t \) where there has not been a recorded inspection or event, \( \lambda_0 \) is the probability of an event at that time. Computation of \( \lambda_0 \) requires trails that start from the
earliest recorded time until the observation of the very first event record or inspection record for each manhole. If the manhole never had an event or inspection, the trail is the full time course. Let \( S_t \) denote the number of manholes that have survived without an event up to and including time \( t - 1 \), and did not have an inspection through time \( t \). Let \( E_t \) denote the number of these manholes that had an event at time \( t \). We first compute an estimate of the baseline event rate \( \lambda_0 \) for each \( t \) separately as \( E_t / S_t \). The later times represent trails that are long, and there are much fewer such trails. Thus we have far less data at later times than at earlier times, and the estimates at later times can be unreliable. For our estimate of \( \lambda_0 \) we take an average of these estimates, weighted by the number of observations for each estimate, \( S_t \). It reduces to a pooled average as follows:

\[
\lambda_0 \approx \frac{\sum_{t=1}^{T_{\text{max}}} (E_t / S_t) \cdot S_t}{\sum_{t=1}^{T_{\text{max}}} S_t} = \frac{\sum_{t=1}^{T_{\text{max}}} E_t}{\sum_{t=1}^{T_{\text{max}}} S_t}
\]

where \( T_{\text{max}} \) is the length of the longest trail that we are willing to consider.

To validate the CF estimator method result, we were able to reproduce the value of \( \lambda_0 \) given in a simulation very closely. In simulation studies described in more detail in the Supplementary Material, we estimated \( \lambda_0 = 0.0097 \), when the true value was \( \lambda_0 = 0.01 \).

As an alternative method for estimating \( \lambda_0 \) in a personalized way, one can use a long-term probabilistic model (such as the one we had developed for the project previously in [Rudin et al. (2010)]) that incorporates features for each manhole.

2. **Estimate** \( C_1 \). After finding the baseline hazard rate, we estimate \( C_1 \) from trails in the data where the effects of \( g_2 \) and \( g_4 \) are hypothesized to have approximately vanished, and where the manhole had a prior event. In other words, the trails that we consider have the following properties: \( i \) there was a past event in the manhole, \( ii \) the start of the trail is a time at which the effect of a previous event/inspection is fully decayed, and \( iii \) the end of the trail is the observation of an event or inspection record in that manhole if there was one. There are some manholes for which we have multiple trails by these criteria, and we use all of them. Defining \( E_{\Delta t} \) and \( \bar{E}_{\Delta t} \) as the total number of trails in the data with and without an event after \( \Delta t \) of the start of the trail (and no inspections during the trail), we again use a pooled
average to get:

\[
\lambda_0 (1 + C_1) = \frac{\sum_{\Delta t=1}^T E_{\Delta t}}{\sum_{\Delta t=1}^T (E_{\Delta t} + \bar{E}_{\Delta t})}
\]

\[
C_1 = \left[ \frac{1}{\lambda_0} \sum_{\Delta t=1}^T E_{\Delta t} \left( E_{\Delta t} + \bar{E}_{\Delta t} \right) \right] - 1.
\]

3. **Estimate** \(g_2\): To obtain the CF estimates, we consider events that are observed at steady-state, meaning that the effect of previous events/inspections, if any, should have fully decayed. In other words, we choose trails starting at events that occurred at baseline vulnerability levels. Furthermore, to keep the computation consistent, we consider only trails where there was already a previous serious event in that manhole (so that we can consistently use \(C_1\) in our computations). The trail ends at the immediate next event/inspection record, if any.

Redefining \(E_{\Delta t}\) and \(\bar{E}_{\Delta t}\) as the total number of trails in the database with and without an event after \(\Delta t\) of the start of the trail (and no inspections within that time), we compute an estimated failure rate \(\frac{E_{\Delta t}}{E_{\Delta t} + \bar{E}_{\Delta t}}\) for each \(\Delta t\) interval and a CF estimate for \(g_2\) at time \(\Delta t\):

\[
g_{\text{CFestimator}}^2(\Delta t) = \frac{1}{\lambda_0} \frac{E_{\Delta t}}{E_{\Delta t} + \bar{E}_{\Delta t}} - C_1 - 1.
\]

Figure 4(a) shows \(g_{\text{CFestimator}}^2\) traced for the Manhattan data (dots in the figure), and the parameterized \(g_2\) function we fitted afterwards. The fitted curve uses the form given in (2).

We also evaluated the shape of \(g_2\) curves estimated using this method using data simulated with known \(g_2\) curves. The results, given in detail in the Supplementary Materials, indicate that this method accurately recovers the true \(g_2\) function.
4. **Estimate** \( g_4 \): The computation of \( g_4^{\text{CFestimator}} \) values across various trail lengths follows the same procedure as for \( g_2^{\text{CFestimator}} \), with the distinction that the trails now start with an inspection record. That is, we use only trails where the manhole was inspected at the baseline vulnerability level and there was a history of at least one serious event in that manhole. Figure 1(d) in the Supplementary Material shows that we are able to recover \( g_4 \) for such trails in the simulated data. We cannot well-estimate \( g_4 \) for the Manhattan dataset since the inspection program is relatively new, and since the effect of an inspection and repair can take much longer to wear off than the effect of an event (in fact, longer than the span of the data that we have presently). Engineers at Con Edison assisted us to estimate \( g_4 \) using domain knowledge, with a band of uncertainty that we take into account in Section 6.

5. **Estimate** \( g_1 \): The computation of \( g_1 \) involves focusing on the trails with event cascades; that is, we are interested in the trails that start with a serious event that is under the influence of a prior serious event. These are events for which a previous event’s effect has not fully decayed. These cascades of multiple events close together in time “activate” the \( g_1 \) component of our model.

Accordingly, we find trails that start at the time of a serious event that closely follows at least one prior serious event. These trails end at the next event (when a new trail begins) or at the next inspection. We exclude trails under the influence of a past inspection. We first discretize possible values for \( \sum g_2 \) into bins. At each timestep of each trail, we will determine which is the corresponding bin, then estimate a probability of event for each bin. We denote a trail as \( Tr \), its start time as \( \text{time}(Tr) \), and \( E_{\Delta t}^{Tr} = \{e_1, e_2, \ldots, e_k\} \) as the set of previous recent events whose effects have not completely decayed at time \( \text{time}(Tr) + \Delta t \). This includes the event at time \( \text{time}(Tr) \). At each timestep \( \Delta t \) after the start of trail \( Tr \), we compute

\[
V^{Tr} = \sum_{e \in E_{\Delta t}^{Tr}} g_2(\text{time}(Tr) + \Delta t - \text{time}(e)).
\]

We then determine which bin \( V^{Tr} \) corresponds to, which is denoted by \( b \). Depending on whether or not we observed a serious event at time \( T_i + \Delta t \), we add 1 to either \( S_{\text{Event},b} \) or \( S_{\text{NoEvent},b} \). Once \( S_{\text{Event},b} \) and \( S_{\text{NoEvent},b} \) have been computed using all timesteps for all trails, the value of \( g_1 \) in bin \( b \) equals \( S_{\text{Event},b} / (S_{\text{Event},b} + S_{\text{NoEvent},b}) \).

For the Manhattan dataset, our estimate of \( g_1 \) was afterwards parameterized using the func-
tional form in (2), as shown in Figure 4(b).

6. **Estimate** $g_3$: The computation of $g_3$ follows the same steps defined for $g_1$, with the distinction that we now consider trails that start with an inspection where that inspection is still under the influence of an earlier inspection, and no previous events.

This procedure yields the following estimates for the Con Edison data:

- $\lambda_0 = 2.4225\times10^{-4}$ (baseline rate of approximately 1 in 4070)
- $C_1 = 0.0512$ (baseline rate changes to $\lambda_0(1 + C_1)$ which is approximately 1 in 3930)
- $g_2(t) = \frac{1.162}{1 + e^{0.039t}}$
- $g_1(t) = 16.98 \times \left(1 - \log(1 + e^{-0.15t}) \times \frac{1}{\log 2}\right)$.

For the likelihood-based and Bayesian RPP models discussed below, we will use the parametric forms of $g_1$, $g_2$, $g_3$, and $g_4$ that we discovered using the CF estimator method, though we will specialize the parameters to each manhole adaptively.

4.2. **Likelihood-based RPP.** We start by considering a simple version of (3) where there are no inspections, and $\beta$ does not depend on covariates. The manholes are indexed by $p$. Here, the log likelihood is then expressed using the standard likelihood for a non-homogeneous Poisson process over the time interval $[0, T_{\text{max}}]$:

$$\log L \left( \left\{ t_{(p)}^{(1)}, \ldots, t_{(p)}^{(N_E^{(p)})} \right\}_p ; \beta \right) = \sum_{p=1}^{P} \left[ \sum_{e=1}^{N_E^{(p)}} \log(\lambda(t_e^{(p)})) - \int_0^{T_{\text{max}}} \lambda(u) du \right],$$

where $\lambda$ is given by (1) and (2). This formula extends directly to the feature-based case where covariates are present. In terms of $\upsilon$, the likelihood is:

$$\log L \left( \left\{ t_{(1)}^{(p)}, \ldots, t_{(N_E^{(p)})}^{(p)} \right\}_p ; \upsilon, a_1, M \right) = \sum_{p=1}^{P} \left[ \sum_{e=1}^{N_E^{(p)}} \log(\lambda_p(t_e^{(p)})) - \int_0^{T_{\text{max}}} \lambda_p(u) du \right],$$

where

$$\lambda_p(t) = \lambda_0 \left[1 + g_1 \left( \sum_{\upsilon_e^{(p)} < t} g_2(t - t_e^{(p)}) \right) + C_1 1_{[N_E^{(p)} \geq 1]} \right]$$

and

$$g_1(t) = a_1 \times \left(1 - \log(1 + e^{-a_1t}) \times \frac{1}{\log 2}\right), \quad g_2(t - t_e^{(p)}) = \frac{1}{1 + e^{(t - t_e^{(p)}) \log(1 + \exp(\sum_j M_{pj}v_j))}}.$$
We sampled the coefficients $\nu$ from a normal distribution, displaying in Figure 5 the region of maximum likelihood (shown in dark red), where the data have the highest likelihood with respect to the sampled values of $\nu$. The covariates are the number of main phase cables in the manhole (number of current carrying cables between two manholes), the total number of cable sets (total number of bundles of cables) including main, service, and streetlight cables, and the age of the oldest cable set within the manhole. All covariates were normalized to be between -0.5 and 0.5.

Fig 5: Log Likelihood for Manhattan Dataset. Each axis represents possible values of a coefficient. The “Age” axis is for the age of the oldest cable set within the manhole, “Main PH” axis for the number of main phase cables in the manhole and “Total Num. Sets” axis is for the total number of cable sets including main, service, and streetlight cables.

4.3. Bayesian RPP. Developing a Bayesian framework facilitates sharing of information between observably similar manholes, thus making more efficient use of available covariate information. The RPP model encodes much of our prior information into the shape of the rate function given in Equation 1. In developing a Bayesian framework, therefore, we opt for a simple, parsimonious model that imposes mild regularization and information sharing without adding substantial additional information. Specifically, beginning with the non-homogeneous Poisson Process likelihood in Equation 4, we use diffuse Gaussian priors on the log scale for each regression coefficient.

We fit the model using Approximate Bayesian Computation (Diggle and Gratton, 1984). The principle of Approximate Bayesian Computation (ABC) is to randomly choose proposed parameter values, use those values to generate data, and then compare the generated data to the observed data. If the difference is sufficiently small, then we accept the proposed parameters as draws from the approximate posterior. To do ABC, we need two things: (i) to be able to simulate from the model and (ii) a summary statistic. To compare the generated and observed data, the summary
statistic from the observed data, \( S(\{t_1^{(p)}, ..., t_{N_E}^{(p)}\}_p) \), is compared to that of the data simulated from the proposed parameter values, \( S(\{t_1^{(p),\text{sim}}, ..., t_{N_E}^{(p),\text{sim}}\}_p) \). If the values are similar, it indicates that the proposed parameter values may yield a useful model for the data.

A critical difference between updating a parameter value in an ABC iteration versus, for example, a Metropolis-Hastings step is that ABC requires simulating from the likelihood whereas Metropolis-Hastings requires evaluating the likelihood. In our context, we are able to both evaluate and simulate from the likelihood with approximately the same computational complexity. ABC has some advantages, namely that we have meaningful summary statistics, discussed below. Further, in our case it is not particularly computationally challenging, as we already extensively simulate from the model as a means of evaluating hypothetical inspection policies. We evaluated the adequacy of this method extensively in simulation studies presented in the Supplementary Material.

A key conceptual aspect of ABC is that one can choose the summary statistic to best match the problem. The sufficient statistic for the RPP is the vector of event times, and thus gives no data reduction - so we choose other statistics. One important insight in constructing our summary statistic is that changing the parameters in the RPP model alters the distribution of times between events. The histogram of time differences for a homogenous Poisson Process, for example, has an exponential decay. The self-exciting process, on the other hand, has a distribution resembling a lognormal because of the positive association between intensities after an event occurs. Altering the parameters of the RPP model changes the intensity of self-excitation and self-regulation, thus altering the distribution of times between events. We construct our first statistic, therefore, by examining the KL divergence between the distribution of times between events in the data and the distribution between event times in the simulated data. We do this for each of our proposed parameters. Examining the distribution of times between events, though not the true sufficient statistic, captures a concise and low dimensional summary of a key feature of the process. This statistic does not, however, capture the overall prevalence of events in the process. Since we focus only on the distribution of times between events, various processes with different overall intensity could produce distributions with similar KL divergence to the data distribution. We therefore introduce a second statistic that counts the total number of events. We contend that together these statistics represent both the spacing and the overall scale (or frequency) of events. Thus, the two summary measures we use are:
1. DNE: The difference in the number of events in the simulated and observed data.

2. KL: The Kullback-Leibler divergence between two histograms, one from the observed data, and one from the real data. These are histograms of time differences between events.

For the NYC data, we visualized three-dimensional parameter values, both for DNE (in Figure 6) and KL (in Figure 7) metrics. In both figures, smaller values (dark blue) are better. As seen, the regions where KL and DNE are optimized are very similar.

Denoting the probability distribution of the actual data as $P$ and the probability distribution of the simulated data as $Q_{\nu}$, KL Divergence is computed as

$$KL(P||Q_{\nu}) = \sum_{\text{bin}} \ln \left( \frac{P(\text{bin})}{Q_{\nu}(\text{bin})} \right) P(\text{bin}).$$

As mentioned previously in this section, the prior information in our model enters primarily through the likelihood, which means that, though still critical for regularization and information sharing, the Bayesian portion of our model is relatively parsimonious. We require a distribution $\pi$ over parameter values. If covariates are not used, $\pi$ is a distribution over $\beta$ (and $\gamma$ if inspections are present). If covariates are used, $\pi$ is a distribution over $\omega$ and $\nu$. One option for $\pi$ is a uniform distribution across a grid of reasonable values. Another option, which was used in our experiments, is to simulate from diffuse Gaussian/weakly informative priors on the log scale (for instance, draw $\log(\nu_j) \sim N(0, 5)$). We assumed that $C_1$ and $a_1$ can be treated as tuning constants to be estimated using the CF estimator method, though it is possible to define priors on these quantities as well if desired.

There is an increasingly large literature in both the theory and implementation of ABC (see for
example Fearnhead and Prangle, 2012; Beaumont et al., 2009; Drovandi, Pettitt and Faddy, 2011) that could be used to produce estimates of the full posterior. In the Supplementary Material, we present an importance sampling algorithm as one possible approach. In our work, however, the goal is to estimate the posterior mode, which we then use for prediction.

We estimate the posterior mode by using a manifold approximation to the region of high posterior density. We begin by generating a set of proposed parameter values using the prior distributions. Consistent with ABC, we simulate data from each set of candidate values and compare the simulated data to our observed data using the KL and DNE statistics described above. From here, we could, for example, define a kernel and accept draws with a given probability as in importance sampling. Instead, since our goal is estimating the posterior mode, we concentrate only on the points that have the highest similarity to the observed data. Specifically, we fit a manifold to the points that were in the bottom 10% of KL values and the bottom 10% of DNE values. We then choose a particularly desirable point on the manifold as our set of parameter values to use for the policy simulation. In our case we chose the points on the manifold closest to the origin, which encourages additional regularization. To verify the procedure, we used simulated ground truth data with known $\beta$ and $\gamma$ values, and attempted to recover these values with the ABC method, for both the DNE and KL metrics. We performed extensive simulation studies to evaluate this method and full results are given in the Supplementary Materials.

4.4. Choosing Parameter Values for the Policy Decision. For the policy simulation in Section 6 we wish to choose one set of parameter values to inform our decision. We use the parameter values

Fig 7: KL for Manhattan Dataset. Each axis corresponds to the coefficient for one of the covariates. The magnitude of KL is indicated by the color.
from our Bayesian RPP approach which incorporates both the key structural features of the RPP through the likelihood and regularization through the prior structure.

In order to choose a single best value of the parameters, we fitted a polynomial manifold to the intersection of the bottom 10% of KL values and the bottom 10% of DNE values. Defining $v_1$, $v_2$ and $v_3$ as the coefficients for “number of main phase cables”, “age of oldest main cable set” and “total number of sets” features, the formula for the manifold is:

$$v_3 = -9.6 - 0.98v_1 - 0.13v_2 - 1.1 \times 10^{-3}(v_1)^2 - 3.6 \times 10^{-3}v_1v_2 + 4.67 \times 10^{-2}(v_2)^2,$$

which is determined by a least squares fit to the data. The fitted manifold is shown in Figure 8 along with the data.

We then optimized for the point on the manifold closest to the origin. This implicitly adds regularization, as it chooses the parameter values closest to the origin. This point is $v_1 = -4.6554$, $v_2 = -0.5716$, and $v_3 = -4.8028$. Note that cable age (corresponding to the second coefficient) is not the most important feature defining the manifold. As previous studies have shown (Rudin et al., 2010), even though there are very old cables in the city, the age of cables within a manhole is not alone the best predictor of vulnerability. Now we also know that it is not the best predictor of the rate of decay of vulnerability back to baseline levels. This supports Con Edison’s goal to prioritize the most vulnerable components of the power grid, rather than simply replacing the oldest components.

5. Predicting events on the NYC power grid. Our first experiment aims to evaluate whether the CF Estimator for the RPP or feature-based strategies introduced above is better in
terms of identifying the most vulnerable manholes. To do this, we selected 5,000 manholes (rank 1,001-6,000 from the project’s current long-term prediction model). These manholes have similar vulnerability levels, which allows us to isolate the self-exciting effect without modeling the baseline level. Using both the feature-based $\beta$ (ABC, with KL metric) and constant $\beta$ (CF estimator method) strategies, the models were trained on data through 2009, and then we estimated the vulnerabilities of the manholes on December 31$^{st}$, 2009. These vulnerabilities were used as the initial vulnerabilities for an evaluation on the 2010 event data. 2010 is a relevant year because the first inspection cycle ended in 2009. All manholes had been inspected at least once, and many were inspected towards the end of 2009, which stabilizes the inspection effects. For each of the 53K manholes and at each of the 365 days of 2010, when we observed a serious event in a manhole $p$, we evaluated the rank of that manhole with respect to both the feature-based and non-feature-based models, where rank represents the number of manholes that were given higher vulnerabilities than manhole $p$. As our goal is to compare the relative rankings provided by the two strategies, we consider only events where the vulnerabilities assigned by both strategies are different than the baseline vulnerability. Figure 9 displays the ranks of the manholes on the day of their serious event. A smaller rank indicates being higher up the list, thus lower is better. Overall, we find that the feature-based $\beta$ strategy performs better than the non-feature-based strategy over all of the rank comparisons in 2010 (pvalue .09, sign test).

In the second experiment, we compared the feature-based $\beta$ strategy to the Cox Proportional-
Hazard Model, which is commonly used in survival analysis to assess the probability of failure in mechanical systems. We employed this model to assess the likelihood of a manhole having a serious event on a particular day. For each manhole, we used the same three static covariates as in the feature-based $\beta$ model, and developed four time-dependent features. The time-varying features for day $t$ are 1) the number of times the manhole was a trouble hole (source of the problem) for a serious event until $t$, 2) the number of times the manhole was a trouble hole for a serious event in the last year, 3) the number of times the manhole was a trouble hole for a precursor event (less serious event) until $t$, and 4) the number of times the manhole was a trouble hole for a precursor event in the last year. The feature-based $\beta$ model currently does not differentiate serious and precursor events, though it is a direct extension to do this if desired. The model was trained using the coxph function in the R survival package using data prior to 2009, and then predictions were made on the test set of 5,000 manholes in the 2010 dataset. These predictions were transformed into ranked lists of manholes for each day. We then compared the ranks achieved by the Cox model with the ranks of manholes at the time of events. The difference of aggregate ranks was in favor of the feature-based $\beta$ approach (p-value $7e-06$, sign test), indicating that the feature-based $\beta$ strategy provides a substantial advantage in its ability to prioritize vulnerable manholes.

6. Making Broader Policy Decisions Using RPP’s. Because the RPP model is a generative model, it can be used to simulate the future, and thus assist with broader policy decisions regarding how often inspections should be performed. This can be used to justify allocation of spending. Con Edison’s existing inspection policy is a combination of targeted periodic inspections and ad-hoc inspections. The targeted inspections are planned in advance, whereas the ad hoc inspections are unscheduled. An ad hoc inspection could be performed while a utility worker is in the process of, for instance, installing new service cable to a building or repairing an outage. Either source of inspection can result in an urgent repair (Type I), an important but not urgent repair (Type II), a suggested structural repair (Types III and IV), or no repair, or any combination of repairs. Urgent repairs need to be completed before the inspector leaves the manhole, whereas Type IV repairs are placed on a waiting list. According to the current inspections policy, each manhole undergoes a targeted inspection every 5 years. The choice of inspection policy to simulate can be determined very flexibly, and any inspection policy and hypothesized effect of that policy can be
examined through simulation.

As a demonstration, we conducted a simulation over a 20 year future time horizon that permits a cost-benefit analysis of the inspection program, when targeted inspections are performed at a given frequency. To do this simulation we require the following:

- A characterization of manhole vulnerability. For Manhattan, this is learned from the past using the ABC RPP feature-based $\beta$ training strategy for $g_1$ and $g_2$ discussed above. Functions $g_3$ and $g_4$ for the inspection program cannot yet be learned due to the newness of the inspection program and are discussed below.
- An inspection policy. The policy can include targeted, ad hoc, or history-based inspections. We chose to evaluate “bright line” inspection policies, where each manhole is inspected once in each $Y$ year period, where $Y$ is varied (discussed below). We also included an ad hoc inspection policy that visits 3 manholes per day on average.

**Effect of Inspections:** The effect of inspections on the overall vulnerability of manholes were designed in consultation with domain experts. The choices are somewhat conservative, so as to give a lower bound for costs. The effect of an urgent repair (Type I) is different from the effect of less urgent repairs (Types II, III, and IV). For all inspection types, after 1 year beyond the time of the inspection, the effect of the inspection decays to, on average, 85% of its initial effect, in agreement with a short-term empirical study on inspections. (There is some uncertainty in this initial effect, and the initial drop in vulnerability is chosen from a normal distribution so that after one year, the effect decays to a mean of 85%.) For Type I inspections, the effect of the inspection decays to baseline levels after approximately 3000 days, and for Type II, III, and IV, which are more extensive repairs, the effect fully decays after 7000 days. In particular, we use the following $g_4$ functions:

\[
g_{\text{Type I}}(t) = -83.7989 \times (r \times 5 \times 10^{-4} + 3.5 \times 10^{-3}) \times \frac{1}{1 + e^{0.0018t}}
\]

\[
g_{\text{Type II,III,IV}}(t) = -49.014 \times (r \times 5 \times 10^{-4} + 7 \times 10^{-3}) \times \frac{1}{1 + e^{0.00068t}}
\]

where $r$ is randomly sampled from a standard normal distribution. For all inspection types, we used the following $g_3$ saturation function:

\[
g_3(t) = 0.4 \times \left(1 - \log(1 + e^{-3.75t}) \times \frac{1}{\log 2}\right)
\]

which ensures that subsequent inspections do not lower the vulnerability to more than 60% of the
baseline vulnerability. Sampled \( g_4 \) functions for Type I and Type II, II, IV, along with \( g_3 \) are shown in Figure 10.

One targeted inspection per manhole was distributed randomly across \( Y \) years for the bright line \( Y \)-year inspection policies, and \( 3 \times 365 = 1095 \) ad-hoc inspections for each year were uniformly distributed, which corresponds to 3 ad-hoc inspections per day for the whole power grid on average. During the simulation, when we arrived at a time step with an inspection, the inspection outcome was Type I with 25% probability, or one of Types II, III, or IV, with 25% probability. In the rest of cases (50% probability), the inspection was clean, and the manhole’s vulnerability was not affected by the inspection. If the inspection resulted in a repair, we sampled \( r \) randomly and randomly chose the inspection outcome (Type I or Types II, III, IV). This percentage breakdown was observed approximately for a recent year of inspections in NYC.

To initialize manhole vulnerabilities for a bright line policy of \( Y \) years, we simulated the previous \( Y \)-year inspection cycle, and started the simulation with the vulnerabilities obtained at the end of this full cycle.

**Simulation Results:** We simulated events and inspections for 53.5K manholes for bright line policies ranging from \( Y = 1 \) year to \( Y = 20 \) years. A longer inspection cycle corresponds to fewer daily inspections, which translates into an increase in overall vulnerabilities and an increase in the number of events. This is quantified in Figure 11, which shows the projected number of inspections and events for each \( Y \) year bright line policy. If we change from a 6 year bright line inspection policy to a 4 year policy, we estimate a reduction of approximately 100 events per year. The relative costs of inspections and events can thus be considered in order to justify a particular choice of \( Y \) for the bright line policy.
Fig 11: Number of events and inspections based on bright line policy. Number of years $Y$ for the bright line policy is on the horizontal axis in both figures. The left figure shows the number of inspections, the right figure shows the number of events.

7. Conclusion. Keeping our electrical infrastructure safe and reliable is of critical concern, as power outages affect almost all aspects of our society including hospitals, financial centers, data centers, transportation, and supermarkets. If we are able to combine historical data with the best available statistical tools, it will be possible to impact our ability to maintain an ever aging and growing power grid. In this work, we presented a methodology for modeling power grid failures that is based on natural assumptions: (i) that power failures have a self-exciting property, which was hypothesized by Con Edison engineers, (ii) that the power company’s actions are able to regulate vulnerability levels, (iii) that the effects on the vulnerability level of past events or repairs can saturate, and (iv) that vulnerability estimates should be similar between similar entities. We have been able to show directly (using the CF estimator for the RPP) that the self-exciting and saturation assumptions hold. We demonstrated through experiments on past power grid data from NYC, and through simulations, that the RPP model is able to capture the relevant dynamics well enough to predict power failures better than the current approaches in use.

The modeling assumptions that underlie RPP’s can be directly ported to other problems. RPP’s are a natural fit for problems in healthcare, where medical conditions cause self-excitation, and treatments provide regulation. Through the Bayesian framework we introduced, RPP’s extend to a broad range of problems where predictive power can be pooled among multiple related entities, such as manholes or medical patients.

The results presented in this work show for the first time that manhole events can be predicted in the short term, which was previously thought not to be possible. Knowing how one might do this permits us to take preventive action to keep vulnerability levels low, and can help make
broader policy decisions for power grid maintenance through simulation of many uncertain futures, simulated over any desired policy.

Acknowledgements. Funding for this work provided by Con Edison and National Science Foundation CAREER grant IIS-1053407 to C. Rudin.

References.


S. Ertekin and C. Rudin
MIT Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139
E-mail: seyda@mit.edu
rudin@mit.edu

T. McCormick
Department of Statistics
University of Washington
Seattle, WA 98195
E-mail: tylermc@u.washington.edu