Estimating the Historical and Future Probabilities of Large Terrorist Events [CW13]

Paper Presentation as part of the course Statistics Comprehensive

Purushottam Saha (BS2119)

B.Stat. 3rd Year
Indian Statistical Institute, Kolkata

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Introduction

Quantities with right-skewed distributions are ubiquitous in complex social systems, including political conflict, economics and social networks, and these systems sometimes produce extremely large events.
For example. The 9/11 terrorist attack were the largest such events in modern history, killing nearly 3000 people [MIP08]. Given their severity, should these attacks be considered statistically unlikely or even outliers? What is the likelihood of another September 11^{th}-sized or larger terrorist event, worldwide, over the next decade?

Why this is important?

Accurate answers to such questions would shed new light both on the global trends and risks of terrorism and on the global social and political processes that generate these rare events, which depends in part on determining whether the same processes generate both rare, large events and smaller, more common events. Insights would also provide objective guidance for our long-term expectations in planning, response and insurance efforts, and for estimating the likelihood of even larger events, including mass casualty Chemical, Biological, Radioactive or Nuclear (CBRN) events.

Major Issues

The rarity of events like 9/11 poses two technical problems:

1. We typically lack quantitative mechanism-based models with demonstrated predictive power at the global scale (which is particularly problematic for CBRN events)
2. The global historical record contains few large events from which to estimate mechanism-agnostic statistical models of large events alone. That is, the rarity of big events implies large fluctuations in the distribution's upper tail, precisely where we wish to have the most accuracy. These fluctuations can lead to poor out-of-sample predictive power in conflict and can complicate both selecting the correct model of the tail's structure and accurately estimating its parameters. Misspecification can lead to severe underestimates of the true probability of large events, for example, in classical financial risk models.

General Approaches

- Little research on terrorism has focused on directly modelling the number of deaths ("severity") in individual terrorist events. When deaths are considered, they are typically aggregated and used as a covariate to understand other aspects of terrorism, for example, trends over time.
- Such efforts have used time series analysis, qualitative models or human expertise of specific scenarios, actors, targets or attacks, or quantitative models based on factor analysis (i.e. in terms of latent variables) etc.
- Most of this work focuses on modelling central tendencies, treats large events like 9/11 as outliers and says little about their quantitative probability.

Our Approach

- In this paper, we present a generic statistical algorithm for making such estimates, which combines semi-parametric models of tail behaviour and a nonparametric bootstrap.
- Our approach combines maximum-likelihood methods, multiple models of the distribution's tail and computational techniques to account for both parameter and model uncertainty. It provides a quantitative estimate of the probability, with uncertainty, of a large event.
- We then use this procedure to make a data-driven statistical forecast of at least one similar event over the next decade.
- The algorithm also naturally generalizes to include certain event covariates, which can shed additional light on the probability of large events of different types.

Overview

- Using this algorithm to analyze a database of 13,274 deadly terrorist events worldwide from 1968–2007, we estimate the global historical probability of at least one 9/11-sized or larger terrorist event over this period to be roughly 11–35%.
- Furthermore, we find the nontrivial magnitude of this historical probability to be highly robust, a direct consequence of the highly right-skewed or "heavy-tailed" structure of event sizes.
- Thus, an event the size or severity of the September 11th terrorist attacks, compared to the global historical record, should not be considered a statistical outlier or even statistically unlikely.
- Using three potential scenarios for the evolution of global terrorism over the next decade, we then estimate the worldwide future probability of a similarly large event as being not significantly different from the historical level.
- We close by discussing the implications for forecasting large terrorist events in particular and for complex social systems in general.

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Methodology

- The problem of estimating the probability of some observed large event is a kind of tail-fitting problem, in which we estimate parameters for a distributional model using only the several largest observations.
- Here, we aim specifically to deal with several sources of uncertainty in this task: uncertainty in the location of the tail, uncertainty in the tail's true structure, and uncertainty in the model parameters.
- Our approach is based on three key insights:
  1. Tail Model Estimating
  2. Model Comparison and Model Averaging
  3. Tests of the Method's Accuracy

Tail Model Estimating

- First, because we are interested only in rare large events, we need only model the structure of the distribution's right or upper tail, which governs their frequency.
- This replaces the difficult problem of modelling both the distribution's body and tail with the less difficult problem of identifying a value x_{min} above which a model of the tail alone fits well.
- So the problem becomes choosing some x_{min} and a tail model $\mathbb{P}(x|\theta, x_{min})$ defined on $x \in [x_{min}, \inf)$.
- We will revisit the problem of choosing x_{min} later.

Tail Model Estimating (Contd.)

- Here, our goal is to estimate the probability that we would observe at least ℓ "catastrophic" events of size x or greater in an empirical sample.
- In principle, any size x and any value may be chosen, but, in practice, we typically choose x as the largest (and thus rarest) event in the empirical data and set $\ell = 1$.
- To ensure that our estimate is meaningful from a historical perspective, we remove the catastrophic event(s) from the empirical sample before applying the algorithm.

Tail Model Estimating (Contd.)

- Let $Pr(x|\theta, x_{min})$ denote a particular tail model with parameters θ, let $\{x_i\}$ denote the n empirical event sizes (leaving the catastrophic events), and let Y $= \{y_j\}$ be a bootstrap of these data (n samples drawn from $\{x_i\}$ with replacement).
- To begin, we assume a fixed x_{min}, the smallest value for which the tail model holds, and later describe the generalization to variable x_{min}.
- The fraction of empirical events with values in the tail region is $p_{tail} = \#\{x_i \geq x_{min}\}/n$, and in each bootstrap, the number is a binomial random variable with probability p_{tail}, i.e.

$$n_{tail} \sim Binomial(n, p_{tail})$$

- The maximum likelihood estimate $\hat{\theta}$ is a deterministic function of the portion of Y above x_{min}, which we denote $\theta(Y, x_{min})$.

Tail Model Estimating (Contd.)

- The probability under the fitted model that not one of $n'_{tail} = 1 + n_{tail}$ events is at least as big as x is

$$F(x|\theta(Y, x_{min}))^{n'_{tail}} = \left(\int_{x_{min}}^{x} Pr(y|\hat{\alpha}, x_{min}) dy \right)^{n'_{tail}}$$

- Thus $1 - F(x|\theta(Y, x_{min}))^{n'_{tail}}$ is the probability of observing at least 1 catastrophic event. Because the bootstrap Y is itself a random variable, to derive the marginal probability of observing at least one catastrophic event, we must integrate the conditional probability over the domain of the bootstrap distribution: $p(n_{tail}, \theta) = p(n_{tail}, Y) =$

$$\int dy_1 dy_2 \ldots dy_{n_{tail}} (1 - F(x|\theta(Y, x_{min}))^{n'_{tail}}) \prod_{i=1}^{n_{tail}} r(y_i|n_{tail})$$

where the trailing product series here is the probability of drawing the specific sequence of values $y_1, \ldots, y_{n_{tail}}$ from the fixed bootstrap distribution r.

Tail Model Estimating (Contd.)

- Finally, the total probability p of at least one catastrophic event is given by a binomial sum over this equation. (over n_{tail})
- As x_{min} is not known, it must be jointly estimated with θ from each bootstrap sample. MLE can not be used to estimate x_{min}(as it truncates Y, and hence the likelihood cannot be written). We use the K-S goodness of fit statistic minimization technique to estimate x_{min}. See Appendix A for details.
- Though the above-mentioned complete calculation is difficult to calculate analytically even for simple tail models, it is straightforward to estimate via Monte Carlo:

---

1. get bootstrap sample Y of same size n of X (leaving ℓ many catastrophic events).
2. jointly estimate x_{min} *and* θ from Y, see Appendix A.
3. Set $\rho = 1 - F(x; \hat{\theta})^{\ell + n_{tail}}$, the probability of observing at least catastrophic events under this bootstrap model.

Tail Model Estimating (Contd.)

- Averaging over the bootstraps yields the estimated probability $\hat{p} = < \rho >$ of observing at least ℓ catastrophic-sized events. The convergence of $\hat{p}$ is guaranteed so long as the number of bootstraps (step 1) tends to infinity.
- Confidence intervals on $\hat{p}$ may be constructed from the distribution of the ρ values.
- If the tail model's c.d.f. $F(x; \theta)$ in step 3 cannot be computed analytically, it can often be constructed numerically; failing that, ρ may always be estimated by sampling directly from the fitted model.

Model Comparison and Model Averaging

- Second, in complex social systems, the correct tail model is typically unknown and a poor choice may lead to severe misestimates of the true probability of a large event.
- We control for this model uncertainty by considering multiple tail models. Given these models and a common choice of x_{min},
- Here we use a likelihood ratio test to identify and discard the statistically implausible ones.
- In principle, the remaining models could be averaged to produce a single estimate with confidence intervals, but it has to be done with severe caution, as misspecification may harm the quality of the confidence interval by significant amount. Here we choose not to average over models, and present results from each model.

Model Comparison and Model Averaging (Contd.)

- Comparing the results from multiple tail models provides a test of robustness against model misspecification, for example, agreement across models that p ¿ˆ0.01 strengthens the conclusion that the event is not statistically unlikely. However, wide confidence intervals and disagreements on the precise probability of a large event reflect the inherent difficulty of identifying the correct tail structure.
- To select reasonable models to compare, standard model comparison approaches may be used. Here, we use a goodness-of-fit test to establish the plausibility of the power-law distribution and Vuong's likelihood ratio test (test for Model Selection based on Kullback–Leibler information criterion) to compare it with alternatives. This approach has the advantage that it can fail to choose one model over another if the difference in their likelihoods is statistically insignificant, given the data.

Tests of the Method's Accuracy

- Finally, large fluctuations in the distribution's upper tail occur precisely where we wish to have the most accuracy, leading to parameter uncertainty.
- To test the accuracy of our estimation algorithm, we examine its ability to recover the true probability of a rare event from synthetic data with known structure. To generate these synthetic data, we use the power-law distribution. By defining a catastrophic event x to be the largest generated event within the n synthetic values, we make the test particularly challenging because the largest value exhibits the greatest fluctuations of all. Detailed results are given in Appendix B.

Methodology (Contd.)

This combination of techniques provides a statistically principled and data-driven solution for estimating the probability of observing rare events in empirical data with unknown tail structure. If such an event is observed, the algorithm provides a measure of whether its occurrence was in fact unlikely, given the overall structure of the distribution's tail.
For instance, if the estimated probability is negligible (say, $p < 0.01$), the event may be judged statistically unlikely. When several tail models are plausible and agree that the probability is away from $p = 0$, the event can be judged to be statistically likely, despite the remaining uncertainty in the tail's structure.

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Historical Probability of 9/11

Having described our statistical approach, we now use it to estimate the historical probability of observing worldwide at least one 9/11-sized or larger terrorist event.

- Global databases of terrorist events show that event severities (number of deaths) are highly right-skewed or "heavy-tailed". We use the RAND-MIPT database [MIP08], which contains 13,274 deadly events worldwide from 1968–2007.
- The power law is a statistically plausible model of this distribution's tail, with $\hat{\alpha} = 2.4 \pm 0.1$, for $x \geq \hat{x}_{min} = 10$.
- A goodness-of-fit test fails to reject this model of tail event severities ($p = 0.40 \pm 0.03$ via Monte Carlo), implying that the deviations between the power-law model and the empirical data are indistinguishable from sampling noise.
- These give us the ability to treat the severity of the events as i.i.d. random variables. This perspective has much in common with stat. physics, in which particular population-level patterns emerge from a sea of individual interactions. We discuss limitations of this model later.

Historical Probability of 9/11 (Contd.)

- This apparent power-law pattern in global terrorism turns out to be remarkably robust. Although the estimated value of α varies somewhat with time the power-law pattern itself seems to persist over the 40 years despite large changes in the international system. It also appears (in past works) to be independent of the type of weapon (explosives, firearms, arson, knives, etc.), the emergence and increasing frequency of suicide attacks, the demise of many terrorist organizations, the economic development of the target country and organizational covariates like size (number of personnel), age and experience (total number of attacks).
- Comparing the power-law tail model against log-normal and stretched exponential (Weibull) distributions, via a likelihood ratio test, yields log-likelihood ratios of $\mathcal{R}$ = -0.278 (p = 0.78) and 0.772 (p = 0.44), respectively. However, neither of these values is statistically significant, as indicated by the large p-values for a test against $\mathcal{R} = 0$. So, while the power-law model is plausible, so are these alternatives.

Historical Probability of 9/11 (Contd.)

This ambiguity illustrates the difficulty of correctly identifying the tail's structure and reinforces the need to use multiple tail models in estimating the likelihood of a rare event like 9/11. Furthermore, it implies that slight visual deviations in the empirical distribution's upper tail (see Figure 1) should not be interpreted as support either for or against any of these models. In what follows, we consider estimates derived from all three.

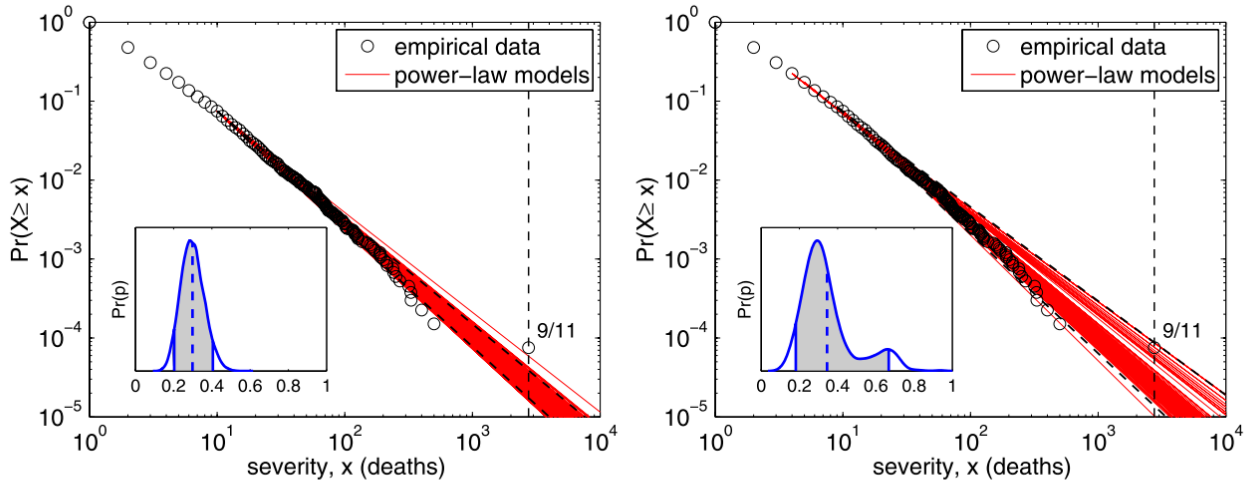


FIG. 1. *Empirical severity distribution with* 100 *bootstrap power-law models for* (a) *fixed* $x_{\min} = 10$ *and* (b) *estimated* $x_{\min}$. *Overprinting illustrates the ensemble of estimated models (dashed lines show* 90% *CI on* $\hat{\alpha}$*) and the inherent uncertainty in the tail structure. Insets show the* 90% *confidence intervals for the estimated probability of observing at least one* 9/11*-sized event.*

Historical Probability of 9/11 (Contd.)

Finally, using the RAND-MIPT event data (other sources [START (2011)] yield similar results; see Appendix C.2), we define $x \geq 2749$ to be a "catastrophic" event—the reported size of the New York City 9/11 events. Removing this event from the empirical data leaves the largest event as the 14 August 2007 coordinated truck bombing in Sinjar, Iraq, which produced approximately 500 fatalities. To illustrate the robustness of our results, we consider estimates derived from fixed and variable x_{min} and from our three tail models. We also analyze the impact of covariates like domestic versus international, the economic development of the target country and the type of weapon used.

Historical Probability of 9/11 (Contd.)

Uncertainty in the scaling parameter

- Let $x_{min} = 10$ be fixed. Figure 1(a) shows 100 of the fitted bootstrap models, illustrating that by accounting for the uncertainty in α, we obtain an ensemble of tail models and thus an ensemble of probability estimates for a catastrophic-sized event. The bootstrap parameter distribution $Pr(\hat{\alpha})$ has a mean $< \hat{\alpha} > = 2.40$, which agrees with the maximum likelihood value $\hat{\alpha} = 2.4$.
- To estimate the historical probability of 9/11, we use 10,000 bootstraps with x_{min} fixed. Letting p denote the overall probability from the algorithm, we find $\hat{p} = 0.299$, with 90% confidence intervals of [0.203, 0.405], or about a 30% chance over the 1968–2007 period.
- An event that occurs with probability 0.299 over 40 years is not a certainty. However, for global terrorism, this value is uncomfortably large and implies that, given the historical record, the size of 9/11 should not be considered a statistical fluke or outlier.

Historical Probability of 9/11 (Contd.)

Uncertainty in the tail location

- A fixed choice of x_{min} underestimates the uncertainty in p due to the tail's unknown structure. Jointly estimating α and x_{min} yields similar results, but with some interesting differences. Figure 1(b) shows 100 of the bootstrap models. The distribution of $\hat{x}_{min}$ is concentrated at $x_{min} = 9$ or 10 (48% of samples), with an average scaling exponent of $< \hat{\alpha} > = 2.40$. However, 15% of models choose $x_{min} = 4$ or 5, and these produce much heavier-tailed models, with $< \hat{\alpha} > = 2.21$.
- This bimodal distribution in $\hat{\alpha}$ is caused by slight curvature in the empirical mid-to-upper tail, which may arise from aggregating multiple types of local events into a single global distribution. The algorithm, however, accounts for this curvature by automatically estimating a slightly wider ensemble of models, with correspondingly greater density in the catastrophic range. As a result, the estimated probability is larger and the confidence intervals are wider. Using 10,000 bootstraps, we find $\hat{p} = 0.347$, with 90% confidence intervals of [0.182, 0.669], or about a 35% chance over the 1968–2007 period.

Historical Probability of 9/11 (Contd.)

Alternative Tail Models

- Comparing our estimates with those derived using log-normal and stretched exponential tail models provides a check on their robustness, especially if the alternative models yield dramatically different estimates.
- The mathematical forms of the alternatives are:

$$\textit{log-normal } Pr(x) \propto x^{-1} \exp[-(\ln x - \mu)^2/2\sigma^2],$$

$$\textit{stretched exp. } Pr(x) \propto x^{\beta-1} e^{-\lambda x^{\beta}},$$

where we restrict each to a "tail" domain $x_{min} \leq x < \inf$. Although both decay asymptotically faster than any power law, for certain parameter choices, these models can track a power law over finite ranges, which may yield only marginally lower estimates of large events.

Historical Probability of 9/11 (Contd.)

Alternate Tail Models (Contd.)

- To simplify the comparison between the tail models, we fix $x_{min} = 10$ and use 10,000 bootstraps for each fitted alternative tail model. This yields $\hat{p} = 0.112$ (CI: [0.063, 0.172]) for the log-normal and $\hat{p} = 0.187$ (CI: [0.115, 0.272]) for the stretched exponential, or roughly an 11% and 19% chance, respectively. These values are slightly lower than the estimates from the power-law model, but they too are consistently away from p = 0, which reinforces our conclusion that the size of 9/11 should not be considered a statistical outlier.
- Figure 2(a) shows the fitted ensembles for all three fixed-x_{min} tail models, and Figure 2(b) shows the bootstrap distributions $Pr(\hat{p})$ for these models, as well as the one with x_{min} free. Although the bootstrap distributions for the log-normal and stretched exponential are shifted to the left relative to the two power-law models, all distributions overlap and none place significant weight below p = 0.01.

Historical Probability of 9/11 (Contd.)

Alternate Tail Models (Contd.)

- The failure of the alternatives to disagree with the power law can be attributed to their estimated forms roughly tracking the power law's over the empirical data's range, which leads to similar probabilistic estimates of a catastrophic event.

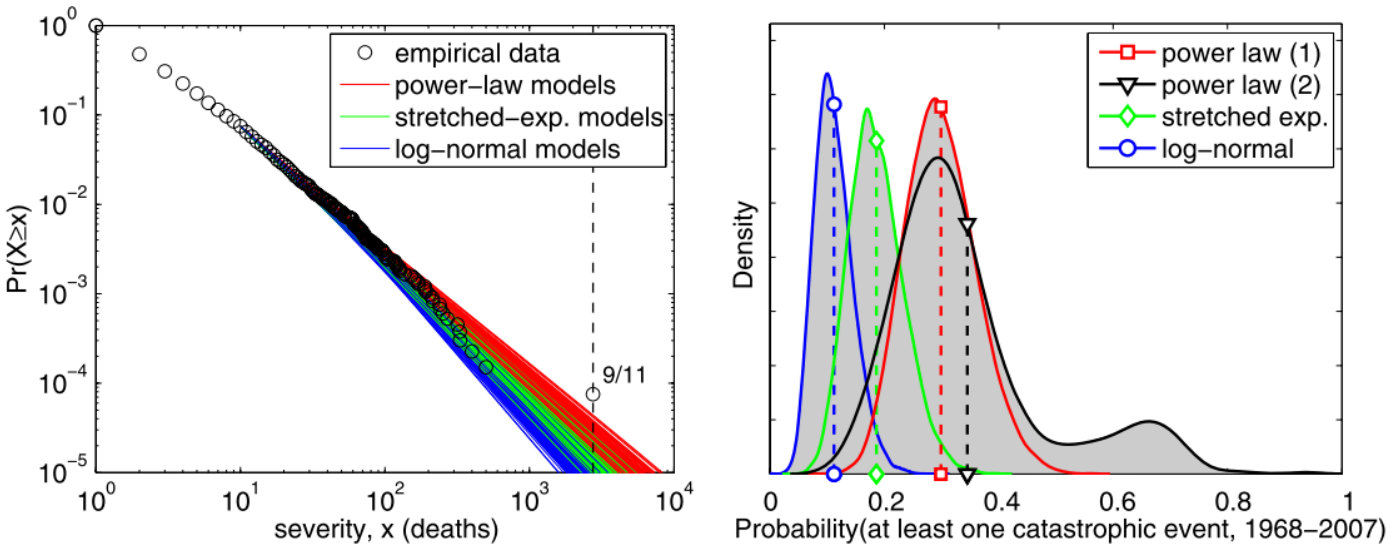


FIG. 2. (a) *Empirical event severities with* 100 *bootstrap models for the power-law, log-normal and stretched exponential tail models, with* $x_{\min} = 10$ *fixed.* (b) *Bootstrap distributions of* $\hat{p}$ *for each model, with overall estimates* (*Table* 1) *given by dashed lines.*

Historical Probability of 9/11 (Contd.)

Impact of Covariates

- Not all large terrorist events are of the same type, and thus our overall estimate is a function of the relative empirical frequency of different covariates and the structure of their marginal distributions. Here, we apply our procedure to the distributions associated with a few illustrative categorical event covariates to shed some additional light on the factors associated with large events. A generalization to and systematic analysis of arbitrary covariates is left for future work.
- For instance, international terrorist events, in which the attacker and target are from different countries, comprise 12% of the RAND-MIPT database and exhibit a much heavier-tailed distribution, with$\hat{\alpha} = 1.93 \pm 0.04$ and $\hat{x}_{min} = 1$ (see Appendix C.3.1). This heavier tail more than compensates for their scarcity, as we estimate $\hat{p} = 0.475$ (CI: [0.309, 0.610]; Figure 6(a)) for at least one such catastrophic event from 1968–2007.

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Forecasts: Another event like 9/11

If the social and political processes that generate terrorist events worldwide are roughly stationary [1], our algorithm can be used to make principled statistical forecasts about the future probability of a catastrophic event. A simple forecast requires estimating the number of events n expected over the fixed forecasting horizon t. Using the RAND-MIPT data as a starting point, we calculate the number of annual deadly events worldwide n_{year} over the past 10 years. Figure 3 shows the empirical trend for deadly terrorist events worldwide from 1998–2007, illustrating a 20-fold increase in n_{year}, from a low of 180 in 1999 to a high of 3555 in 2006. Much of the increase is attributable to conflicts in Iraq and Afghanistan; excluding events from these countries significantly reduces the increase in n_{year}, with the maxima now being 857 deadly events in 2002 and 673 in 2006. However, the fraction of events that are severe ($x \geq 10$) remains constant, averaging $< p_{tail} > = 0.082684$ (or about 8.3%) in the former case and 0.072601 (or about 7.3%) in the latter.

[1] An improvement on this is discussed in the Improvements

Forecasts: Another event like 9/11 (Contd.)

An estimated trend over the next decade could be obtained via fitting standard statistical models to annual data or by soliciting judgements from domain experts about specific conflicts. For instance, Iraq and Afghanistan may decrease their production rates of new events over the next decade, leading n_{year} to decrease unless other conflicts replace their contributions. Rather than make potentially overly specific predictions, we instead consider three rough scenarios (the future's trajectory will presumably lay somewhere between):

1. an optimistic scenario, in which the average number of terrorist attacks worldwide per year returns to its 1998–2002 level, at about $< n_{year} > = 400$ annual events.
2. a status quo scenario, where it remains at the 2007 level, at about 2000 annual events.
3. a pessimistic scenario, in which it increases to about 10,000 annual events.

Forecasts: Another event like 9/11 (Contd.)

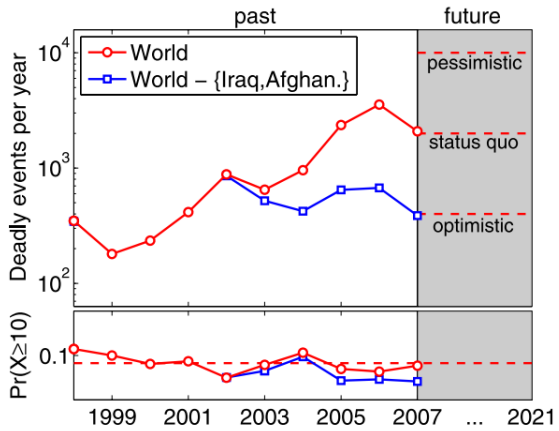


FIG. 3. *(Upper) number of deadly (domestic and international) terrorist events worldwide for the 10-year period 1998–2007, and three forecast scenarios. (Lower) fraction of events that are severe, killing at least 10 individuals and its 10-year average (dashed line).*

TABLE 2

Forecast estimates of at least one catastrophic event worldwide over a 10-year period, using three tail models in each of three forecast scenarios

	Pr($x \geq 2749$) forecast, 2012–2021		
Tail model	**"Optimistic"** $n_{\text{year}} \approx 400$	**"Status quo"** $n_{\text{year}} \approx 2000$	**"Pessimistic"** $n_{\text{year}} \approx 10{,}000$
Power law	0.117	0.461	0.944
Stretched exp.	0.072	0.306	0.823
log-normal	0.043	0.193	0.643

Forecasts: Another event like 9/11 (Contd.)

A quantitative statistical forecast is then obtained by applying the estimation algorithm to the historical data (now including the 9/11 event) and then generating synthetic data with the estimated number of future events n_{tail}. For each scenario, we choose $n_{decade} = 10 \times n_{year}$ and choose n_{tail} with $p_{tail} = 0.082684$ (historical average). Finally, we fix $x_{min} = 10$ to facilitate comparison with our alternative tail models.

Table 2 summarizes the results, using 100,000 bootstraps for each of the three tail models in the three forecast scenarios. Under the status quo scenario, all three models forecast a 19-46% chance of at least one catastrophic event worldwide in the next decade. In the optimistic scenario, with events worldwide being about 5 times less common, the models forecast a 4-12% chance. These estimates depend strongly on the overall frequency of terrorist events n_{year}. Thus, the greater the popularity of terrorism worldwide, that is, the more often terrorist attacks are launched, the greater the general likelihood that at least one will be catastrophic.

Table of Contents

1. Introduction
2. Methodology
3. Historical Probability of 9/11
4. Statistical Forecast
5. Improvements

Limitations & Improvements

During the process of our modelling, we took a few assumptions, which might have affected the estimates. Some of them are discussed below with probable improvements:

- **Stationary Event Generation process**: In the forecast section, we assumed Stationary Event Generation as our data generation model. But like technology, population and culture which exhibit non-stationary dynamics, Terrorism also has similar traits. Hence, maybe Non-stationary Event Generation processes like ARIMA, can be used for improvements in forecasting.
- **Silence reg. efforts to prevent events or mitigate severity**: The data we have been basing all our estimates on, is collected on the occurrence of such events. This makes our estimates conditioned on efforts to prevent such events or mitigate the severity, which in turn makes our estimates biased. More accurate estimates may be achievable by incorporating models of policy consequences or interactions between different factors.

Limitations & Improvements (Contd.)

- We have also neglected the data on where the incident occurred. It may be a significant misspecification that a 9/11 sized incident in Syria or other fighting countries, has the same probability of occurrence and similar effects. Incorporating more fine-grained spatial structure, for example, to make country-level estimates, or incorporating tactical information, for example, about specific CBRN attacks, may be possible. Such refinements will likely require strong assumptions about many context-specific factors, and it remains unclear whether accurate estimates at these scales. At the worldwide level of our analysis, such contingencies appear to play a relatively small role in the global pattern, perhaps because local-level processes are roughly independent. This independence may allow large-scale general patterns to emerge from small-scale contingent chaos via a Central Limit Theorem averaging process, just as regularities in birth rates exist in populations despite high contingency for any particular conception that can be made.

References

[MIP08] RAND MIPT. *RAND Database of Worldwide Terrorism Incidents*. https://www.rand.org/nsrd/projects/terrorism-incidents/download.html. 2008.

[CW13] Aaron Clauset and Ryan Woodard. "Estimating the historical and future probabilities of large terrorist events". In: *The Annals of Applied Statistics* 7.4 (Dec. 2013). ISSN: 1932-6157. DOI: 10.1214/12-aoas614. URL: http://dx.doi.org/10.1214/12-AOAS614.

[STA17] START. *START Dataset on Global Terrorism Dataset*. http://www.start.umd.edu/gtd/. 2017.

Appendix A: Tail Models

Estimates of α

The functional form and normalization of the tail model should follow the type of empirical data used. For example, if the empirical data are real-valued, the power-law tail model has the form:

$$Pr(y|\alpha, x_{min}) = \left(\frac{\alpha - 1}{x_{min}}\right)\left(\frac{y}{x_{min}}\right)^{-\alpha}, \alpha > 1, y_{min} > 0.$$

Then the MLE for α would be, $\hat{\alpha} = 1 + \frac{n}{\sum_{i=1}^{n} \ln(\frac{x_i}{x_{min}})}$.

Though the severity of a terrorist attack is an integer. Thus, in our analysis of terrorist event severities, we use the discrete form of the power-law distribution, $Pr(y|\alpha, x_{min}) = \frac{y^{-\alpha}}{\zeta(\alpha, x_{min})}, \alpha > 1, y_{min} > 0$, where $\zeta(\alpha, x_{min}) = \sum_{i=x_{min}}^{\inf} i^{-\alpha}$ is the generalized or incomplete zeta function. $\hat{\alpha}$ can be estimated by directly maximising the log-likelihood

$$\mathcal{L}(\alpha) = -n \log(\zeta(\alpha, x_{min})) - \alpha \sum_{i=1}^{n} \ln x_i$$

Appendix A: Tail Models (Contd.)

Alternative Tail Models

Our alternative tail models are the log-normal and the stretched exponential distributions, modified to include a truncating parameter x_{min}. These distributions are normally defined on continuous variables. The structure of their extreme upper tails for $x_{min} = 10$, however, is close to that of their discrete versions, and the continuous models are significantly easier to estimate from data. For the results presented in the main text, we used the continuous approximation of the upper tails for these models.

Appendix A: Tail Models (Contd.)

Estimation of x_{min}

For the estimation of x_{min}, we use the Kolmogorov–Smirnov goodness-of-fit statistic minimization (K-S minimization) technique. This method falls in the general class of distance minimization techniques for selecting the size of the tail.
The K-S statistic is defined as:

$$D = \max_{x \geq x_{min}} |S(x) - P(x)|$$

where S(x) is the empirical CDF of data, and P(x) is the CDF of the maximum-likelihood power-law model for the region $x \geq x_{min}$. Our estimate $\hat{x}_{min}$ is s then the value of x_{min} that minimizes D. In the event of a tie between several choices for x_min, we choose the smaller value, which improves the statistical power of subsequent analyses by choosing the larger effective sample size.

Appendix B: Estimator Accuracy

We quantify the expected accuracy of our estimates under two experimental regimes in which the true probability of a catastrophic event can be calculated analytically.

1. Draw n values i.i.d. from a power-law distribution with $x_{min} = 10$ and some α; define $x = \max_i\{x_i\}$, the largest value within that sample. This step ensures that we treat the synthetic data exactly as we treated our empirical data and provides a particularly challenging test, as the largest generated value exhibits the greatest statistical fluctuations.
2. Draw n - 1 i.i.d. values from a power-law distribution with $x_{min} = 10$ and some α, and then add a single value of size x whose true probability of appearing under the generative model is $p = 0.001$, that is, we contaminate the data set with a genuine outlier.

The next figure shows the results of both experiments, where we measure the mean absolute error (MAE) and the mean ratio between $\hat{p}$ and the true p.

Appendix B: Estimator Accuracy (Contd.)

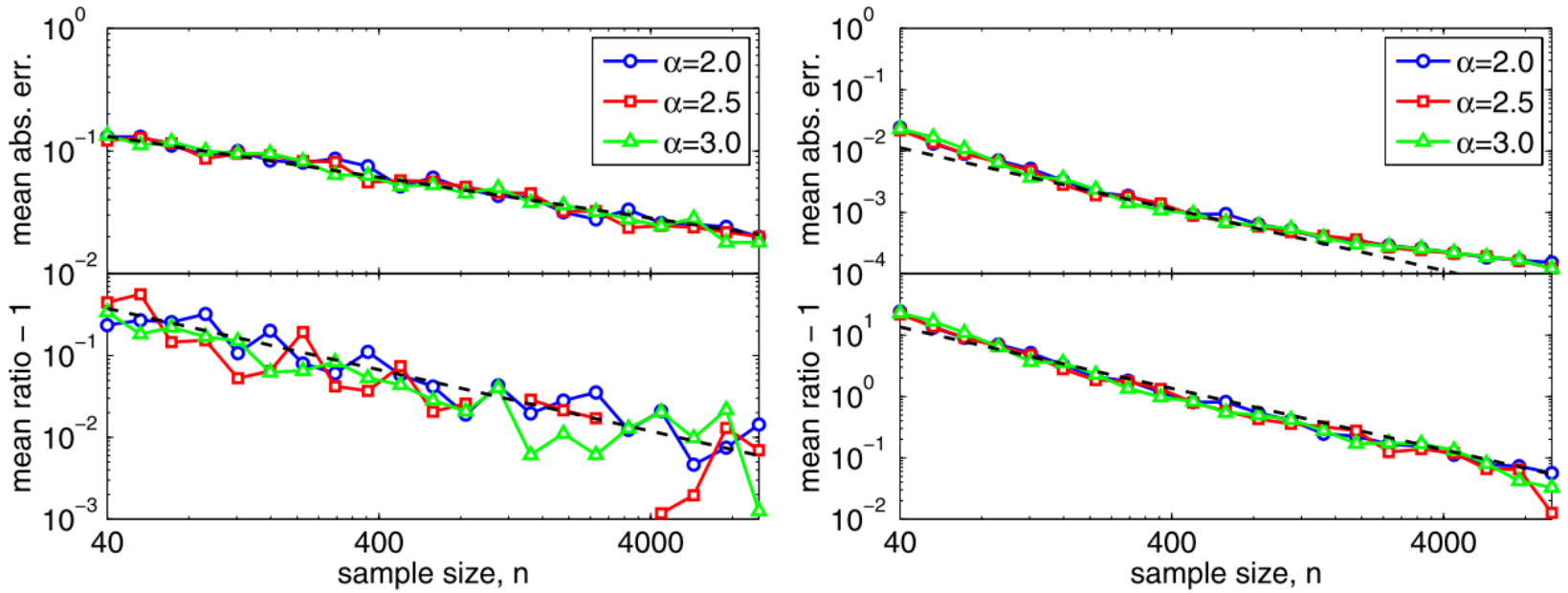


FIG. 4. *The mean absolute error* $\langle|\hat{p}-p|\rangle$ *and mean relative error* $\langle\hat{p}/p\rangle-1$ *for* (a) n *values drawn i.i.d. from a stationary power-law distribution with* $x_{\min}=10$ *and some* α*, with the target size being the single largest value in the draw, and for* (b) $n-1$ *values to which we add a single outlier (with true* $p=0.001$*). In both experiments, both types of errors are small even for fairly small sample sizes and decay further as* n *increases.*

In the first experiment, the error rate decays like $O(n^{-\frac{1}{3}})$, approaching 0.01 error rates as n approaches 5000, while in the second it decays like $O(n^{-1})$ up to about n = 4000, above which the rate of decay gets slightly thinner.

Appendix C: Robustness Checks (Simpler Models)

We present three checks of the robustness of our probability estimates:

1. using simple parametric models without the bootstrap
2. using an alternative source of terrorist event data
3. using event covariates to refine the estimates

In each case, we find roughly similar-sized estimates.

Using Simpler models

A simpler model for estimating the historical probability of a 9/11-sized or larger terrorist event assumes the following: (i) a stationary generative process for event severities worldwide, (ii) event sizes are i.i.d. random variables drawn from (iii) a power-law distribution that (iv) spans the entire range of possible severities ($x_m in = 1$), and (v) has a precisely-known parameter value $\alpha = 2.4$.

Appendix C: Robustness Checks (Simpler Models) (Contd.)

Our historical probability estimate is then
$\hat{p} = 1 - (F(2749))^{n_{tail}} = 1 - (1 - (Pr(x > 2749)))^{n} \approx 1 - e^{n \times Pr(x > 2749)}$
$= 1 - exp[n \left(\frac{2749}{x_{min}}\right)^{1-\alpha}] = 1 - exp[13274 \times 0.000015316] = 0.184$ or 18%.
However, this calculation underestimates the true probability of a large event because the empirical distribution decays more slowly than a power law with $\alpha = 2.4$ at small values of x. Empirically 7.5% of the 13,274 fatal events have at least 10 fatalities, but a simple application of (C.2) using x $= 10$ shows that our model predicts that only 4.0% of events should be this severe. Thus, events with x ≥ 10 occur empirically almost twice as often as expected, which leads to a significant underestimate of p.

By restricting the power-law model to the tail of the distribution, setting $x_{min} = 10$ and noting that only n $= 994$ events had at least this severity over the 40-year period, we can make a more accurate estimate. Repeating the analysis above, we find q(2749) $= 0.0000288098$ and $\hat{p} =$ 0.318, or about a 32% chance of a catastrophic event, a value more in line with the estimates derived using our bootstrap-based approach.

Appendix C: Robustness Checks (Alternate Data Source)

An alternative source[STA17] of global terrorism event data is the Global Terrorism Database [START (2011)] , which contains 98,112 events worldwide from 1970–2007. Of these, 38,318 were deadly (x $>$ 0). Some events have fractional severities due to having their total fatality count divided evenly among multiple event records; we recombined each group of fractional-severity events into a single event, yielding 38,255 deadly events over 38 years. Analyzing the GTD data thus provides a check on our results for the RAND-MIPT data.

The best fitting power-law model obtained using the methodology of Clauset, Shalizi and Newman (2009) is $\hat{\alpha} = 2.91 \pm 0.22$ and $\hat{x}_{min} = 39$. The p ¡ 0.1 for this model may be attributable to the unusually large number of perfectly round number severities in the data set, for example, 10, 20, 100, 200, etc., which indicates rounding effects in the reporting. (These appear in Figure 5 as small discontinuous drops in the complementary CDF at round-number locations; true powerlaw distributed data have no preference for round numbers and thus their presence is a statistically significant deviation from the power-law form.)

Appendix C: Robustness Checks (Alternate Data Source)

Using the algorithm described in the main text with 10,000 bootstraps, we estimate a 38-year probability of at least one catastrophic event as $\hat{p} = 0.534$ (with 90% CI [0.115, 0.848]) or about a 53% chance. Repeating our analysis using the two alternative tail models yields only a modest decrease, as with the RAND-MIPT data.

Figure 5 shows the empirical fatality distribution along with 100 fitted power-law models, illustrating the heavy-tailed structure of the GTD severity data. Notably, the maximum likelihood estimate for α is larger here (indicating a less heavy tail) than for the RAND-MIPT data. However, the marginal distribution $Pr(\hat{\alpha})$ is bimodal, with one mode centered on $\alpha = 2.93$ and a second larger mode centered at roughly $\alpha = 2.4$, in agreement with the RAND-MIPT data. Furthermore, the failure of the GTD-estimated $\hat{p}$ to be dramatically lower than the one estimated using RAND-MIPT data supports our conclusion that the size of 9/11 was not statistically unlikely.

Appendix C: Robustness Checks (Alternate Data Source)

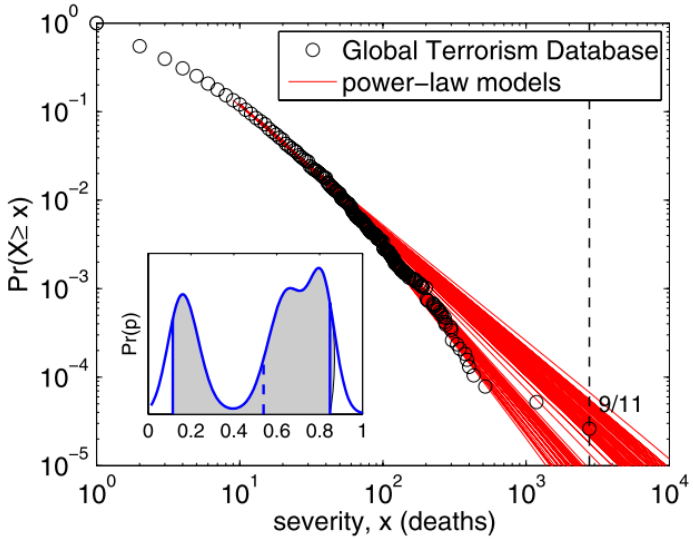
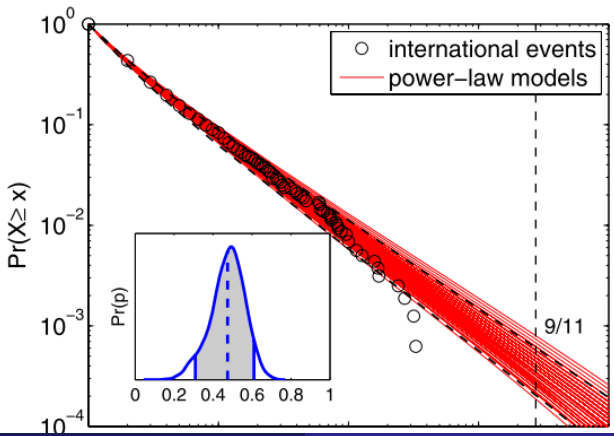
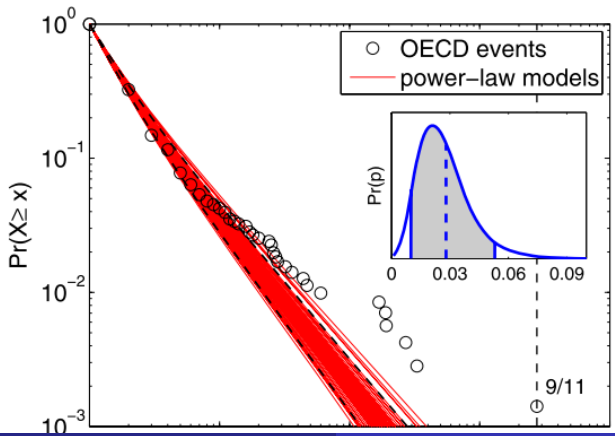


FIG. 5. *Empirical distribution of event severities from the GTD* [*START* (2011)] *with* 100 *power-law models, fitted to bootstraps of the data. Inset shows the estimated distribution of binomial probabilities* $\Pr(\hat{p})$ *for one or more catastrophic events.*





Appendix C: Robustness Checks (Different Machinery)

Economic Development: For economically developed nations, defined here as the member countries of the Organisation for Economic Co-operation and Development (OECD), as of the end of the period covered by the RAND-MIPT data, which are 5.3% of all deadly events, we analyse our data conditioned to this only.

The empirical distribution [Figure 6(b)] of event severities shows an unusual structure, with the upper tail ($x \geq 10$) decaying more slowly than the lower tail. To handle this oddity, we conduct two tests.

First, we consider the entire OECD data set, estimating both α and x_{min}. Using 10,000 bootstraps yields $\hat{p} = 0.028$ (with 90% CI [0.010, 0.053]) or roughly a 3% chance over the 40-year period, which is slightly above our $p = 0.01$ cutoff for a statistically unlikely event. Figure 6(b) shows the resulting ensemble of fitted models, illustrating that the algorithm is placing very little weight on the upper tail.

Second, we apply the algorithm with a fixed $x_{min} = 10$ in order to focus explicitly on the distribution's upper tail. In this case, 10,000 bootstraps yield $\hat{p} = 0.225$, with 90% CI as [0.037, 0.499].

Appendix C: Robustness Checks (Different Machinery) (Contd.)

Type of weapon

Finally, we consider the impact of the attack's weapon type, and we generalize the estimation algorithm to the multi-covariate case. Events are classified as (i) chemical or biological, (ii) explosives (includes remotely detonated devices), (iii) fire, arson and firebombs, (iv) firearms, (v) knives and other sharp objects, and (vi) other, unknown or unconventional. Given the empirically observed distributions over these covariates, we would like to know the probability of observing at least one catastrophic-sized event from any weapon type.

This requires generalizing our Monte Carlo algorithm: let $(x, c)_i$ denote the severity x and categorical covariate c for the i^{th} event. Thus, denote the empirical data by $\mathrm{X} = \{(x, c)_i\}$.

Appendix C: Robustness Checks (Different Machinery) (Contd.)

Algorithm

1. Generate Y by a drawing $(y, c)_j$, j = 1,...,n, uniformly at random, with replacement, from the original data $\{(x, c)_i\}$ (apart from the catastrophic events).
2. For each covariate type c in Y, jointly estimate $\hat{x}_{min}^{(c)}$ and the tail-model parameters $\theta(c)$, and compute $n_{tail}^{(c)} = \#\{y_j \geq \hat{x}_{min}^{(c)}\}$.
3. For each covariate type c in Y, generate a synthetic data set by drawing $n_{tail}^{(c)}$ random deviates from the fitted tail model with parameters $\hat{\theta}^{(c)}$.
4. If any of the covariate sequences of synthetic events includes at least events of size x or greater, set $\rho = 1$; otherwise, set it to zero.

Appendix C: Robustness Checks (Different Machinery) (Contd.)

In applying this algorithm to our data, we choose = 1 and x = 2749, as with our other analyses. In step 2, we again use the KS-minimization technique to choose x_{min} and estimate θ for a power-law tail model via maximum likelihood. Finally, as with the univariate version of the algorithm, bootstrap confidence intervals may be obtained, both for the general hazard and the covariate-specific hazard, by repeating steps 3 and 4 many times for each bootstrap and tracking the distribution of binomial probabilities.

Using 10,000 bootstraps and drawing 1000 synthetic data sets from each bootstrap, we estimate $\hat{p} = 0.564$, with 90% confidence intervals of [x,y]. Again, this value is well above the cutoff for a 9/11-sized attack being statistically unlikely. As a side effect of this calculation, we may also calculate the probability that a catastrophic event will be generated by a particular type of weapon.

Appendix C: Robustness Checks (Different Machinery) (Contd.)

The following table gives these marginal probability estimates, which are greatest for explosives, fire, firearms and unconventional weapon types.

Weapon type	**Historical $\hat{p}$**	**90% CI**
Chem. or bio.	0.023	[0.000, 0.085]
Explosives	0.374	[0.167, 0.766]
Fire	0.137	[0.012, 0.339]
Firearms	0.118	[0.015, 0.320]
Knives	0.009	[0.001, 0.021]
Other or unknown	0.055	[0.000, 0.236]
Any	0.564	[0.338, 0.839]

(The sum of marginal probabilities exceeds that of the "any" column because in some trials, catastrophic events are generated in multiple categories.)

Any Questions?

Thank You