Estimating the historical and future probabilities of large terrorist events

Aaron Clauset[1, 2, 3, *] and Ryan Woodard[4, †]

[1]*Department of Computer Science, University of Colorado, Boulder CO, 80309, USA*
[2]*BioFrontiers Institute, University of Colorado, Boulder CO, 80303, USA*
[3]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, 87501, USA*
[4]*Department of Management, Technology and Economics, ETH Zürich, Kreuzplatz 5, CH-8032 Zürich, Switzerland*

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 instance, the 9/11 terrorist events produced nearly 3000 fatalities, nearly six times more than the next largest event. But, was this enormous loss of life statistically unlikely given modern terrorism's historical record? Accurately estimating the probability of such an event is complicated by the large fluctuations in the empirical distribution's upper tail. We present a generic statistical algorithm for making such estimates, which combines semi-parametric models of tail behavior and a non-parametric bootstrap. Applied to a global database of terrorist events, we estimate the worldwide historical probability of observing at least one 9/11-sized or larger event since 1968 to be 11–35%. These results are robust to conditioning on global variations in economic development, domestic versus international events, the type of weapon used and a truncated history that stops at 1998. We then use this procedure to make a data-driven statistical forecast of at least one similar event over the next decade.

The September 11th terrorist attacks were the largest such events in modern history, killing nearly 3000 people [1, 2]. Given their severity, should these attacks be considered statistically unlikely or even outliers? What is the likelihood of another September 11th-sized or larger terrorist event, worldwide, over the next decade?

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 [3–5], 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 [6, 7], and for estimating the likelihood of even larger events, including mass-casualty chemical, biological, radioactive or nuclear (CBRN) events [8, 9].

The rarity of events like 9/11 poses two technical problems: (i) we typically lack quantitative mechanism-based models with demonstrated predictive power at the global scale (which is particularly problematic for CBRN events) and (ii) 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 (see [10–15]) and can complicate both selecting the correct model of the tail's structure and accurately estimating its parameters [16]. Misspecification can lead to severe underestimates of the true probability of large events, e.g., in classical financial risk models [17, 18].

Little research on terrorism has focused on directly modeling the number of deaths ("severity")[1] in individual terrorist events [5]. When deaths are considered, they are typically aggregated and used as a covariate to understand other aspects of terrorism, e.g., trends over time [19, 20], the when, where, what, how and why of the resort to terrorism [21–23], differences between organizations [24], or the incident rates or outcomes of events [19, 25]. Such efforts have used time series analysis [19, 20, 25], qualitative models or human expertise of specific scenarios, actors, targets or attacks [26] or quantitative models based on factor analysis [27, 28], social networks [29, 30] or formal adversarial interactions [22, 31, 32]. Most of this work focuses on modeling central tendencies, treats large events like 9/11 as outliers, and says little about their quantitative probability [33] or their long-term hazard.

Here, we describe a statistical algorithm for estimating the probability of large events in complex social systems in general, and in global terrorism in particular. Making only broad-scale and long-term probabilistic estimates, our approach is related to techniques used in seismology, forestry, hydrology and natural disaster insurance to estimate the probabilities of individual rare catastrophic events [6, 7, 34–37]. 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 quan-

*Electronic address: aaron.clauset@colorado.edu
†Electronic address: rwoodard@ethz.ch

[1] Other notions of event "size" or severity, which we do not explore here, might be the economic cost, number injured, political impact, etc. To the extent that such notions may be quantitatively measured, our algorithm could also be applied to them.

titative estimate of the probability, with uncertainty, of a large event. The algorithm also naturally generalizes to include certain event covariates, which can shed additional light on the probability of large events of different types.

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 non-trivial magnitude of this historical probability to be highly robust, a direct consequence of the highly right-skewed or "heavy-tailed" structure of event sizes [33]. Thus, an event of 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.

I. ESTIMATING THE PROBABILITY OF A LARGE EVENT

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 largest several observations. This task is distinct from estimating the distribution of maxima within a sample [6, 7], and is more closely related to the peaks-over-threshold literature in hydrology, seismology, forestry, finance and insurance [6, 7, 34–39]. 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. 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 modeling both the distribution's body and tail [6, 7, 39] with the less difficult problem of identifying a value $x_{\min}$ above which a model of the tail alone fits well.[2] That is, choose some $x_{\min}$ and a tail model $\Pr(x \,|\, \theta, x_{\min})$ defined on $x \in [x_{\min}, \infty)$. We will revisit the problem of choosing $x_{\min}$ below.

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 $x_{\min}$, we use a likelihood ratio test to identify and discard the statistically implausible ones [16]. In principle, the remaining models could be averaged to produce a single estimate with confidence intervals [40], e.g., to aid decision makers. We return to this point in more detail below.

Finally, large fluctuations in the distribution's upper tail occur precisely where we wish to have the most accuracy, leading to parameter uncertainty. Using a non-parametric bootstrap [41] to simulate the generative process of event sizes, we incorporate the empirical data's inherent variability into the estimated parameters, weight models by their likelihood under the bootstrap distribution and construct extreme value confidence intervals [37].

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.

A. The method

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.[3] In principle, any size x and any value ℓ 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. Here we describe the method in terms of univariate distributions, but its generalization to certain covariates is straightforward (see Appendix C 3 c).

Let $\Pr(x \,|\, \theta, x_{\min})$ denote a particular tail model with parameters θ, let $\{x_i\}$ denote the n empirical event sizes (sans 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

[2] The notation $x_{\min}$ should not be confused with the first order statistic, $x_{(1)} = \min_i x_i$.

[3] Consider events to be generated by a kind of marked point process [42], where marks indicate either the event's severity or that it exceeded some threshold x. Although we assume the number of marks n to be fixed, this could be relaxed to incorporate additional uncertainty into the algorithm's output.

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 $p_{\text{tail}} = \#\{x_i \geq x_{\min}\}/n$, and in each bootstrap the number is a binomial random variable with probability p_{tail}:

$$n_{\text{tail}} \sim \text{Binomial}(n, p_{\text{tail}}) \ . \tag{1}$$

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

Given that choice, the probability under the fitted model that not one of $n'_{\text{tail}} = 1 + n_{\text{tail}}$ events is at least as big as x is

$$F(x \,|\, \theta(Y, x_{\min}))^{n'_{\text{tail}}} = \left(\int_{x_{\min}}^{x} \Pr(y \,|\, \hat{\alpha}, x_{\min}) \mathrm{d}y \right)^{n'_{\text{tail}}} \ . \tag{2}$$

Thus, $1 - F(x \,|\, \theta(Y, x_{\min}))^{n'_{\text{tail}}}$ is the probability that at least one event is of catastrophic size. 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:

$$p(n_{\text{tail}}, \theta) = p(n_{\text{tail}}, Y) = \int \mathrm{d}y_1 \cdots \mathrm{d}y_{n_{\text{tail}}} \left(1 - F(x; \theta(Y, x_{\min}))^{n'_{\text{tail}}}\right) \prod_{i=1}^{n_{\text{tail}}} r(y_i \,|\, n_{\text{tail}}) \tag{3}$$

The trailing product series here is the probability of drawing the specific sequence of values $y_1, \ldots y_{n_{\text{tail}}}$ from the fixed bootstrap distribution r. Finally, the total probability p of at least one catastrophic event is given by a binomial sum over this equation.[4]

When the correct value $x_{\min}$ is not known, it must be estimated jointly with θ on each bootstrap. Maximum likelihood cannot be used for this task, because $x_{\min}$ truncates Y. Several principled methods for automatically choosing $x_{\min}$ exist, e.g., [16, 33, 37, 43–46]. So long as the choice of $x_{\min}$ is also a deterministic function of Y, the above expression still holds. Variation in $x_{\min}$ across the bootstraps, however, leads to different numbers of observations n_{tail} in the tail region. The binomial probability p_{tail} is then itself a random variable determined by Y, and n_{tail} is a random variable drawn from a mixture of these binomial distributions.

Analytically completing the above calculation can be difficult, even for simple tail models, but it is straightforward to estimate numerically via Monte Carlo:

1. Given the n empirical sizes, generate Y by drawing y_j, $j = 1, \ldots, n$, uniformly at random, with replacement, from the observed $\{x_i\}$ (sans the ℓ catastrophic events).

2. Jointly estimate the tail model's parameters θ and $x_{\min}$ on Y, and compute $n_{\text{tail}} = \#\{y_j \geq \hat{x}_{\min}\}$ (see Appendix A).

3. Set $\rho = 1 - F(x; \hat{\theta})^{\ell + n_{\text{tail}}}$, the probability of observing at least ℓ catastrophic events under this bootstrap model.

Averaging over the bootstraps yields the estimated probability $\hat{p} = \langle \rho \rangle$ 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 [41]. Confidence intervals on $\hat{p}$ [37, 41] may be constructed from the distribution of the ρ values. If the tail model's cdf $F(x; \theta)$ 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.

B. Model comparison and model averaging

In complex social systems, we typically do not know *a priori* which particular tail model is correct, and the algorithm described above will give no warning of a bad choice (but see [16]). This issue is partly mitigated by estimating $x_{\min}$, which allows us to focus our modeling efforts on the upper tail alone. But, without additional evidence of the model's statistical plausibility, the estimate $\hat{p}$ should be treated as provisional.

Comparing the results from multiple tail models provides a test of robustness against model misspecification, e.g., agreement across models that $\hat{p} > 0.01$ 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, e.g., a fully Bayesian approach [47], cross-validation [48], or minimum description length [49]. Here, we use a goodness-of-fit test to establish the plausibility of the power-law distribution [16] and Vuong's likelihood ratio test [16, 50] 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 statistically insignificant, given the data.

In some circumstances, we may wish to average the resulting models to produce a single estimate with confidence intervals, e.g., to aid decision makers. However,

---

[4] We may calculate p in either of two ways: (i) we draw n_{tail} events from a tail model alone, or (ii) we draw n events from a conditional model, in which the per-event probability is $q(x) = \Pr(X \geq x \,|\, X \geq x_{\min}) \Pr(X \geq x_{\min}) = p_{\text{tail}}(1 - F(x \,|\, \theta, x_{\min}))$. When the probability of a catastrophic event is small, these calculations yield equivalent results.

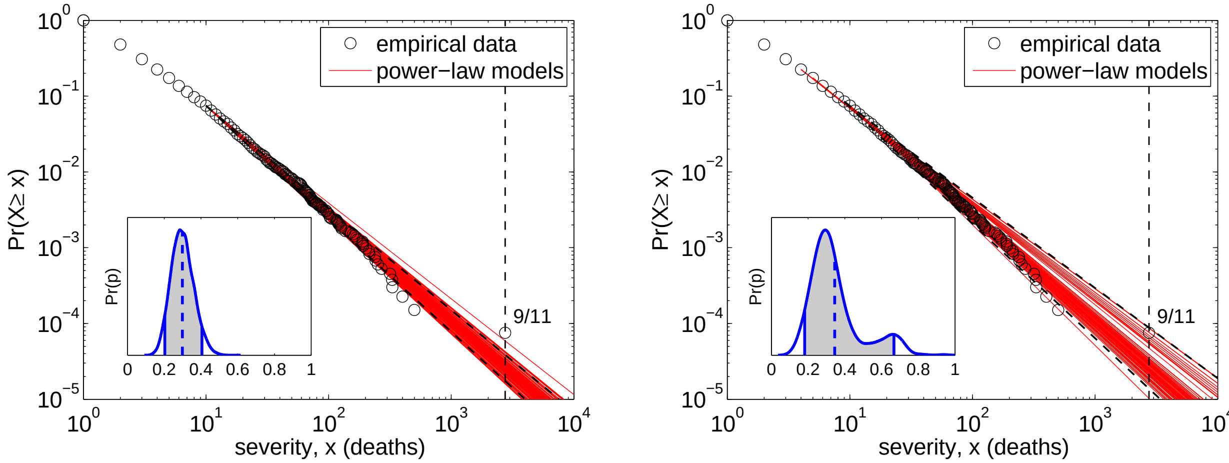

FIG. 1: Empirical severity distribution with 100 bootstrap power-law models for (*a*) fixed $x_{\min} = 10$ and (*b*) estimated $x_{\min}$. Overprinting illustrates the ensemble of estimated models (dashed lines show 90% CI on $\hat{\alpha}$) and the inherent uncertainty in the tail structure. Insets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

averaging poses special risks and technical problems for estimating the probability of large events. For instance, traditional approaches to averaging can obscure the inherent uncertainty in the tail's structure and can produce spuriously precise confidence intervals [40, 51]; a Bayesian approach would be inconsistent with our existing framework; and an appropriate frequentist framework is not currently available, although one may be possible using insights from [52].

Thus, in our application below, we elect not to average and instead we present results for each model. Even without averaging, however, several valuable insights may be drawn.

C. Tests of the method's accuracy

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

$$\Pr(y) \propto y^{-\alpha} \;, \tag{4}$$

where $\alpha > 1$ is the "scaling" parameter and $y \geq x_{\min} > 0$. When $\alpha < 2$, this distribution exhibits infinite variance and produces extreme fluctuations in the upper tail of finite-size samples. 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.

We find that despite the large fluctuations generated by the power-law distribution, the algorithm performs well: the mean absolute error $\langle|\hat{p} - p|\rangle$ is small even for samples with less than 100 events, and decays like $O(n^{-1/3})$. A small absolute deviation, however, may be an enormous relative deviation, e.g., if the true probability tends to zero or one. Our algorithm does not make this type of error: the mean ratio of the estimated and true probabilities $\langle\hat{p}/p\rangle$ remains close to 1 and thus the estimate is close in relative terms, being only a few percent off for $n \gtrsim 100$ events.

II. 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" [1, 2]. We use the RAND-MIPT database [1], 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 $x \geq \hat{x}_{\min} = 10$ [16, 33]. A goodness-of-fit test fails to reject this model of tail event severities ($p = 0.40 \pm 0.03$ via Monte Carlo [16]), implying that the deviations between the power-law model and the empirical data are indistinguishable from sampling noise.

This fact gives us license to treat as iid random variables the severity of these events. This treatment does force a particular and uncommon theoretical perspective on terrorism, in which a single global "process" produces events, even if the actions of individual terrorists or terrorist organizations are primarily driven by local events. This perspective has much in common with statistical physics, in which particular population-level patterns emerge from a sea of individual interactions. We discuss limitations of this perspective in Section IV.

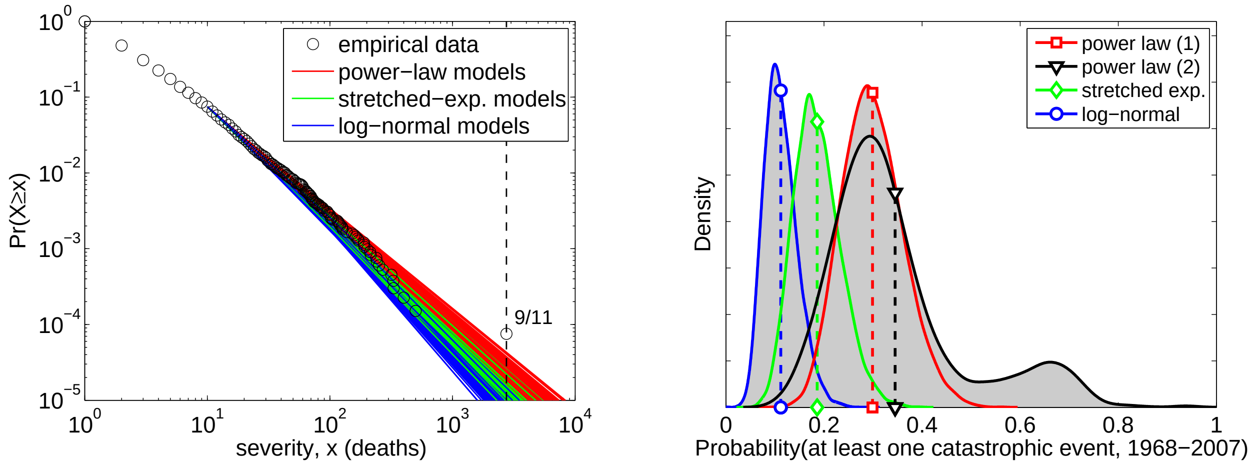

FIG. 2: (a) Empirical event severities with 100 bootstrap models for the power-law, log-normal and stretched exponential tail models, with $x_{\min} = 10$ fixed. (b) Bootstrap distributions of $\hat{p}$ for each model, with overall estimates (Table I) given by dashed lines.

Past work shows that this apparent power-law pattern in global terrorism is remarkably robust. Although the estimated value of α varies somewhat with time [33], the power-law pattern itself seems to persist over the 40 year period despite large changes in the international system. It also appears 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 [33] and organizational covariates like size (number of personnel), age and experience (total number of attacks) [53].

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 [16]. However, neither of these values is statistically significant, as indicated by the large p-values for a test against $\mathcal{R} = 0$. Thus, while the power-law model is plausible, so too are these alternatives. 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 Fig. 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.

To apply our algorithm to this problem, we must make several choices. For consistency with past work on the frequency of severe terrorist events [16, 33], we choose $x_{\min}$ automatically by minimizing the Kolmogorov-Smirnov goodness-of-fit statistic between the tail model and the truncated empirical data.[5] We use the discrete power-law distribution as our tail model (which implies $x_{\min}$ is also discrete; see Appendix A) and compare its estimates to those made using log-normal and stretched exponential models. To avoid the problem of choosing an appropriate event count distribution, we keep the number of events n fixed.

Finally, using the RAND-MIPT event data (other sources [2] 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.[6] 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.

A. Uncertainty in the scaling parameter

Let $x_{\min} = 10$ be fixed. Figure 1a shows 100 of the fitted bootstrap models, illustrating that by accounting for the uncertainty in α, we obtain an ensemble of tail

[5] Ref. [16] provides a thorough motivation of this strategy. Briefly, the KS statistic will be large either when $x_{\min}$ is too small (including non-power-law data in the power-law fit) or when $x_{\min}$ is too large (when sample size is reduced and legitimately power-law data thrown out), but will be small between these two cases.

[6] Official sources differ slightly on the number killed in New York City. Repeating our analyses with other reported values does not significantly change our estimates.

tail model	parameters	est. $\Pr(x \geq 2749)$ per event, $q(x)$	est. prob. p, 1968–2007	90% CI (bootstrap)
power law (1)	$\Pr(\hat{\alpha})$, $x_{\min} = 10$	0.0000270200	0.299	$[0.203, 0.405]$
power law (2)	$\Pr(\hat{\alpha}, \hat{x}_{\min})$	0.0000346345	0.347	$[0.182, 0.669]$
stretched exp.	$\Pr(\hat{\beta}, \hat{\lambda})$, $x_{\min} = 10$	0.0000156780	0.187	$[0.115, 0.272]$
log-normal	$\Pr(\hat{\mu}, \hat{\sigma})$, $x_{\min} = 10$	0.0000090127	0.112	$[0.063, 0.172]$

TABLE I: Estimated per-event and worldwide historical probabilities for at least one catastrophic event over the period 1968–2007, for four tail models.

models and thus an ensemble of probability estimates for a catastrophic-sized event. The bootstrap parameter distribution $\Pr(\hat{\alpha})$ has a mean $\langle\hat{\alpha}\rangle = 2.40$, which agrees with the maximum likelihood value $\hat{\alpha} = 2.4$ [16].

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 $[0.203, 0.405]$ (Fig. 1*a* inset), 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.

B. 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 $\langle\hat{\alpha}\rangle = 2.40$. However, 15% of models choose $x_{\min} = 4$ or 5, and these produce much heavier-tailed models, with $\langle\hat{\alpha}\rangle = 2.21$.

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 (see Appendix C 3 c). 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 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.

C. Alternative tail models

Comparing these 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

$$\begin{array}{ll} \text{log-normal} & \Pr(x) \propto x^{-1} \exp\left[-(\ln x - \mu)^2 / 2\sigma^2\right] \\ \text{stretched exp.} & \Pr(x) \propto x^{\beta-1} \mathrm{e}^{-\lambda x^{\beta}} \quad , \end{array}$$

where we restrict each to a "tail" domain $x_{\min} \leq x < \infty$ (see Appendix A). In the stretched exponential, $\beta < 1$ produces a heavy-tailed distribution; in the log-normal, small values of μ and large values of σ yield heavy tails. 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.[7]

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 $p = 0.01$. 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.

D. Impact of covariates

Not all large terrorist events are of the same type, and thus our overall estimate is a function of the relative em-

[7] The question of power-law versus non-power-law [16] is not always academic; for instance, macroeconomic financial models have traditionally and erroneously assumed non-power-law tails that assign negligible probability to large events like widespread sub-prime loan defaults [18].

pirical 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 a). This heavier tail more than compensates for their scarcity, as we estimate $\hat{p} = 0.475$ (CI: $[0.309, 0.610]$; Fig. 6*a*) for at least one such catastrophic event from 1968–2007.[8] A similar story emerges for events in economically developed nations, which comprise 5.3% of our data (see Appendix C 3 b). Focusing on large such events ($x \geq 10$), we estimate $\hat{p} = 0.225$ (CI: $[0.037, 0.499]$, Fig. 6*b*).

Another important event covariate is the type of weapon used. The tails of the weapon-specific distributions remain well described as power laws, but weapons like guns, knives and explosives exhibit less heavy tails (fewer large events) than unconventional weapons [33], even as the former are significantly more common than the latter. The estimation algorithm used above can be generalized to handle categorical event covariates, and produces both marginal and total probability estimates (see Appendix C 3 c). Doing so yields an overall estimate of $\hat{p} = 0.564$ (CI: $[0.338, 0.839]$; Fig. 7). Examining the marginal hazard rates, we see that the largest contribution comes from explosives, followed by fire and firearms.

III. STATISTICAL FORECASTS

If the social and political processes that generate terrorist events worldwide are roughly stationary, our algorithm can be used to make principled statistical forecasts about the future probability of a catastrophic event. Although here we make the strong assumption of stationarity, this assumption could be relaxed using non-stationary forecasting techniques [54–56].

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

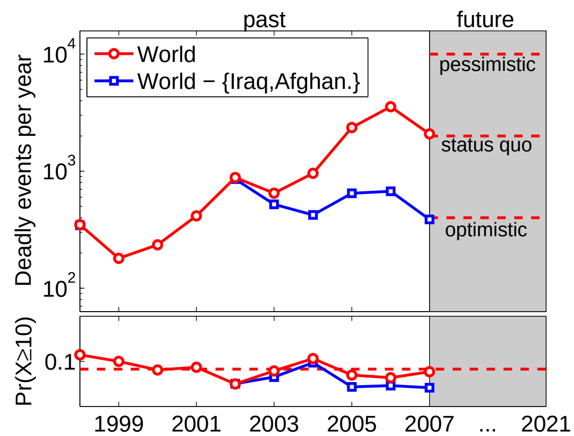

FIG. 3: (upper) Number of deadly (domestic and international) terrorist events worldwide for the ten year period 1998–2007, and three forecast scenarios. (lower) Fraction of events that are severe, killing at least 10 individuals and its 10-year average (dashed line).

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 $\langle p_{\text{tail}} \rangle = 0.082684$ (or about 8.3%) in the former case and 0.072601 (or about 7.3%) in the latter.

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): (i) an optimistic scenario, in which the average number of terrorist attacks worldwide per year returns to its 1998–2002 level, at about $\langle n_{\text{year}} \rangle = 400$ annual events; (ii) a status quo scenario, where it remains at the 2007 level, at about 2000 annual events; and finally (iii) a pessimistic scenario, in which it increases to about 10,000 annual events.[9]

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

[8] The implication of a larger $\hat{p}$ for a covariate distribution, as compared to the full data set, is a smaller p for the excluded types of events. That is, a larger p for international events implies a smaller p for domestic events.

[9] Modeling these rough event counts via a Poisson process with rate $\lambda_{\text{scenario}}$ would refine our forecasts slightly. More detailed event production models could also be used.

tail model	Pr($x \geq 2749$) forecast, 2012-2021 "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

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

n_{tail}. For each scenario, we choose $n_{\text{decade}} = 10 \times n_{\text{year}}$ and choose n_{tail} via Eq. (1) with $p_{\text{tail}} = 0.082684$ (its historical average). Finally, we fix $x_{\min} = 10$ to facilitate comparison with our alternative tail models.

Table II 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, i.e., the more often terrorist attacks are launched, the greater the general likelihood that at least one will be catastrophic. Any progress in moving the general frequency of terrorism toward the more optimistic scenario is likely to reduce the overall, near-term probability of a catastrophic event.

IV. IMPROVED ESTIMATES

Our analysis places the 1968–2007 worldwide historic probability of a catastrophic event in the 11–35% range (see Table I) and none of the alternative or covariate models provide any support for judging the size of 9/11 as statistically unlikely. The wide confidence interval illustrates the difficulty of obtaining precise estimates when accounting for model and parameter uncertainty. That being said, our calculations could be further refined to improve the overall estimates, incorporate additional sources of uncertainty, or address specific questions, by relaxing portions of our iid treatment of event severities. We discuss several such possibilities here, but leave their investigation for the future.

First, our algorithm assumes a stationary event generation process, which is unlikely to be accurate in the long term. Technology, population, culture and geopolitics are believed to exhibit non-stationary dynamics and these likely play some role in event severities. Thus, techniques for statistical forecasting in non-stationary time series [54–56] could be used to identify subtle trends in the relevant covariates to make more accurate forecasts.

Second, our algorithm is silent regarding efforts to prevent events or mitigate their severity [57]. However, the historical impact of these processes is implicitly present in our empirical data because only events that actually occurred were recorded. Thus, our results may be interpreted as probabilities conditioned on historical prevention or mitigation efforts. To the extent that policies have an impact on incidence and severity, more accurate estimates may be achievable by incorporating models of policy consequences or interactions between different actors. Similarly, our algorithm is silent regarding the actors responsible for events, and incorporating models of organizational capabilities, proclivities, etc. [24, 53, 58] may improve the estimates.

Finally, our approach is non-spatial and says little about where the event might occur. Incorporating more fine-grained spatial structure, e.g., to make country-level or theatre-level estimates [59] (as is now being done in seismology [60]), or incorporating tactical information, e.g., about specific CBRN attacks, may be possible. Such refinements will likely require strong assumptions about many context-specific factors [61], and it remains unclear whether accurate estimates at these scales can be made. 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. How far into this chaos we can venture before losing general predictive power remains unclear [13, 62].

V. DISCUSSION

In many complex social systems, although large events have outsized social significance, their rarity makes them difficult to study. Gaining an understanding of such systems requires determining if the same or different processes control the appearance of small, common events versus large, rare events. A critical scientific problem is estimating the true but unknown probability of such large events, and deciding whether they should be classified as statistical outliers. Accurate estimates can facilitate historical analysis, model development and statistical forecasts.

The algorithm described here provides a principled and data-driven solution for this problem that naturally incorporates several sources of uncertainty. Conveniently, the method captures the tendency of highly-skewed distributions to produce large events without reference to particular generative mechanisms or strong assumptions about the tail's structure. When properly applied, it provides an objective estimate of the historical or future probability of a rare event, e.g., an event that has occurred exactly once.

Using this algorithm to test whether the size of the 9/11 terrorist events, which were nearly six times larger than the next largest event, could be an outlier, we estimated the historical probability of observing at least one 9/11-sized event somewhere in the world over the past 40 years to be 11–35%, depending on the particular choice of tail model used to fit the distribution's upper tail. These values are much larger than any reasonable definition of a statistical anomaly and thus the size of 9/11, which was nearly six times larger than the next largest event, should not be considered statistically unlikely, given the historical record of events of all sizes.

This conclusion is highly robust. Conditioning on the relative frequency of important covariates [33], such as the degree of economic development in the target country, whether an event is domestic or international, or the type of weapon used, we recover similar estimates, with additional nuance. Large events are probabilistically most likely to target economically developed nations, be international in character and use explosives, arson, firearms or unconventional weapons. Although chemical and biological events can also be very large [8], historically they are rare at all sizes, and this outweighs the heaviness of their tail.

Furthermore, using only event data prior to 9/11 (as opposed to using all available data sans 9/11), we find a similar overall historical hazard rate. This suggests that the worldwide probability for large events has not changed dramatically over the past few decades. In considering three simple forecast scenarios for the next 10 years, we find that the probability of another large event is comparable to its historical level over the past 40 years. This risk seems unlikely to decrease significantly without a large reduction in the number of deadly terrorist events worldwide.

Of course, all such estimates are only as accurate as their underlying assumptions, and our method treats event sizes as iid random variables drawn from a stationary distribution. For complex social phenomena in general, it would be foolish to believe this assumption holds in a very strong sense, e.g., at the micro-level or over extremely long time scales, and deviations will lower the method's overall accuracy. For instance, non-stationary processes may lower the global rate of large events faster than smaller events, leading to overestimates in the true probability of a large event. However, the iid assumption appears to be statistically justified at the global spatial and long-term temporal scales studied here. Identifying the causes of this apparent iid behavior at the global scale is an interesting avenue for future work.

The relatively high probability of a 9/11-sized event, both historically and in the future, suggests that the global political and social processes that generate large terrorist events may not be fundamentally different from those that generate smaller, more common events. Although the mechanism for event severities remains unclear [63], the field of possible explanations should likely be narrowed to those that generate events of all sizes.

Independent of mechanistic questions, the global probability of another large terrorist event remains uncomfortably high, a fact that can inform our expectations (as with large natural disasters [34–36]) of how many such events will occur over a long time horizon and how to appropriately anticipate or respond to them. This perspective is particularly relevant for terrorism, as classical models aimed at predicting event incidence tend to dramatically underestimate event severity [33].

To conclude, the heavy-tailed patterns observed in the frequency of severe terrorist events suggests that some aspects of this phenomenon, and possibly of other complex social phenomena, are not nearly as contingent or unpredictable as is often assumed. That is, there may be global political and social processes that can be effectively described without detailed reference to local conflict dynamics or the strategic tradeoffs among costs, benefits and preferences of individual actors. Investigating these global patterns offers a complementary approach to the traditional rational-actor framework [12] and a new way to understand what regularities exist, why they exist, and their implications for long-term stability.

Acknowledgments

The authors thank Cosma Shalizi, Nils Weidmann, Julian Wücherpfennig, Kristian Skrede Gleditsch, Victor Asal, Didier Sornette, Lars-Erik Cederman and Patrick Meier for helpful conversations. Implementations of our numerical methods are available online at `http://www.santafe.edu/~aaronc/rareevents/`.

Appendix A: Tail models

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

$$\Pr(y \mid \alpha, x_{\min}) = \left(\frac{\alpha-1}{x_{\min}}\right)\left(\frac{y}{x_{\min}}\right)^{-\alpha} \quad , \qquad \text{(A1)}$$

with $\alpha > 1$ and $y \geq x_{\min} > 0$. Given a choice of $x_{\min}$, the maximum likelihood estimator for this model is

$$\hat{\alpha} = 1 + n \Bigg/ \sum_{i=1}^{n} \ln(x_i/x_{\min}) \quad . \qquad \text{(A2)}$$

The severity of a terrorist attack, however, is given by an integer. Thus, in our analysis of terrorist event severities, we use the discrete form of the power-law distribution

$$\Pr(y \mid \alpha, x_{\min}) = y^{-\alpha} / \zeta(\alpha, x_{\min}) \quad , \qquad \text{(A3)}$$

with $\alpha > 1$ and $y \geq x_{\min} > 0$, and where $\zeta(\hat{\alpha}, x_{\min}) = \sum_{i=x_{\min}}^{\infty} i^{-\alpha}$ is the generalized or incomplete zeta function. The MLE in for the discrete power law is less straightforward, being the solution to the transcendental equation

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n}\sum_{i=1}^{n} x_i \quad . \qquad \text{(A4)}$$

However, it is straightforward to directly maximize the log-likelihood function for the discrete power law in order to obtain $\hat{\alpha}$:

$$\mathcal{L}(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i \ . \tag{A5}$$

Past work shows that the continuous model given by Eq. (A3) provides a reasonably good approximation to the discrete case when $x_{\min}$ takes moderate values. In our own experiments with this approximation, we find that when $x_{\min} \gtrsim 10$ the difference in estimated probabilities for observing one or more 9/11-sized events between using the discrete versus continuous model is at most few percent.

Estimates of $x_{\min}$ may be obtained using any of several existing automatic methods [37, 43–46]. We use the Kolmogorov-Smirnov goodness-of-fit statistic minimization (KS-minimization) technique [16, 33]. This method falls in the general class of distance minimization techniques for selecting the size of the tail [7], and was previously used to analyzing event severities in global terrorism. The KS statistic [64] is the maximum distance between the CDFs of the data and the fitted model:

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \ , \tag{A6}$$

where $S(x)$ is the CDF of the data for the observations with value at least $x_{\min}$, 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 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.

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 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 iid 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 $n-1$ iid 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 $p = 0.001$, i.e., we contaminate the dataset with a genuine outlier.

Figure 4 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. Even for samples as small as $n = 40$ observations, the absolute error is fairly small and decreases with increasing sample size n. In the first experiment, the error rate decays like $O(n^{-1/3})$, approaching 0.01 error rates as n approaches 5000 (Fig. 4a), while in the second it decays like $O(n^{-1})$ up to about $n = 4000$, above which the rate of decay attenuates slightly (Fig. 4b).

Absolute deviations may mask dramatic relative errors, e.g., if the true probability is very close to one or zero (as in our contaminated samples experiment). The mean ratio of $\hat{p}$ to p would reveal such mistakes. The lower panels in Fig. 4 show that this is not the case: the estimation procedure is close both in absolute and in relative terms. As the sample size increases, the estimated probability converges on the true probability. For contaminated data sets, the $\hat{p}/p$ can be fairly large when n is very small, but for sample sizes of a few hundred observations, the method correctly estimates the relative size of the outlier's probability.

Appendix C: Robustness checks

We present three checks of the robustness of our probability estimates, (i) using simple parametric models without the bootstrap, (ii) using an alternative source of terrorist event data, and (iii) using event covariates to refine the estimates. In each case, we find roughly similar-sized estimates.

1. Estimates using simple models

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

A version of this model was used in a 2009 Department of Defense-commissioned JASON report on "rare events" [5], which estimated the historical probability of a catastrophic (9/11-sized or larger) terrorist event as 23% over 1968–2006. The report used a slightly erroneous estimate of the power law's normalizing constant,

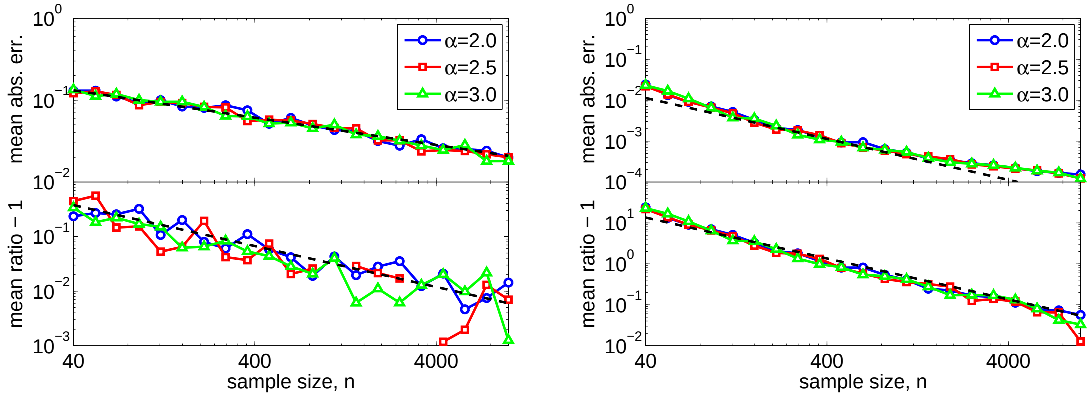

FIG. 4: The mean absolute error $\langle|\hat{p}-p|\rangle$ and mean relative error $\langle\hat{p}/p\rangle-1$ for (a) n values drawn iid from a stationary power-law distribution with $x_{\min}=10$ and some α, with the target size being the single largest value in the draw, and for (b) $n-1$ values 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.

a slightly different estimate of α and a smaller value of n. Here, we repeat the JASON analysis, but with more accurate input values.

Let $q(x)$ be the probability of observing a catastrophic event of size x. With event severities being iid random variables drawn from a fixed distribution $\Pr(y)$, the generation of catastrophic events can be described by a continuous-time Poisson process with rate $q(x)$ [65]. Approximating x as a continuous variable, in a sequence of n such events, the probability $\hat{p}$ of observing at least one of catastrophic severity is

$$\begin{aligned}\hat{p} &= 1-[1-q(x)]^n \\ &\approx 1-\mathrm{e}^{-n\,q(x)} \quad . \end{aligned} \tag{C1}$$

The rate $q(x)$ is simply the value of the complementary CDF at x, and for a power-law distribution is given by

$$\begin{aligned} q(x) &= \int_x^\infty \Pr(y)\mathrm{y} \\ &= (\alpha-1)x_{\min}^{\alpha-1}\int_x^\infty y^{-\alpha}\mathrm{y} \\ &= \left(\frac{x}{x_{\min}}\right)^{1-\alpha} \quad , \end{aligned} \tag{C2}$$

for $x \geq x_{\min}$. Substituting $x_{\min}=1$ and $\alpha=2.4$ yields the per-event probability of a catastrophic event $q(2749)=0.0000153164$.

The RAND-MIPT database records $n=13274$ deadly events worldwide from 1968–2007; thus, substituting n and $q(x)$ into Eq. (C1) yields a simple estimate of the probability of observing at least one catastrophic event over the same time period $\hat{p}=1-\mathrm{e}^{-13274\,q(2749)}=0.184$, or about 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 Eq. (C2) using $x=10$ shows that our model predicts that only 4.0% of events should be this severe. Thus, events with $x\geq 10$ 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,[10] a value more in line with the estimates derived using our bootstrap-based approach in the main text.

2. Estimates using the Global Terrorism Database

An alternative source of global terrorism event data is the Global Terrorism Database [2], 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.

[10] To make our reported per-event probabilities $q(x)$ consistent across models, we report them as $q(x)=\Pr(X\geq x\,|\,X\geq x_{\min})\Pr(X\geq x_{\min})$, i.e., the probability that a tail event is catastrophic times the probability that the event is a tail event. These values can be used with Eq. (C1) to make rough estimates if the corresponding n is the total number of deadly events.

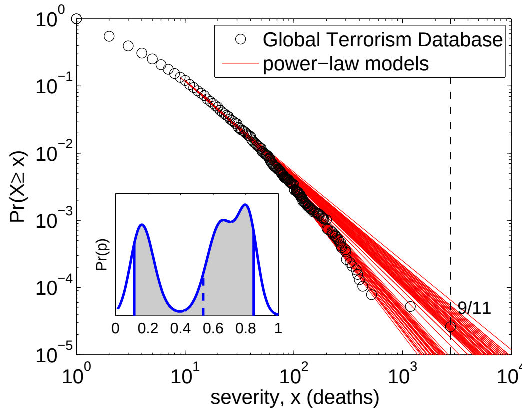

FIG. 5: Empirical distribution of event severities from the GTD [2] with 100 power-law models, fitted to bootstraps of the data. Inset shows the estimated distribution of binomial probabilities $\Pr(\hat{p})$ for one or more catastrophic events.

The largest event in the GTD is 9/11, with severity 2763, and the second largest is the 13 April 1994 Rwandan massacre of Tutsi refugees, with 1180 reported fatalities. This event is absent from the RAND-MIPT data; its inclusion in the GTD highlights this data set's broader definition of terrorism, which includes a number of genocidal or criminal events.

The best fitting power-law model obtained using the methodology of [16] 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 dataset, e.g., 10, 20, 100, 200, etc., which indicates rounding effects in the reporting. (These appear in Fig. 5 as small discontinuous drops in the complementary CDF at round-number locations; true power-law distributed data have no preference for round numbers and thus their presence is a statistically significant deviation from the power-law form.)

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.

3. Impact of event covariates

a. International versus domestic, and events prior to 1998

Events in the RAND-MIPT database with dates before 1 January 1998 are mainly international events, i.e., the attacker's country of origin differed from the target's country. Subsequent to this date, both domestic and international events are included but their domestic versus international character is not indicated. Analyzing events that occurred before this breakpoint thus provides a natural robustness check for our overall estimate: (i) we can characterize the effect that domestic versus international events have on the overall estimate and (ii) we can test whether the probability estimates have changed significantly in the past decade.

The pre-1998 events comprise 12% of the RAND-MIPT database and exhibit a more heavy-tailed distribution ($\hat{\alpha} = 1.92 \pm 0.04$ and $x_{\min} = 1$). Using 10,000 bootstraps, we estimate $\hat{p} = 0.475$ (90% CI: $[0.309, 0.610]$) for at least one catastrophic international event over the target period. Figure 6a shows the empirical distribution for international events and the ensemble of fitted models, illustrating good visual agreement with the empirical distribution.

The estimate for international-only data ($\hat{p} = 0.475$) is larger than the estimate derived using the full data set ($\hat{p} = 0.347$), although these values may not be as different as they seem, due to their overlapping confidence intervals. Fundamentally, the larger estimate is caused by the heavier-tailed distribution of the international-only data. Because the full data set includes these international events, this result indicates that domestic events tend to exhibit a lighter tail, and thus generate large terrorist events with smaller probability. As a general guideline, subsets of the full data set should be analyzed with caution, as their selection is necessarily conditioned. The full data set provides the best estimate of large events of all types.

b. Economic development

A similar story emerges for deadly events in 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. The empirical distribution (Fig. 6b) of event severities shows unusual structure, with the upper tail ($x \geq 10$ fatalities) decaying more slowly than 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 6b shows the resulting ensemble of fitted models, il-

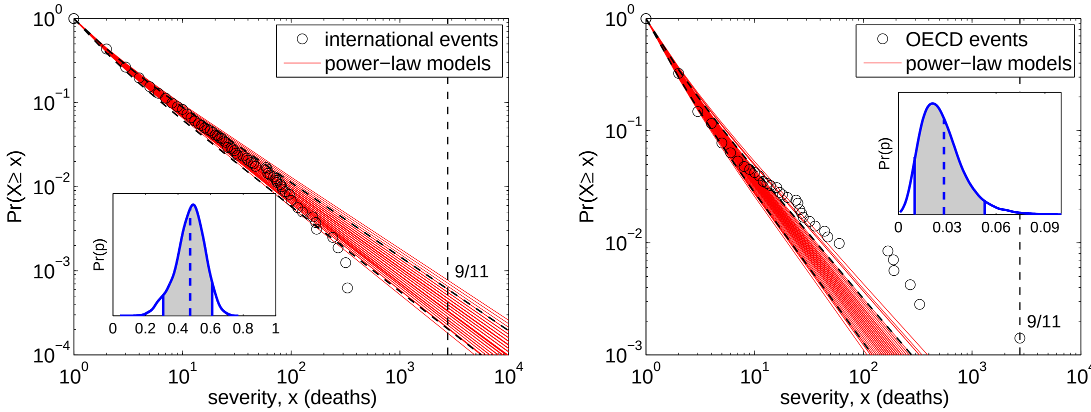

FIG. 6: Empirical distributions, with 100 power-law bootstrap models, for (a) international events (events from 1968–1997 in the RAND-MIPT database) and (b) events within the OECD nations; dashed lines show the 90% CI on $\hat{\alpha}$. Insets show the estimated distribution $\Pr(\hat{p})$ with 90% confidence intervals (shaded area) and overall estimate (dashed line).

lustrating 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 yields $\hat{p} = 0.225$, with 90% CI as $[0.037, 0.499]$.

c. *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 ith event. Thus, denote the empirical data by $X = \{(x, c)_i\}$.

1. Generate Y by drawing $(y, c)_j$, $j = 1, \ldots, n$, uniformly at random, with replacement, from the original data $\{(x, c)_i\}$ (sans the ℓ catastrophic events).

2. For each covariate type c in Y, jointly estimate $\hat{x}^{(c)}_{\min}$ and the tail-model parameters $\theta^{(c)}$, and compute $n^{(c)}_{\text{tail}} = \#\{y_j \geq \hat{x}^{(c)}_{\min}\}$.

3. For each covariate type c in Y, generate a synthetic data set by drawing $n^{(c)}_{\text{tail}}$ 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 ℓ events of size x or greater, set $\rho = 1$; otherwise, set it to zero.

In applying this algorithm to our data, we choose $\ell = 1$ and $x = 2749$, as with our other analyses. In step 2, we again use the KS-minimization technique of [16] 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 [41], 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, drawing 1000 synthetic data sets from each, 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. Figures 7a-f show the ensembles for each weapon-specific severity distribution. 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. The following table gives these marginal probability estimates, which are greatest for explosives, fire, firearms and unconventional weapon types.

It is emphasized that these are historical estimates, based on the relative frequencies of weapon covariates in the historical RAND-MIPT data. If the future exhibits similar relative frequencies and total number of attacks, then they may also be interpreted as future hazards, but we urge strong caution in making these assumptions.

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.

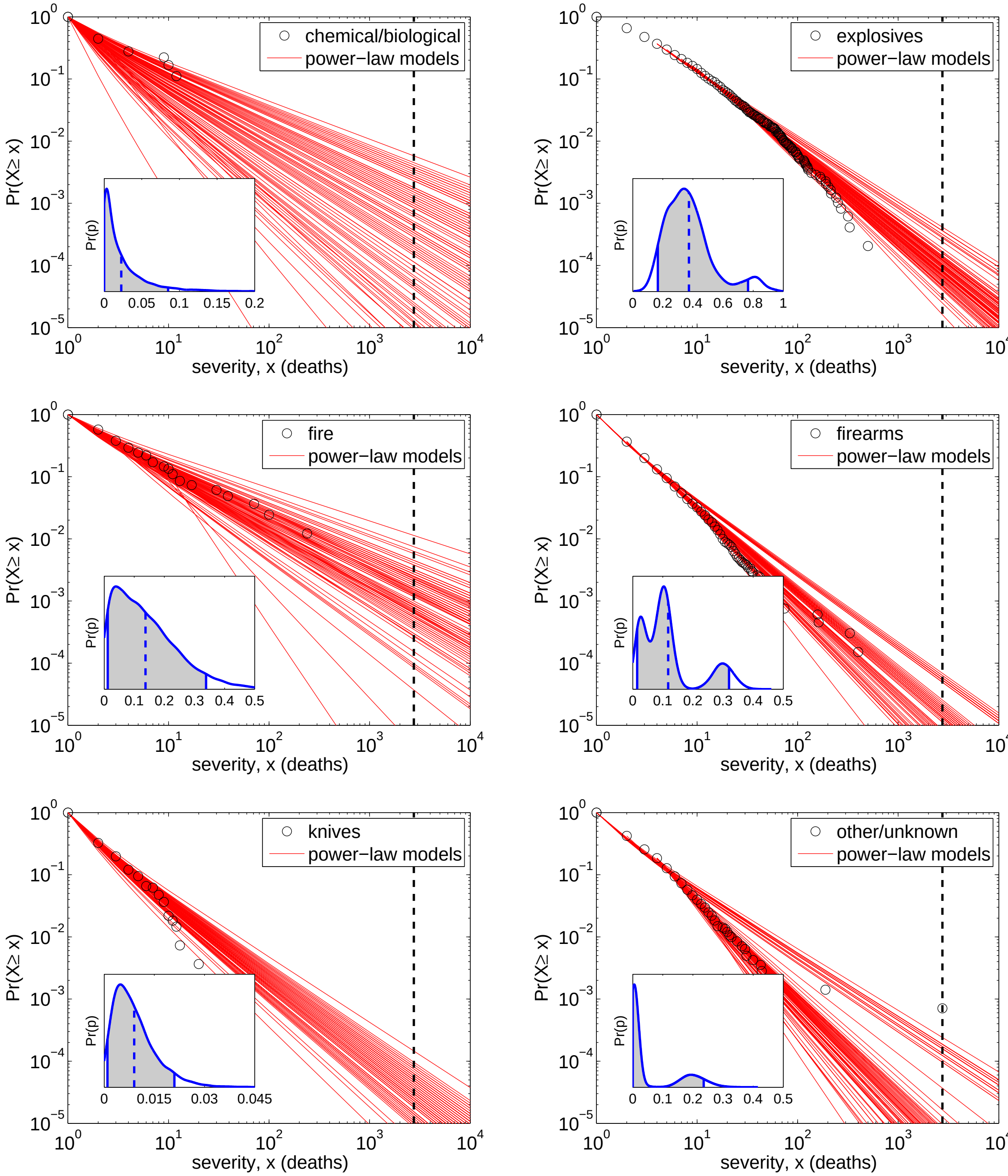

FIG. 7: Empirical distribution, with 100 power-law bootstrap models, for events using (a) chemical or biological, (b) explosives (includes remote detonation), (c) fire, arson and firebombs, (d) firearms, (e) knives or sharp objects, and (f) other, unknown or unconventional weapons. Insets: marginal distributions of estimated hazard rates $\Pr(\hat{p})$, with the region of 90% confidence shaded and the mean value indicated by the dashed line.

[1] MIPT (2008), www.tkb.org (accessed 29 January 2008).
[2] National Consortium for the Study of Terrorism and Responses to Terrorism (START) (2011), www.start.umd.edu/gtd/ (access date October 31, 2011).
[3] D. Sornette, International Journal of Terraspace Science and Engineering **2**, 1 (2009).
[4] Sornette, D. and Ouillon, G., eds., *Eur. Phys. J.*

Special Topics—Dragon-kings: Mechanisms, statistical methods and empirical evidence, vol. 205 (2012), URL http://dx.doi.org/10.1140/epjst/e2012-01559-5.
[5] D. McMorrow, *Rare Events* (The MITRE Corporation, McLean, VA, 2009), JASON Report JSR-09-108.
[6] L. de Hann and A. Ferreira, *Extreme Value Theory: An Introduction* (Springer-Verlag, New York, 2006).
[7] R.-D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields* (Birkhäuser, Basel, Switzerland, 2007).
[8] G. Cameron, Nonproliferation Review **7**, 162 (2000).
[9] M. P. Drennan, in *The Economic Costs and Consequences of Terrorism*, edited by H. W. Richardson, P. Gordon, and J. E. Moore (Edward Elgar Publishing, Northampton, MA, 2007).
[10] G. King and L. Zeng, International Organization **55**, 693 (2001).
[11] N. Beck, G. King, and L. Zeng, American Political Science Review **94**, 21 (2000).
[12] B. Bueno de Mesquita, Rationality and Society **15**, 136 (2003).
[13] M. D. Ward, B. D. Greenhill, and K. M. Bakke, Journal of Peace Research **47**, 363 (2010).
[14] S. C. A. Rustad, H. Buhaug, Å. Falch, and S. Gates, Conflict Management and Peace Science **28**, 14 (2011).
[15] B. Bueno de Mesquita, Conflict Management and Peace Science **28**, 64 (2011).
[16] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Review **51**, 661 (2009).
[17] J. D. Farmer and F. Lillo, Quantitative Finance **4**, 7 (2004).
[18] Financial Crisis Inquiry Commission, *Financial Crisis Inquiry Report* (U.S. Government Printing Office, Washington, 2011), http://bit.ly/hEsCwo.
[19] W. Enders and T. Sandler, Journal of Conflict Resolution **44**, 307 (2000).
[20] W. Enders and T. Sandler, International Studies Quarterly **46**, 145 (2002).
[21] D. Brown, J. Dalton, and H. Hoyle, Lecture Notes in Computer Science **3073**, 426 (2004).
[22] W. E. Enders and T. Sandler, *The Political Economy of Terrorism* (Cambridge University Press, Cambridge, 2006).
[23] M. L. Valenzuela, C. Feng, P. Reddy, F. Momen, J. W. Rozenblit, B. T. Eyck, and F. Szidarovszky, in *Proc. 17th IEEE International Conference and Workshops on the Engineering of Computer-Based Systems* (2010), pp. 311–315.
[24] V. Asal and R. K. Rethemeyer, Journal of Politics **70**, 437 (2008).
[25] W. Enders, T. Sandler, and K. Gaibulloev, Journal of Peace Research **48**, 319 (2011).
[26] W. A. Wulf, Y. Y. Haimes, and T. A. Longstaff, Risk Analysis **23**, 429 (2003).
[27] Q. Li, Journal of Conflict Resolution **49**, 278 (2005).
[28] R. A. Pape, American Political Science Review **97**, 343 (2003).
[29] M. Sageman, *Understanding Terror Networks* (University of Pennsylvania Press, Philadelpha, 2004).
[30] B. A. Desmarais and S. J. Cranmer, in *Proc. European Intelligence and Security Informatics Conference* (2011), pp. 171–177.
[31] J. A. Major, Journal of Risk Finance **4**, 15 (1993).
[32] E. Kardes and R. Hall, *Survey of Literature on Strategic Decision Making in the Presence of Adversaries* (Center for Risk and Economic Analysis of Terrorism Events, Los Angeles CA, 2005).
[33] A. Clauset, M. Young, and K. S. Gleditsch, Journal of Conflict Resolution **51**, 58 (2007).
[34] B. Gutenberg and C. F. Richter, Bulletin of the Seismological Society of America **34**, 185 (1944).
[35] E. J. Gumbel, The Annals of Mathematical Statistics **12**, 163 (1941).
[36] W. J. Reed and K. S. McKelvey, Ecological Modelling **150**, 239 (2002).
[37] L. Breiman, C. J. Stone, and C. Kooperberg, J. Statistical Computation and Simulation **37**, 127 (1990).
[38] R. J. Adler, R. E. Feldman, and M. S. Tuqqu, eds., *A Practical Guide to Heavy Tails: Statistical Techniques and Applications* (Birkhäuser, Boston, 1998).
[39] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling* (Springer-Verlag, New York, 2006).
[40] G. Claeskens and N. L. Hjort, *Model Selection and Model Averaging* (Cambridge University Press, Cambridge, 2008).
[41] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, NY, 1993).
[42] G. Last and A. Brandt, *Marked Point Processes on the Real Line* (Springer, Berlin, 1995).
[43] J. Danielsson, L. de Hann, L. Peng, and C. G. de Vries, Journal of Multivariate Analysis **76**, 226 (2001).
[44] A. L. M. Dekkers and L. de Hann, Journal of Multivariate Analysis **47**, 173 (1993).
[45] H. Drees and E. Kaufmann, Stochastic Processes and their Applications **75**, 149 (1998).
[46] M. S. Hancock and J. H. Jones, Theoretical Population Biology **65**, 413 (2004).
[47] R. E. Kass and A. E. Raftery, J. Amer. Statist. Assoc. **90**, 773 (1995).
[48] M. Stone, J. Royal Statistical Society Series B **36**, 111 (1974).
[49] P. D. Grünwald, *The Minimum Length Description Principle* (MIT Press, Cambridge, MA, 2007).
[50] Q. H. Vuong, Econometrica **57**, 307 (1989).
[51] N. L. Hjort and G. Claeskens, J. American Stat. Assoc. **98**, 879 (2003).
[52] N. Cesa-Bianchi, A. Conconi, and C. Gentile, IEEE Transactions on Information Theory **50**, 2050 (2004).
[53] A. Clauset and K. S. Gleditsch (2009), E-print, http://arxiv.org/abs/0906.3287.
[54] S. Caires and J. A. Ferreira, Statistical Inference for Stochastic Processes **8**, 151 (2005), correction, vol. 9 (2006), pp. 109–110.
[55] M. P. Clements and D. F. Hendry, *Forecasting Non-stationary Economic Time Series* (MIT Press, Cambridge, Massachusetts, 1999).
[56] C. R. Shalizi, A. Z. Jacobs, K. L. Klinkner, and A. Clauset (2011), E-print http://arxiv.org/abs/1103.0949.
[57] D. Kilcullen, *Counterinsurgency* (Oxford University Press, Oxford, 2010).
[58] B. A. Jackson, J. C. Baker, K. Cragin, J. Parachini, H. R. Trujillo, and P. Chalk, *Aptitude for Destruction: Organizational Learning in Terrorist Groups and Its Implications for Combating Terrorism*, vol. 1 (RAND Corporation, Arlington, 2005).
[59] A. Zammit-Mangion, M. Dewar, V. Kadirkamanathan,

and G. Sanguinetti, Proc. Nat. Acad. Sci. USA **Early Edition**, 1 (2012).

[60] Y.-T. Lee, D. L. Turcotte, J. R. Holliday, M. K. Sachs, J. B. Rundle, C.-C. Chen, and K. F. Tiampo, Proc. Natl. Acad. Sci. USA **108**, 16533 (2011).

[61] E. Gartzke, International Organization **53**, 567 (1999).

[62] J. B. Rundle, J. R. Holliday, M. Yoder, M. K. Sachs, A. Donnellan, D. L. Turcotte, K. F. Tiampo, W. Klein, and L. H. Kellogg, Geophysical Journal International **187**, 225 (2011).

[63] A. Clauset, M. Young, and K. S. Gleditsch, Peace Economics, Peace Science and Public Policy **16**, article 12 (2010).

[64] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, UK, 1992).

[65] M. L. Boas, *Mathematical Models in the Physical Sciences* (Wiley, Hoboken, NJ, 2005).