Submitted to the Annals of Applied Statistics

POWER-LAW DISTRIBUTIONS IN BINNED EMPIRICAL DATA

By Yogesh Virkar[†] and Aaron Clauset[†,‡,*]

University of Colorado at Boulder[†]
Santa Fe Institute[‡]

Many man-made and natural phenomena, including the intensity of earthquakes, population of cities, and size of international wars, are believed to follow power-law distributions. The accurate identification of power-law patterns has significant consequences for developing an understanding of complex systems. However, statistical evidence for or against the power-law hypothesis is complicated by large fluctuations in the empirical distribution's tail, and these are worsened when information is lost from binning the data. We adapt the statistically principled framework for testing the power-law hypothesis, developed by Clauset, Shalizi and Newman, to the case of binned data. This approach includes maximum-likelihood fitting, a hypothesis test based on the Kolmogorov-Smirnov goodness-of-fit statistic and likelihood ratio tests for comparing against alternative explanations. We evaluate the effectiveness of these methods on synthetic binned data with known structure and apply them to twelve real-world binned data sets with heavy-tailed patterns.

1. Introduction. Power-law distributions have attracted broad scientific interest [36] both for their mathematical properties, which sometimes lead to surprising consequences, and for their appearance in a wide range of natural and man-made phenomena, spanning physics, chemistry, biology, computer science, economics and the social sciences [21, 23, 33, 13].

Quantities that follow a power-law distribution are sometimes said to exhibit "scale invariance", indicating that common, small events are not qualitatively distinct from rare, large events. Identifying this pattern in empirical data can indicate the presence of unusual underlying or endogenous processes, e.g., feedback loops, network effects, self-organization or optimization, although not always [29]. Knowing that a quantity does or does not follow a power law provides important theoretical clues about the underlying generative mechanisms we should consider. It can also facilitate statistical extrapolations about the likelihood of very large events [7].

*To whom correspondence should be addressed.

AMS 2000 subject classifications: Primary 62G32, 62-07; secondary 65C05
Keywords and phrases: power-law distribution, heavy-tailed distributions, model selection, binned data

The task of deciding if some quantity does or does not plausibly follow a power law is complicated by the existence of large fluctuations in the empirical distribution's upper tail, precisely where we wish to have the most accuracy. These fluctuations are amplified when the empirical data are binned, i.e., converted into a series of counts over a set of non-overlapping ranges in event size. As a result, the upper tail's true shape is often obscured and we may be unable to distinguish a power-law pattern from alternative heavy-tailed distributions like the stretched exponential or the log-normal. Here, we present a set of principled statistical methods, adapted from [6], for answering these questions when the data are binned.

Statistically, power-law distributions generate large events orders of magnitude more often than would be expected under a Normal distribution, and thus such quantities are not well-characterized by quoting a typical or average value. For instance, the 2000 U.S. Census indicates that the average population of a city, town or village in the United States contains 8226 individuals, but this value gives no indication that a significant fraction of the U.S. population lives in cities like New York and Los Angeles, whose populations are roughly 1000 times larger than the average. Extensive discussions of this and other properties of power laws can be found in reviews by Mitzenmacher [21], Newman [23], Sornette [33] and Gabaix [13].

Mathematically, a quantity x obeys a power law if it is drawn from a probability distribution with a form

$$\Pr(x) \propto x^{-\alpha} \quad , \tag{1.1}$$

where $\alpha > 1$ is the *exponent* or *scaling parameter* and $x > 0$. In practice, few empirical phenomena obey power laws for all values of x. More often, the power-law pattern holds only above some value $x_{\min}$, in which case we say that the *tail* of the distribution follows a power law.

Recently, Clauset, Shalizi and Newman [6] introduced a set of statistically principled methods for fitting and testing the power-law hypothesis for continuous or discrete-valued data. Their approach combines maximum-likelihood techniques for fitting a power-law model to the distribution's upper tail, a distance-based method [30] for automatically identifying the point $x_{\min}$ above which the power-law behavior holds [12], a goodness-of-fit test based on the Kolmogorov-Smirnov (KS) statistic for characterizing the fitted model's statistical plausibility, and a likelihood ratio test [37] for comparing it to alternative heavy-tailed distributions.

Here, we adapt these methods to the less common but important case of binned empirical data, i.e., when we see only the frequency of events within a set of non-overlapping ranges. Our goal is not to provide an exhaustive evaluation of all possible principled approaches to considering power-law distributions in binned empirical data, but rather the more narrow aim of adapting the popular framework of [6] to binned data. Such data often occur when direct measurements are impractical or impossible and only the order of magnitude is known, or when we recover measurements from an existing histogram. Sometimes, when the original measurements are unavailable, this is simply the form of the data we receive and despite the loss of information due to binning, we would still like to make strong statistical inferences about power-law distributions. This requires specialized tools not currently available.

Toward this end, we present maximum-likelihood techniques for fitting the power-law model to binned data, for identifying the smallest bin $b_{\min}$ for which the power-law behavior holds, for testing its statistical plausibility, and for comparing it with alternative distributions.[1] These methods make no assumptions about the type of binning scheme used, and can thus be applied to linear, logarithmic or arbitrary bins. We evaluate the effectiveness of our techniques on synthetic data with known structure, showing that they are highly accurate when given a sample of sufficient size. Their effectiveness does depend on the amount of information lost due to binning, and we quantify this loss of accuracy and statistical power in several ways.

Following [6], we advocate the following recipe for investigating the power-law hypothesis in binned empirical data.

1. *Fit the power law.* Section 3. Estimate the parameters $b_{\min}$ and α of the power-law model.
2. *Test the power law's plausibility.* Section 4. Conduct a hypothesis test for the fitted model. If $p \geq 0.1$, the power-law is a plausible statistical hypothesis for the data; otherwise, it is rejected.
3. *Compare against alternative distributions.* Section 5. Compare the power law to alternative heavy-tailed distributions via a likelihood ratio test. For each alternative, if the log-likelihood ratio is significantly away from zero, then its sign indicates whether or not the alternative is favored over the power-law model.

The model comparison step could be replaced with another statistically principled approach for model comparison, e.g., fully Bayesian, cross-validation or minimum description length. We do not describe these techniques here.

Practicing what we advocate, we then apply our methods to 12 real-world data sets, all of which exhibit heavy-tailed, possibly power-law behavior. Many of these data sets were obtained in their binned form. We also include a

[1] Our code is available at `http://www.santafe.edu/~aaronc/powerlaws/bins/`

few examples from [6] to demonstrate consistency with their results. Finally, to highlight the concordance of our binned methods with the continuous or discrete-valued methods of [6], we organize our paper in a similar way.

2. Binned Power-Law Distributions. Conventionally, a power-law distributed quantity can be either continuous or discrete. For continuous values, the probability density function (pdf) of a power law is defined as

$$p(x)\,\mathrm{d}x = \Pr(x \le X < x + \mathrm{d}x) = C\,x^{-\alpha}\,\mathrm{d}x \ , \tag{2.1}$$

where X is the observed value and C is the normalization constant. This density diverges as $x \to 0$ and so Equation (2.1) cannot hold for all $x \ge 0$. Instead, there must exist some lower-bound to the power-law behavior, which we denote $x_{\min}$. In this case, so long as $\alpha > 1$, it is easy to calculate the normalizing constant, yielding

$$\Pr(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha} \ . \tag{2.2}$$

For discrete values, the probability mass function is defined as

$$\Pr(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})} \ , \tag{2.3}$$

where $\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$ is the generalized or Hurwitz zeta function, and serves as the normalization constant.

Because formulae for continuous distributions, like Eq. (2.2), tend to be simpler than those for discrete distributions, which often involve special functions, in the remainder of the paper, we present analysis only of the continuous case. The methods, however, are entirely general and can easily be adapted to the discrete case.

A binned data set is sequence of counts of observations over a set of non-overlapping ranges. Let $\{x_i\}$ denote our N original empirical observations. After binning, we discard these observations and retain only the given ranges or bin boundaries B and the counts within them H. Letting k be the number of bins, the bin boundaries B are denoted

$$B = (b_1, b_2, \ldots, b_k) \ , \tag{2.4}$$

where $b_1 > 0$, $k > 1$, the i^{th} bin covers the interval $x \in [b_i, b_{i+1})$, and by convention we assume the k^{th} bin extends to $+\infty$. The bin counts H are denoted

$$H = (h_1, h_2, \ldots, h_k) \ , \tag{2.5}$$

where $h_i = \#\{b_i \leq x_i < b_{i+1}\}$ counts the number of raw observations in the i^{th} bin.

The probability that some observation falls within the i^{th} bin is the fraction of total density in the corresponding interval:

$$\begin{aligned} \Pr(b_i < x < b_{i+1}) &= \int_{b_i}^{b_{i+1}} p(x)\,\mathrm{d}x \\ &= x_{\min}^{\alpha-1}\left[{b_i}^{1-\alpha} - {b_{i+1}}^{1-\alpha}\right] \quad . \end{aligned} \tag{2.6}$$

Subsequently, we assume that the binning scheme B is fixed by an external source, as otherwise we would have access to the raw data and we could apply direct methods to test the power-law hypothesis [6].

To test the power-law hypothesis using binned data, we must first estimate the scaling exponent α, which requires choosing the smallest bin for which the power law holds, which must be a member of the sequence B, i.e., it must be a bin boundary. We denote this choice $b_{\min}$ in order to distinguish it from $x_{\min}$.

3. Fitting Power Laws to Binned Empirical Data.

Many studies of empirical distributions and power laws use poor statistical methods for this task. The most common approach is to first tabulate the histogram and then fit a regression line to the log-frequencies. Taking the logarithm of both sides of Equation (2.1), we see that the power-law distribution obeys the relation $\ln p(x) = \ln C - \alpha \ln x$, implying that it follows a straight line on a doubly logarithmic plot. Fitting such a straight line may seem like a reasonable approach to estimate the scaling parameter α, perhaps especially in the case of binned data where binning will tend to smooth out some of the sampling fluctuations in the upper tail. Indeed, this procedure has a long history, being used by Pareto in the analysis of wealth distributions in the late 19th century [1], by Richardson in analyzing the size of wars in the early 20th century [31], and by many researchers since.

This method and its variations, however, generate significant errors under relatively common conditions and give no warning of their mistakes, and their results should not be trusted (see [6] for a detailed explanation). In this section, we describe a generally accurate method for fitting a power-law distribution to binned data, based on maximum likelihood. Using synthetically generated binned data, we illustrate its accuracy and the inaccuracy of the linear regression approach.

3.1. *Estimating the Scaling Parameter.*

First, we consider the task of estimating the scaling parameter α. Correctly estimating α requires a good choice for the lower bound $b_{\min}$, but for now we will assume that this value is known. In cases where it is not known, we may estimate it using the methods given in Section 3.3.

The chosen method for fitting parameterized models to empirical data is the *method of maximum likelihood*, which provably gives accurate parameter estimates in the limit of large sample size [3, 38]. Specifically, it can be shown that the maximum likelihood estimate $\hat{\theta}$ is asymptotically consistent, i.e., in the limit of large sample size $n \to \infty$, the estimate converges on the truth $\hat{\theta} \to \theta$. Details of our derivations are given in Appendix A. In this section, we focus on the resulting formula's use. Here and elsewhere, we use "hatted" symbols to denote estimates derived from data; hatless symbols denote the true values, which are typically unknown.

Assuming that our observations are drawn from a power-law distribution above the value $b_{\min}$, the log-likelihood function is

$$\mathcal{L} = n(\alpha - 1)\ln b_{\min} + \sum_{i=\min}^{k} h_i \ln\left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)}\right] \ , \tag{3.1}$$

where $n = \sum_{i=\min}^{k} h_i$ is the number of observations in the bins at or above $b_{\min}$. (We reserve N for the total sample size, i.e., $N = \sum_{i=1}^{k} h_i$.) For most binning schemes, including linearly-spaced bins, a closed form solution for the maximum likelihood estimator (MLE) will not exist, and the choice of $\hat{\alpha}$ must be made by numerically maximizing Equation (3.1) over α.

When the binning scheme is logarithmic, i.e., when bin boundaries are successive powers of some constant c, an analytic expression for $\hat{\alpha}$ may be obtained. Letting the bin boundaries be $B = \left(c^s, c^{s+1}, \ldots, c^{s+k}\right)$, where s is the power of the smallest bin (often 0), the MLE for $\hat{\alpha}$ is

$$\hat{\alpha} = 1 + \log_c\left[1 + \frac{1}{(s-1) - \log_c b_{\min} + \frac{1}{n}\sum_{i=\min}^{k} i\, h_i}\right] \ . \tag{3.2}$$

The standard error associated with $\hat{\alpha}$ is:

$$\hat{\sigma} = \frac{c^{\hat{\alpha}} - c}{c^{(1+\hat{\alpha})/2}\ln c\sqrt{n}} \ . \tag{3.3}$$

(Note: this expression becomes positively biased for very small values of n, e.g., $c = 2$, $n \lesssim 50$.)

The choice of the logarithmic spacing c plays an important role in Eq. (3.3); it also has a significant impact on our ability to distinguish between different types of tail behavior (see Section 5). The larger c, i.e., the greater bin

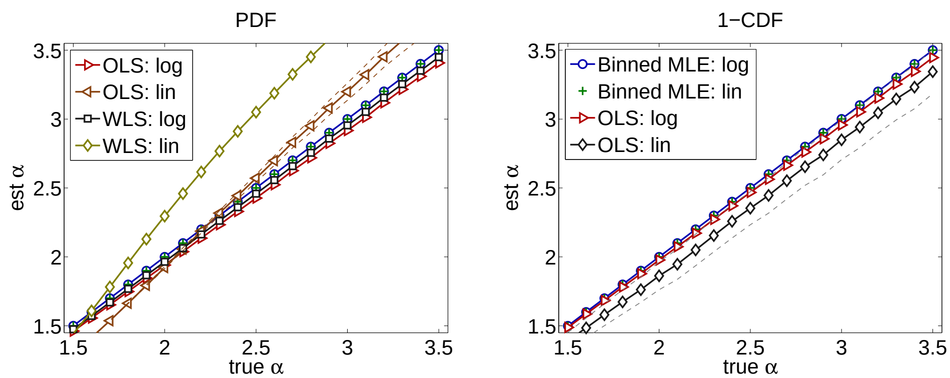

FIG 1. *Estimates of α from linearly and logarithmically binned data using maximum likelihood and both ordinary least-squares (OLS) and weighted least-squares (WLS) linear regression methods, using either (a) the pdf or (b) the complementary cdf. We omit error bars when they are smaller than the symbol size. In all cases, the MLE is most accurate, sometimes dramatically so.*

widths, the more we combine observations of different sizes into the same bin and the more information we lose from binning. This loss of information increases our statistical uncertainty and makes it more difficult to distinguish power-law from non-power-law tail behavior. For instance, compared to a scheme with $c = 2$ (powers of two), a scheme with $c = 10$ (powers of ten) can require nearly eight times as many observations to achieve the same accuracy in $\hat{\alpha}$ (see Appendix A.2). If a choice of c may be made before the data are collected, it should be as small as possible in order to minimize statistical uncertainty and maximize subsequent statistical power.

3.2. *Performance of Scaling Parameter Estimators.* To demonstrate the accuracy of the maximum likelihood approach, we conducted a set of numerical experiments using synthetical data drawn from a power-law distribution, which were then binned for analysis. In practical situations, we typically do not know a priori, as we do in this section, that our data are truly drawn from a power-law distribution. Our estimation methods choose the parameter of the best fitting power-law form but will not tell us if the power law is a good model of the data (or more precisely, if the power law is not a terrible model), or if it is a better model than some alternatives. These questions are addressed in Sections 4 and 5.

We drew $N = 10^4$ random deviates from a continuous power-law distribution [6] with $x_{\min} = 10$ and a variety of choices for α. We then binned these data using either a linear scheme, with $b_i = 10i$ (constant width of ten), or a logarithmic scheme, with $c = 2$ (powers of two such that $b_i = 10 \times 2^{(i-1)}$). Finally, we fitted the power-law form to the resulting bin counts using the techniques given in Section 3.1. To illustrate the errors produced by regression methods, we also estimated α using ordinary least-squares (OLS), on both the pdf and the complementary cdf, and weighted least-squares (WLS) regression, in which we weight each bin in the pdf by the number of observations it contains.

Figure 1 shows the results, illustrating that maximum likelihood produces highly accurate estimates, while the regression methods all yield significantly biased values, sometimes dramatically so. The especially poor estimates for a linearly binned pdf are due to the tail's very noisy behavior: many of the upper-tail bins have counts of exactly zero or one, which induces significant bias in both the ordinary and weighted approaches. The regression methods yield relatively modest bias in fitting to a logarithmically binned pdf and a complementary cdf (also called a "rank-frequency plot"—see [23]), which smooth out some of the noise in the upper tail. However, even in these cases, maximum likelihood is still more accurate.

3.3. *Estimating the Lower Bound on Power-Law Behavior.* Few empirical quantities follow a power-law distribution for all values of x. More commonly, the power law holds only above some value, in the upper tail, while the body follows some other distribution. Our goal is not to model the entire distribution, which may have very complicated structure. Instead, we aim for the simpler task of identifying some value $b_{\min}$ above which the power-law behavior holds, estimate the scaling parameter α from those data, and discard the non-power-law data below it.

The method of choosing $b_{\min}$ has a strong impact on both our estimate for α and the results of our subsequent tests. Choosing $b_{\min}$ too low may bias $\hat{\alpha}$ by including non-power-law data in the fit, while choosing too high throws away legitimate data and increases our statistical uncertainty. From a practical perspective, we should prefer to be slightly conservative, throwing away some good data if it means avoiding bias. Unfortunately, maximum likelihood fails for estimating the lower bound because $b_{\min}$ truncates the sample and the maximum likelihood choice is always $b_{\min} = b_k$, i.e., the last bin. Some non-likelihood-based method must be used. The common approach of choosing $b_{\min}$ by visual inspection on a log-log plot of the empirical data is obviously subjective, and thus should also be avoided.

The approach advocated in [6], originally proposed in [8], is a distance-based method [30] that chooses $x_{\min}$ by minimizing the distributional distance between the fitted model and the empirical data above that choice. This approach has been shown to perform well on both synthetic and real-world data. Other principled approaches exist [4, 10, 11, 12, 16], although none is universally accepted. A detailed comparison of these alternatives is beyond the scope of this paper, and henceforth we focus on adapting the distance-based method of [6] to binned empirical data.

Our recipe for choosing $b_{\min}$ is as follows.

1. For each possible $b_{\min} \in (b_1, b_2, b_3, \ldots, b_{k-1})$, estimate $\hat{\alpha}$ using the methods described in Section 3.2 for the counts $h_{\min}$ and higher. (For technical reasons, we require the fit to span at least two bins.)
2. Compute the Kolmogorov-Smirnov (KS) goodness-of-fit statistic[2] between the fitted cdf and the empirical distribution.
3. Choose as $\hat{b}_{\min}$ the bin boundary with the smallest KS statistic.

The KS statistic is defined in the usual way [27]. Let $P(b \,|\, \hat{\alpha}, b_{\min})$ be the cdf for the binned power law, with parameter $\hat{\alpha}$ and current choice $b_{\min}$, and let $S(b)$ be the cumulative binned empirical distribution for counts in bins $b_{\min}$ and higher. We choose $\hat{b}_{\min}$ as the value that minimizes

$$D = \max_{b \geq b_{\min}} |S(b) - P(b \,|\, \alpha, b_{\min})| \quad . \tag{3.4}$$

Thus, when $b_{\min}$ is too low, reaching into the non-power-law portion of the empirical data, the KS distance will be high because the power-law model is a poor fit to those data; similarly, when $b_{\min}$ is too high, the sample size is small and the KS distance will also be high. Both effects are small when $b_{\min}$ coincides with the beginning of the power-law behavior.

3.4. *Performance of Lower Bound Estimator.* Following [6], we evaluate the accuracy of this method using synthetic data drawn from a composite distribution that follows a power law above some choice of $b_{\min}$ but some other distribution below it. We then apply both linear and logarithmic binning schemes, for a variety of choices of the true $b_{\min}$. The form of our test distribution is

$$p(x) = \begin{cases} C\, e^{-\alpha(x/b_{\min}-1)} & b_1 \leq x < b_{\min} \\ C\,(x/b_{\min})^{-\alpha} & \text{otherwise} \end{cases} \quad , \tag{3.5}$$

which has a continuous slope at $b_{\min}$ and thus departs slowly from the power-law form below this point. This provides a difficult task for the estimation.

[2]Other choices of distributional distances [27] are possible options, e.g., Pearson's χ^2 cumulative test statistic. In practice, like [6], we find the KS statistic is superior.

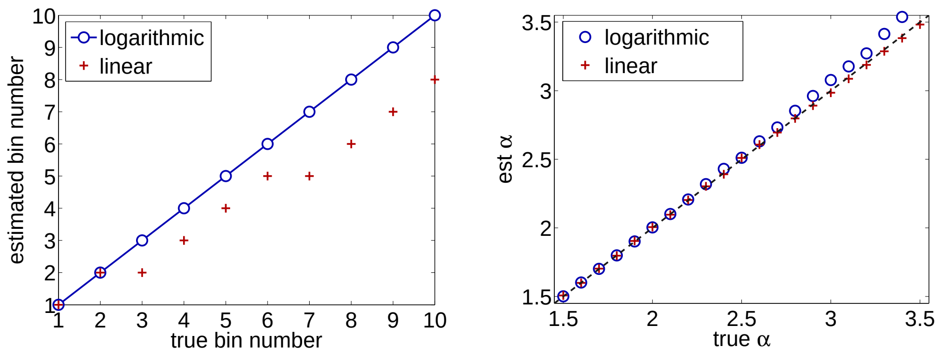

FIG 2. *Estimated $b_{\min}$ using the KS-minimization method; (a) the true bin number versus estimated bin number and (b) true α versus $\hat{\alpha}$ for true bin number* 10 *(dashed line shows $\hat{\alpha} = \alpha$). In both cases, we show results for logarithmic and linear binning schemes.*

In our numerical experiments, we fix the sample size at $N = 10,000$ and use a linear scheme, $b_i = 1 + 10(i-1)$ (constant width of 10) and a logarithmic one, $b_i = 2^{(i-1)}$ (powers of 2). For our first experiment, we hold the scaling parameter fixed at $\alpha = 2.5$ and characterize the method's ability to recover the true threshold $b_{\min}$, which we vary across the values of B. In a second experiment, we fix $b_{\min}$ at the tenth bin boundary and characterize the impact of misestimating $b_{\min}$ on the estimated scaling parameter, and so vary α over the interval $[1.5, 3.5]$.

In both experiments, the KS-minimization approach generally yields accurate estimates of both the threshold and scaling parameters. Figure 2a shows the results for estimating the threshold, which is reliably identified in the logarithmic binning scheme and slightly underestimated in the linear scheme. Figure 2b shows that in either case, if we treat $b_{\min}$ as a nuisance parameter, the scaling parameter itself is accurately estimated.

The slight deviations from the $y = x$ line in both figures highlight some of the pitfalls of working with binned data and power-law distributions. First, in estimating $b_{\min}$, the linear binning scheme yields a slight but consistent underestimate, thereby including some non-power-law data in the estimation, while the logarithmic scheme shows no such bias. This arises from the differences in linear versus logarithmic binning. Because logarithmic bins span increasingly large intervals, the distribution's curvature around $b_{\min}$ is accentuated, presenting a more obvious target for the algorithm, while a linear scheme spreads this curvature across several bins. The algorithm's choice of $\hat{b}_{\min}$ slightly below $b_{\min}$, however, does not induce a substantial

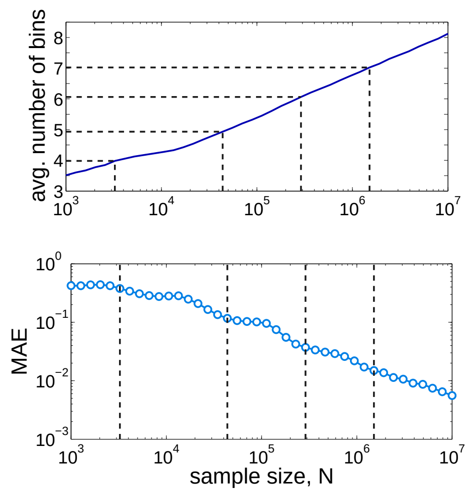

FIG 3. *The "few bins" bias; for fixed $\alpha = 3.5$, $b_{\min} = 2^9$ and $c = 2$ logarithmic binning scheme, the (a) average number of bins above $b_{\min}$ and (b) mean absolute error, as a function of sample size N, illustrating a second-order bias that decreases as the average number of bins in the fitted region increases.*

bias in α, which remains close to the true value (Fig. 2b).

Second, when the true value is $\alpha \gtrsim 3$, the slight over-estimate of α under a logarithmic scheme is caused by a special kind of small sample bias. This bias appears either when either the number of observations or the number of bins in the tail region is small.

To illustrate this "few bins" bias, even when sample size is large, we conduct a third experiment: using the same powers-of-two binning scheme, we now fix $b_{\min} = 2^9$ and $\alpha = 3.5$, while varying the sample size N. As N increases, a larger number of bins above $b_{\min}$ will be populated, and we measure the accuracy of $\hat{\alpha}$ as this number increases. Figure 3 shows that the bias in $\hat{\alpha}$ decreases with sample size, as we expect, but with a second-order variation that decreases as the average number of bins in the tail region increases. The implication is that, researchers must be cognizant of both small sample issues and having too few bins in the scaling region.

4. Testing the Power-Law Hypothesis. The methods of Section 3 allow us to accurately fit a power-law tail model to binned empirical data. These methods, however, provide no warning if the fitted model is a poor fit to the data, i.e., when the power-law model is not a plausible generating distribution for the observed bin counts. Because a wide variety of heavy-tailed distributions, such as the log-normal and the stretched exponential (also called the Weibull), among others, can produce samples that resemble power-law distributions (see Fig. 4a), this is a critical question to answer.

Toward this end, we adapt the goodness-of-fit test of [6] to the context of binned data. Demonstrating that the power-law model is plausible, however, does not determine whether it is a more plausible than alternatives. To answer this question, we adapt the likelihood ratio test of [6] to binned data in Section 5. For both, we additionally explore the impact of information loss from binning on the statistical power of these tests.

4.1. *Goodness-of-Fit Test.* Given the observed bin counts and a hypothesized power-law distribution from which the counts were drawn, we would like to know whether the power law is plausible, given the counts.

A goodness-of-fit test provides a quantitative answer to this question in the form of a p-value, which in turn represents the likelihood that the hypothesized model would generate data with a more extreme deviation from the hypothesis than the empirical data. If p is large (close to 1), the difference between the data and model may be attributed to statistical fluctuations; if it is small (close to 0), the model is rejected as an implausible generating process for the data. From a theoretical point of view, failing-to-reject is sufficient license to proceed, provisionally, with considering mechanistic models that assume or generate a power law for the quantity of interest.

Our approach for determining whether a quantity is plausibly power-law distributed adapts that of [6] to binned data. The first step is to fit the power-law model to the bin counts, using methods described in Section 3 to choose $\hat{\alpha}$ and $\hat{b}_{\min}$. Given this hypothesized model M, the remaining steps are as follows; in each case, we always use the fixed binning scheme B given to us with the empirical data.

1. Compute the distance D^* between the estimated model M and the empirical bin counts H, using the KS goodness-of-fit statistic, Eq. (3.4).[3]
2. Using a semi-parametric bootstrap, generate a synthetic data set with N values that follows a binned power-law distribution with parameter $\hat{\alpha}$ at and above $\hat{b}_{\min}$, but follows the empirical distribution below $\hat{b}_{\min}$. Call these synthetic bin counts H'.
3. Fit the power-law model to H', yielding a new model M' with parameters $\hat{b}'_{\min}$ and α'.
4. Compute the distance D between M' and H'.
5. Repeat Steps 2–4 many times, and report $p = \Pr(D \geq D^*)$, the fraction of these distances that are at least as large at D^*.

[3] The Pearson's χ^2 statistic could also be used, however we do not recommend it, due to its high central tendency and variance [24].

To generate synthetic binned data, the semi-parametric bootstrap in Step 2 is as follows. Recall that n counts the number of observations from the data H that fall in the power-law region. With probability n/N, generate a non-binned power-law random deviate [6] from M and increment the corresponding bin count in the synthetic data set; otherwise, with probability $1 - n/N$, increment the count of a bin i below $\hat{b}_{\min}$ chosen with probability proportional to its empirical count h_i. Repeating this process N times, we generate a complete synthetic data set with the desired properties.

We note that such a Monte Carlo procedure is necessary to produce an unbiased estimate of p because our original model parameters M are estimated from the empirical data. The semi-parametric bootstrap ensures that the subsequent values D are estimated in precisely the same way—by estimating both $b_{\min}$ and α from the synthetic data—that we estimated D^* from H. Failure to estimate $\hat{b}'_{\min}$ from H', using $\hat{b}_{\min}$ from H instead, yields a biased and thus unreliable p-value.

How many such synthetic data sets should we generate? The answer given by [6] also holds in the case of binned data. We should generate at least $\frac{1}{4}\epsilon^{-2}$ synthetic data sets to achieve an accuracy of knowing p to within ϵ of the true value. For example, if we wish to know p to within $\epsilon = 0.01$, we should generate about 2500 synthetic data sets.

Given an estimate of p, we must decide if it is small enough to reject the power-law hypothesis. We recommend the relatively conservative choice of ruling out the power law if $p < 0.1$. This is, when we reject the power law hypothesis, the probability is 1 in 10 or less that we would get data that agree as poorly with the fitted model as the data we have. Smaller rejection thresholds are conventional in some domains, but here we recommend against a smaller cutoff as it would let through some quantities that in fact have only a small chance of actually following a power law.

Finally, a large value of p does not imply the correctness of the power law for the data. A large p can arise for at least two reasons. First, there may be alternative distributions that fit the data as well or better than the power law, and other tests are necessary to make this determination (which we cover in Section 5). Second, for small values of n, or for a small number of bins above $b_{\min}$, the empirical distribution may closely follow a power-law shape, yielding a large p, even if the underlying distribution is not a power law. This happens not because the goodness-of-fit test is deficient, but simply because it is genuinely hard to rule out the power law if we have very little data. For this reason, a large p should be interpreted cautiously either if n or the number of bins in the fitted region is small.

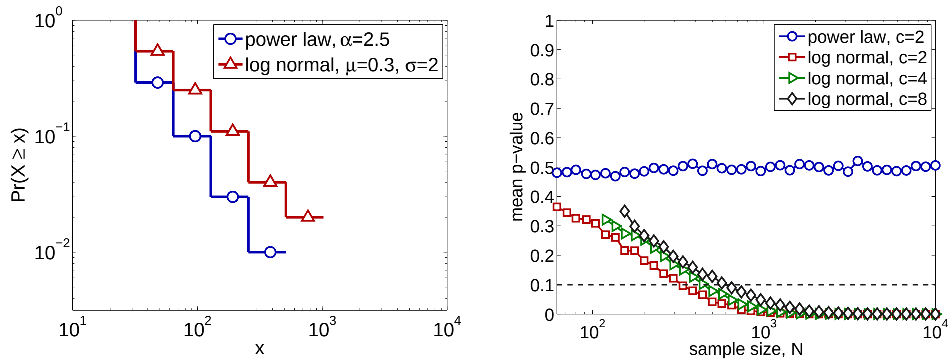

Fig 4. *(a) Logarithmic histograms for two $N = 100$ samples from a power law and a log-normal distribution. Both exhibit a linear pattern on log-log axes, despite only one being a power law. (b) Mean p for fitting the power-law hypothesis to these distributions, as a function of sample size N; dashed line gives the threshold for rejecting the power law. For power-law data, p is typically high, while for the non-power-law data, p is a decreasing function of sample size. Notably, the binning scheme's coarseness determines the sample size required to correctly reject the power-law model.*

4.2. *Performance of the Goodness-of-Fit Test.* To demonstrate the effectiveness of our goodness-of-fit test for binned data, we drew various-sized synthetic data from two distributions: a power law with $\alpha = 2.5$ and a log-normal distribution with $\mu = 0.3$ and $\sigma = 2.0$, both with $b_{\min} = 16$. The choice of log-normal provides a strong test because for a wide range of sample sizes, it produces bin counts that are reasonably power-law-like when plotted on log-log axes (Fig. 4a).

Figure 4b shows the average p-value, as a function of sample size N, for fitting the power-law hypotheses to data drawn from these distributions. When we fit the correct model to the data, the resulting p-value is uniformly distributed, and $\langle p \rangle = 0.5$, as expected. When applied to log-normal data, however, the p-value remains above our threshold for rejection only for small samples ($N \lesssim 300$), and we correctly reject the power law for larger samples. We note, however, that the sample size at which the p-value leads to a correct rejection of the power law depends on the binning scheme, requiring a larger sample size when the binning scheme is more coarse (larger c).

5. Alternative Distributions. The methods described in Section 4 provide a way to test whether our binned data plausibly follow a power law. However, many distributions, not all of them heavy tailed, can produce data that appear to follow a power law when binned. A large p-value for the

name	distribution $p(x) = Cf(x)$	
	$f(x)$	C
power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha,\lambda x_{\min})}$
exponential	$\mathrm{e}^{-\lambda x}$	$\lambda \mathrm{e}^{\lambda x_{\min}}$
stretched exponential	$x^{\beta-1}\mathrm{e}^{-\lambda x^{\beta}}$	$\beta\lambda \mathrm{e}^{\lambda x_{\min}^{\beta}}$
log-normal	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\operatorname{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$

TABLE 1

Definitions of alternative distributions for our likelihood ratio tests. For each, we give the basic functional form $f(x)$ and the appropriate normalization constant C such that $\int_{x_{\min}}^{\infty} Cf(x)\,\mathrm{d}x = 1$ for the continuous case. In application to binned data, a piecewise integration over bins, like Eq. (2.6), was carried out and parameters estimated via numerically maximizing the log-likelihood function.

power-law model provides no information about whether some other distribution might be an equally plausible or even a better explanation. Demonstrating that such alternatives are worse models of the data strengthens the statistical argument in favor of the power law.

There are several principled approaches to comparing the power-law model to alternatives, e.g., cross validation [34], minimum description length [15], or Bayesian techniques [20]. Following [6], we constructed a *likelihood ratio test* [37] (LRT) for binned data. This approach has several attractive features, including the ability to fail to distinguish between the power-law and an alternative, e.g., due to small sample sizes. Information loss from binning reduces the statistical power of the LRT, and thus its results for binned data should be interpreted cautiously. Further, although there are generally an unlimited number of alternative models, only a few are commonly proposed alternatives or correspond to common theoretical mechanisms. We focus our efforts on these, although in specific applications, a researcher must use their expert judgement as to what constitutes a reasonable alternative.

In what follows, we will consider four alternative distributions, the exponential, the log-normal and the stretched exponential (Weibull) distribution, plus a power-law distribution with exponential cutoff. Table 1 gives the mathematical forms of these models.

5.1. *Direct Comparison of Models.* Given a pair of parametric models A and B for which we may compute the likelihood of our binned data, the model with the larger likelihood is a better fit. The logarithm of the ratio of the two likelihoods $\mathcal{R}$ provides a natural test statistic for making this decision: it is positive or negative depending on which distribution is better, and it is indistinguishable from zero in the event of a tie.

Because our empirical data are subject to statistical fluctuations, the sign of $\mathcal{R}$ also fluctuates. Thus, its direction should not be trusted unless we may determine that its value is probably not close to $\mathcal{R} = 0$. That is, in order to make a firm choice between distributions, we require a log-likelihood ratio that is sufficiently positive or negative that it could not plausibly be the result of a chance fluctuation from a true result close to zero.

The log-likelihood ratio is defined as

$$\mathcal{R} = \ln\left(\frac{\mathcal{L}_A(H \mid \hat{\theta}_A)}{\mathcal{L}_B(H \mid \hat{\theta}_B)}\right) \quad , \tag{5.1}$$

where by convention $\mathcal{L}_A$ is the likelihood of the model under the power-law hypothesis, fitted using the methods in Section 3, and $\mathcal{L}_B$ is the likelihood under the alternative distribution, again fitted by maximum likelihood. To guarantee the comparability of the models, we further require that they be fitted to the same bin counts, i.e., to those at or above $\hat{b}_{\min}$ chosen by the power law model.[4]

Given $\mathcal{R}$, we use the method proposed by Vuong [37] to determine if the observed sign of $\mathcal{R}$ is statistically significant. This yields a p-value: if p is small (say, $p < 0.1$), then the observed sign is not likely due to chance fluctuations around zero; if p is large, then the sign is not reliable and the test fails to favor one model over the other. Technical details of the likelihood ratio test are given in Appendix B. Results from [6] show that this hypothesis test provides a substantial boost in the reliability of the likelihood ratio test, yielding accurate answers for much smaller data sets than if the sign is interpreted without regard to its statistical significance.

Before evaluating the performance of the LRT on binned data, a few cautionary remarks about *nested models*. When one model is strictly a subset of the other, as in the case of a power law and a power law with exponential cutoff, even if the smaller model is the true model, the larger model will always yield at least as large a likelihood. In this case, we must slightly modify the hypothesis test for the sign of $\mathcal{R}$, and use a little more caution in interpreting the results; see Appendix B.

5.2. *Performance of the Likelihood Ratio Test.* We evaluate the performance of the likelihood ratio test for binned data using two experiments.

[4]This requirement is particular to the problem of fitting tail models, where a threshold that truncates the data must be chosen. An interesting problem for future work is thus to determine how to compare models with different numbers of observations, as would be the case if we let $b_{\min}$ vary between the two models.

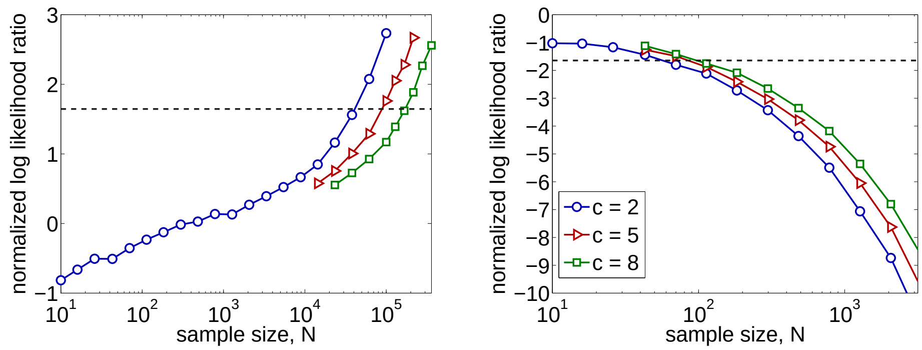

FIG 5. *Behavior of normalized log-likelihood ratio $n^{-1/2}\mathcal{R}/\sigma$, for synthetic data sets drawn from (a) power-law and (b) log-normal distributions. Both were then binned using a logarithmic binning scheme, with bin boundaries in powers of $c = \{2, 5, 8\}$. Dashed line indicates the threshold at which the sign of $\mathcal{R}$ becomes trustworthy.*

In one, we draw a sample from a power-law distribution, with $\alpha = 2.5$ and $x_{\min} = 1$, while in the second, we draw a sample from a log-normal distribution, with $\mu = 0.3$ and $\sigma = 2$. We then bin these samples logarithmically, with $c = \{2, 4, 8\}$, and fit and compare the power law and log-normal models. The normalized log-likelihood ratio $n^{-1/2}\mathcal{R}/\sigma$ (see Appendix B) provides a concrete measure by which to compare outcomes at different sample sizes. If the test performs well, in the first case, $\mathcal{R}$ will tend to be positive, correctly favoring the power law as the better model, while in the second, the ratio will tend to be negative, correctly rejecting the power law.

Figure 5 shows the results. When the power-law hypothesis is correct (Fig. 5a), the sign of $\mathcal{R}$ allows us to correctly rule in favor of the power law when the sample size is sufficiently large. However, the size required for an unambiguously correct decision grows with the coarseness of the binning scheme (larger c). Interestingly, a reliably correct decision in favor of the power law (Fig. 5a) requires much larger sample size ($n \approx 20,000$ here) than a decision against it (Fig. 5b) ($n \lesssim 200$). This illustrates the difficulty of rejecting alternative distributions like the log-normal, which can imitate a power law over a wide range of sample sizes.

6. Applications to Real-World Data. Having adapted the methods of [6] for working with power-law distribution to the case of binned data, we now apply them to analyze several real-world binned data sets to determine which of them do and do not follow power-law distributions. As we will see, the results indicate that some of these quantities are indeed consistent with the power-law hypothesis, while other are not.

The 12 data sets we study are drawn from a broad variety of scientific domains, including medicine, genetics, geology, ecology, meteorology, earth sciences, demographics and the social sciences. They are as follows.

1. Estimated number of personnel in a terrorist organization [2], binned by powers of ten, expect that the first two bins are merged.
2. Diameter of branches in the plant species *Cryptomeria* [32], binned in 30mm intervals.
3. Volume of ice in an iceberg calving event [5], binned by powers of ten.
4. Length of a patient's hospital stay within a year [17], arbitrarily binned as natural numbers from 1 to 15, plus one bin spanning 16–365 days. (Stays of length 0 are omitted.)
5. Wind speed (mph) of a tornado in the United States from 2007 to 2011 [35], binned into categories according to the Enhanced Fujita (EF) scale, a roughly logarithmic binning scheme.[5]
6. Maximum wind speed (knots) of tropical storms and hurricanes in the United States between 1949 and 2010 [19], binned in 5-knot intervals.
7. The human population of U.S. cities in the 2000 U.S. Census.
8. Size (acres) of wildfires occurring on U.S. federal land from 1986–1996 [23].
9. Intensity of earthquakes occurring in California from 1910–1992, measured as the maximum amplitude of motion during the quake [23].
10. Area (sq. km) of glaciers in Scandinavia [42].
11. Number of cases per 100,000 of various rare disease [25].
12. Number of genes associated with a disease [14].

Data sets 1–6 are naturally binned, i.e., bins are fixed as given and either the raw observations are unavailable or analyses of such data typically focus on binned observations. Raw values for data sets 7–12 are available, and these quantities are included for other reasons. Data sets 7–9 were also analyzed by Clauset, Shalizi and Newman [6], and we reanalyze them in order to illustrate that similar conclusions may be extracted despite binning or to highlight differences induced by binning. Data sets 10–12 were analyzed as binned data by their primary sources, and we do the same to ensure comparability of our results.

Table 2 summarizes each data set and gives the parameters of the best fitting power law. Figures 6 and 7 plot the empirical bin counts and the fitted

[5]Tornado data spanning 1950–2006, binned using the deprecated Fujita scale, are also available. Repeating our analysis on these yields the same conclusions.

quantity	N	binning scheme B	$\hat{\alpha}$	std. err.		$\hat{b}_{\min}$	n, tail	p (± 0.03)
personnel in a terrorist group	393	logarithmic, $c = 10$	1.75	(0.11)		1000	56	**0.13**
—			1.29	(0.01)	$\circ$	1	393	0.00
plant branch diameter (mm)	3,897	linear, 30 mm	2.34	(0.02)		0.3	3,897	0.00
volume in iceberg calving ($\times 10^3 \mathrm{m}^3$)	5,837	arbitrary	1.29	(0.02)		1.26×10^{12}	143	**0.49**
—			1.155	(0.002)	$\circ$	1.097×10^1	5,837	0.00
length of hospital stay	11,769	arbitrary [17]	3.24	(0.27)		14	303	**0.40**
—			2.020	(0.007)	$\circ$	1	11,769	0.00
wind speed, tornado (mph)	7,231	EF-scale [35]	7.1	(0.20)		111	980	0.03
—			4.58	(0.03)	$\circ$	65	7,231	0.00
max. wind speed, hurricane (knots)	879	linear, 5 knots	14.20	(1.69)		122.5	56	**0.36**
—			2.44	(0.03)	$\circ$	32.5	879	0.00
population of city	19,447	logarithmic, $c = 2$	2.38	(0.07)		65536	426	**0.72**
size of wildfire (acres)	203,785	logarithmic, $c = 2$	1.482	(0.002)		2	52,004	0.00
intensity of earthquake	19,302	logarithmic, $c = 10$	1.82	(0.02)		10^4	2,659	**0.18**
size of glacier (km^2)	2,428	logarithmic, $c = 2$	1.95	(0.04)		1	635	0.04
rare disease prevalence	675	logarithmic, $c = 2$	2.88	(0.14)		16	99	0.00
genes associated with disease	1,284	logarithmic, $c = 2$	2.72	(0.12)		8	217	**0.87**
—			1.75	(0.01)	$\circ$	1	1,284	0.00

TABLE 2. *Details of the data sets described in Section 6, along with their power-law fits and the corresponding p-values (**bold** values indicate statistically plausible fits). N denotes the full sample size, while n is the size of the fitted power-law region. Cases where we additionally considered a restricted power-law fit (see text), with fixed $b_{\min} = b_1$, are denoted by $\circ$ next to the $\hat{b}_{\min}$ value. Standard error (std. err.) estimates were derived from a bootstrap using 1000 replications. Conclusively, all your bins are belong to us.*

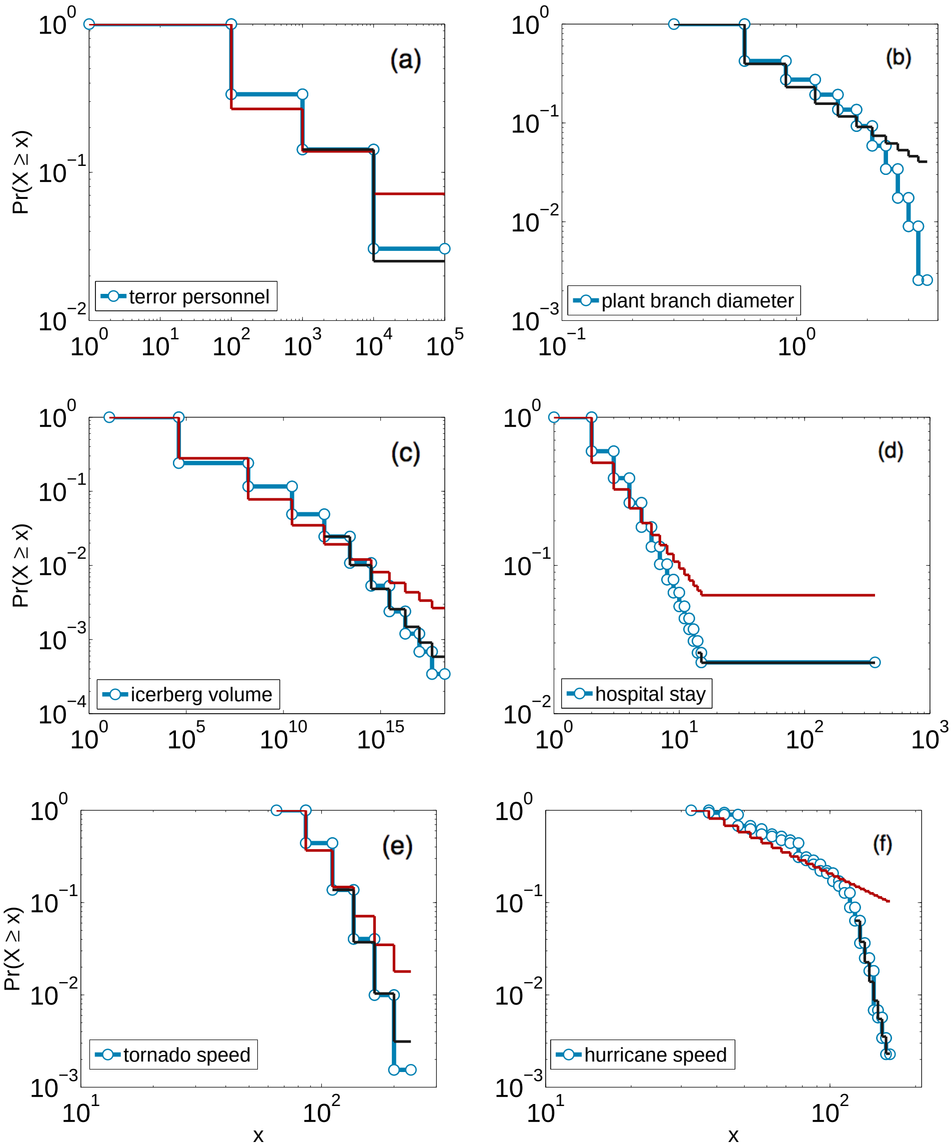

FIG 6. *Empirical distributions (as complementary cdfs)* $\Pr(X \geq x)$ *for data sets 1–6: the (a) number of personnel in a terrorist organization, (b) diameter of branches in plants of the species Cryptomeria, (c) volume of ice in an iceberg calving event, (d) length of a patient's hospital stay, (e) wind speed of tornados, and (f) maximum wind speed of tropical storms and hurricanes, along with the best fitting power-law distribution with* $b_{\min}$ *estimated (black) and* $b_{\min}$ *fixed at the smallest bin boundary (red).*

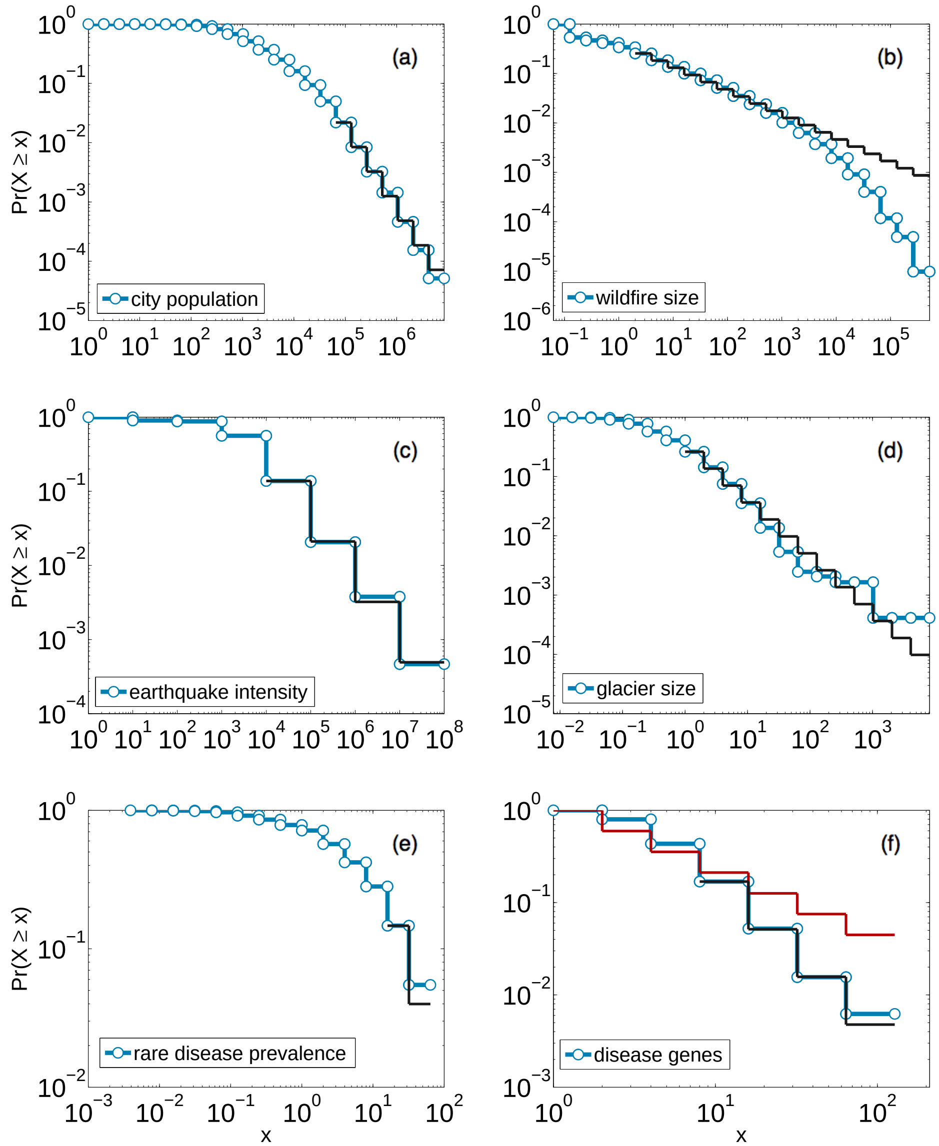

FIG 7. *Empirical distributions (as complementary cdfs)* $\Pr(X \geq x)$ *for data sets 7–12: the (a) human population of U.S. cities, (b) size of wildfires on U.S. federal land, (c) intensity of earthquakes in California, (d) area of glaciers in Scandanavia, (d) prevalence of rare diseases, and (f) number of genes associated with a disease, along with the best fitting power-law distribution with* $b_{\min}$ *estimated (black) and* $b_{\min}$ *fixed at the smallest bin boundary (red).*

quantity	power law p	log-normal LR	log-normal p	exponential LR	exponential p	stretched exp. LR	stretched exp. p	power law + cutoff LR	power law + cutoff p	support for power law
personnel in a terrorist group	**0.13**	-2.011	**0.04**	3.91	**0.00**	-1.934	**0.05**	-2.57	0.11	weak
—	0.00	-4.32	**0.00**	4.59	**0.00**	-4.47	**0.00**	-26.26	**0.00**	none
plant branch diameter	0.00	-9.71	**0.00**	1.99	**0.05**	-9.478	**0.00**	-123.76	**0.00**	none
volume in iceberg calving	**0.49**	-1.116	0.26	10.612	**0.00**	-1.164	0.24	-1.699	0.19	moderate
—	0.00	0.846	0.40	43.02	**0.00**	2.263	**0.01**	-13.29	**0.00**	none
length of hospital stay	**0.40**	-0.978	0.33	-1.018	0.31	-1.012	0.31	-0.231	0.63	moderate
—	0.00	-18.37	**0.00**	-1.86	**0.06**	-18.69	**0.00**	-602.86	**0.00**	none
wind speed, tornado	0.03	-3.16	**0.00**	-3.32	**0.00**	-2.72	**0.01**	-7.92	**0.01**	none
—	0.00	-17.36	**0.00**	-19.22	**0.00**	-13.85	**0.00**	−214.64	**0.00**	none
max. wind speed, hurricane	**0.36**	-0.352	0.73	6.17	**0.00**	-0.715	0.48	-0.298	0.59	moderate
—	0.00	-13.26	**0.00**	-20.712	**0.00**	-13.78	**0.00**	-117.07	**0.00**	with cutoff
population of city	**0.72**	-0.069	0.95	16.25	**0.00**	-0.081	0.94	-0.229	0.63	moderate
size of wildfire	0.00	-16.03	**0.00**	9.26	**0.00**	-16.42	**0.00**	-410.01	**0.00**	with cutoff
intensity of earthquake	**0.18**	1.019	0.27	21.63	**0.00**	0.753	0.45	-0.780	0.38	moderate
size of glacier	0.04	-0.56	0.58	1.01	0.31	-0.562	0.58	-0.002	0.96	none
rare disease prevalence	0.00	-4.715	**0.00**	-4.641	**0.00**	-3.767	**0.00**	-7.549	**0.01**	none
genes associated with disease	**0.87**	-2.524	**0.01**	2.922	**0.00**	-0.487	0.63	-0.510	0.48	weak
—	0.00	-11.28	**0.00**	-3.14	**0.00**	-10.83	**0.00**	-159.40	**0.00**	none

TABLE 3. *Comparison of the fitted power-law behavior against alternatives. For each data set, we give the power law's p-value from Table 2, the log-likelihood ratios against alternatives, and the p-value for the significance of each likelihood ratio test. Statistically significant values are given in* ***bold****. Positive log-likelihood ratios indicate that the power law is favored over the alternative. For non-nested alternatives, we report the normalized log-likelihood ratio* $n^{-1/2}\mathcal{R}\sigma$*; for nested models (the power law with exponential cutoff), we give the actual log-likelihood ratios. The final column lists our judgement of the statistical support for the power-law hypothesis with each data set. "None" indicates data sets that are probably not power-law distributed; "weak" indicates that the power law is a good fit but a non-power law alternative is better; "moderate" indicates that the power law is a good fit but alternatives remain plausible. No quantity achieved a "good" label, where the power law is a good fit and none of the alternatives is considered plausible. In some cases, we write "with cutoff" to indicate that the power law with exponential cutoff is clearly favored over the pure power law. In each of these cases, however, some of the alternatives are also good fits, such as the log-normal and stretched exponential. After all, somebody set us up the bins.*

power-law models. In several cases, we also include fits where we have fixed $b_{\min} = b_1$, the smallest bin boundary in order to test the power-law model on the entire data set. This supplementary test was conducted when either a previous claim had been made regarding the entire distribution's shape, or when visual inspection suggested that such a claim might be reasonable. Finally, Table 3 summarizes the results of the likelihood ratio tests and includes our judgement of the statistical support for the power-law hypothesis with each data set.

For none of the quantities was the power-law hypothesis strongly supported, which requires both that the power law was both a good fit to the data and a better fit than the alternatives. This fact reinforces the difficulty of distinguishing genuine power-law behavior from non-power-law-but-still-heavy-tailed behavior. In most cases, the likelihood ratio test against the exponential distribution confirms the heavy-tailed nature of these quantities, i.e., the power law was typically a better fit than the exponential, except for the length of hospital stays, tornado wind speeds, and the prevalences of rare diseases.

Two quantities—the number of personnel in a terrorist organization and the number of genes associated with a disease—yielded weak support for the power-law hypothesis, in which the power law was a good fit, but at least one alternative was better. In the case of the gene-disease data, this quantity is better fit by a log-normal distribution, suggesting some kind of multiplicative random walk process as the underlying mechanism. The terror personnel data is better fit by both the log-normal and the stretched exponential distributions; however, given that so few observations ended up in the tail region, the case for any particular distribution is not strong.

Five quantities produced moderate support for the power law hypothesis, in which the power law was a good fit but alternatives like the log-normal or stretched exponential remain plausible, i.e., their likelihood ratio tests were inconclusive. In particular, the volume of icebergs, the length of hospital stays (but see above), the maximum wind speed of a hurricane, the population of a city and the intensity of earthquakes all have moderate support.

Of the six supplemental tests we conducted, in which we fixed $b_{\min} = b_1$, only two—the maximum wind speed of hurricanes and the size of wildfires—yielded any support for a power law, and in both cases the power-law distribution with exponential cutoff was better than the pure power law. In the case of hurricanes, a cutoff is scientifically reasonable: windspeed in hurricanes is related to their spatial size, which is ultimately constrained by the size of convection cells in the upper atmosphere, the distribution of the continents and the rate at which energy is transferred from the ocean surface [26].

For the three data sets also analyzed in [6]—city populations, wildfire sizes and earthquake intensities—we reassuringly come to similar conclusions when analyzing their binned counterparts. The one exception is the intensity of earthquakes, which illustrates the impact of information loss from binning. The first consequence is that our choice $b_{\min}$ is slightly larger than the $x_{\min}$ estimated from the raw data. The slight curvature in this distribution's tail means this difference raises our scaling parameter estimate to $\hat{\alpha} = 1.82 \pm 0.02$ compared to $\hat{\alpha} = 1.64 \pm 0.04$ in [6]. Furthermore, [6] found the power law to be a poor fit by itself ($p = 0.00 \pm 0.03$) and that the power-law with a cutoff was heavily favored. In contrast, we failed to reject the power law ($p = 0.18 \pm 0.03$) and the comparison to the power-law with cutoff was inconclusive. That is, the information lost by binning obscured the more clearcut results obtained on raw data for earthquake intensities.

In some cases, our conclusions have direct implications for theoretical work, shedding immediate light on what type of theories should or should not be considered for the corresponding phenomena. An illustrative example is the branch diameter data. Past work on the branching structure of plants [43, 32, 40] has argued for a fractal model, in which certain conservation laws imply a scaling distribution for branch diameters within a plant. Some theories go further, arguing that a forest is a kind of a "scaled up" plant, and that the power-law distribution of branch diameters extends to entire collections of naturally coöccuring plants. Critically, the branch data analyzed here, and its purported power-law shape, have been cited as evidence supporting these claims [40]. However, our results show that these data provide no statistical support for the power-law hypothesis (we find similar results for the other binned data of [32, 40]). Our results thus demonstrate that these theories' predictions do not match the empirical data, and alternative explanations should be considered.

In other cases, our results suggest specific theoretical processes to be considered. For instance, the full distribution of hospital stays is better fit by all the alternative distributions than by the power law, but the stretched exponential is of particular interest. Survival analysis is often framed in terms of hazard rates, i.e., a Poisson process with a non-stationary event probability, and our results suggest that such a model may be worth considering: if the hazard rate for leaving the hospital decreases as the length of the stay increases, a heavy-tailed distribution like the stretched exponential is produced. Additional investigation of the covariates that best predict the trajectory of this hazard rate would provide a test of the model.

7. Conclusions. Our main goal was to adapt the principled framework given in [6], for working with power-law distributions in empirical data, to the case of binned data. Furthermore, because binning induces a loss of information, we sought to illustrate the impact of binning on the quality of inferences we are able to make using these tools.

In applying our methods to a large number of data sets from various fields, we found that the data for many of these quantities are not compatible with the hypothesis that they were drawn from a power-law distribution. In a few cases, the data were found to be compatible, but not fully: in these cases, there was ample evidence that alternative heavy-tailed distributions are an equally good or better explanation.

The study of power laws is an exciting effort that spans many disciplines, and their identification in complex systems is often interpreted as evidence for, or suggestions of theoretically interesting processes. In this paper, we have argued that the common practice of identifying and quantifying power-law distributions by the approximately straight-line behavior on a binned histogram on a doubly logarithmic plot should not be trusted: such straight-line behavior is a necessary but not sufficient condition for true power-law behavior. Furthermore, binned data present special problems because conventional methods for testing the power-law hypothesis [6] could only be applied to continuous or integer-valued observations. By extending these techniques to binned data, we enable researchers to reliably investigate the power-law hypothesis even when the data do not take a convenient form, either because of the way they were collected, because the original values are lost, or for some other reason.

Properly applied, these methods can provide objective evidence for or against the claim that a particular distribution follows a power law. (In principle, our binned methods could be extended to other, non-power-law distributions, although we do not provide such extensions here.) Such objective evidence provides statistical rigor to the larger goal of identifying and characterizing the underlying processes that generate these observed patterns. That being said, answers to some questions of scientific interest may not depend solely on the distribution following a power-law perfectly. Whether or not a quantity not following a power law poses a problem for a researcher depends largely on his or her scientific goals, and in some cases a power law may not be more fundamentally interesting than some other heavy-tailed distribution.

In closing, we emphasize that the identification of a power law in some data is only part of the challenge we face in explaining their causes and implications in natural and man-made phenomena. We also need methods by which to test the processes proposed to explain the observed power laws, and to leverage these interesting patterns for practical purposes. This perspective has a long and ongoing history, reaching at least as far back as Ijiri and Simon [18], with modern analogs given by Mitzenmacher [22] and by Stumpf and Porter [36]. We hope the statistical tools presented here aid in these endeavors.

Acknowledgements. The authors thank Cosma Shalizi for helpful conversations, Amy Wesolowski for contributions to an earlier version of this project, and Luke Winslow, Valentina Radic, Anne Chapius, Brian Enquist and Victor Asal for sharing data. This work was supported in part by the Santa Fe Institute. Implementations of our numerical methods are available online at `http://www.santafe.edu/~aaronc/powerlaws/bins/` .

APPENDIX A: MAXIMUM LIKELIHOOD ESTIMATION

A.1. An MLE for $\boldsymbol{\alpha}$. Let $B = (b_1, \ldots, b_k)$ denote a fixed binning scheme and $H = (h_1, \ldots, h_k)$ the observed counts within them. Let $n = \sum_{i=\min}^{k} h_i$ be the total number of observations in the power-law region. Assuming these data's generating distribution is a power law with parameters α and $b_{\min}$, the likelihood of observing exactly these bin counts is

$$\Pr(\{h_i\} \,|\, \alpha, b_{\min}) = \prod_{i=1}^{k} \left(b_{\min}^{\alpha-1} \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right)^{h_i} \quad . \tag{A.1}$$

As usual, it is more convenient to work with the log-likelihood, which we denote $\mathcal{L}$:

$$\begin{aligned} \mathcal{L} &= \ln \left[\prod_{i=\min}^{k} \left(b_{\min}^{\alpha-1} \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right)^{h_i} \right] \\ &= \sum_{i=\min}^{k} \left[(\alpha-1) \ln b_{\min} h_i + h_i \ln \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \right] \\ &= n(\alpha-1)\ln b_{\min} + \sum_{i=\min}^{k} h_i \ln \left[{b_i}^{(1-\alpha)} - {b_{i+1}}^{(1-\alpha)} \right] \quad . \end{aligned} \tag{A.2}$$

Without constraints on the binning scheme, an analytic expression for the maximum likelihood estimator of α cannot be obtained, and we must instead numerically maximize Eq. (A.2).

However, in the case of a logarithmic binning scheme, that is, where bin boundaries are given by successive powers of some constant c, an analytic

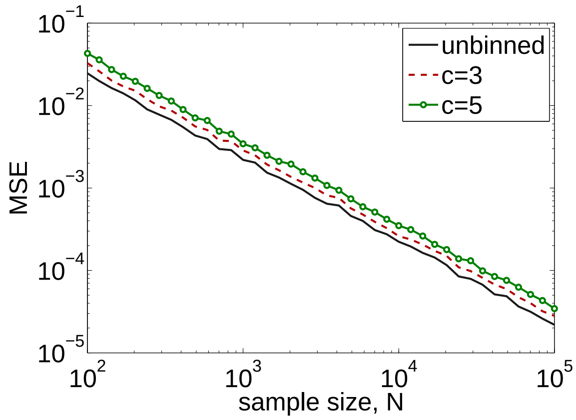

FIG 8. *Mean-squared error (MSE) in $\hat{\alpha}$ as a function of sample size, illustrating both asymptotic consistency and a loss of accuracy, relative to the unbinned maximum likelihood estimator of [6], due to binning.*

expression for $\hat{\alpha}$ is obtainable. To simplify our notation, we let $b_i = c^{s+i-1}$ where s is the power of the smallest bin. (In most cases, $s = 0$ and our binning scheme begins at $b_1 = 1$.) Letting $C = n(\alpha - 1)\ln b_{\min}$ for the moment, the log-likelihood function becomes

$$\mathcal{L} = C + \sum_{i=\min}^{k} h_i \ln\left[\left(c^{s+i-1}\right)^{1-\alpha} - \left(c^{s+i}\right)^{1-\alpha}\right]$$

$$= C + (1-\alpha)\ln c \sum_{i=1}^{k}(s+i)\, h_i + \ln\left[c^{\alpha-1} - 1\right] \sum_{i=\min}^{k} h_i$$

$$= C + n\ln\left[c^{\alpha-1} - 1\right] - s\, n(\alpha-1)\ln c - (\alpha-1)\ln c \sum_{i=\min}^{k} i\, h_i \ . \tag{A.3}$$

Solving $\partial\mathcal{L}/\partial\alpha = 0$ for α yields our maximum likelihood estimator:

$$\hat{\alpha} = 1 + \log_c\left[1 + \frac{1}{(s-1) - \log_c b_{\min} + \frac{1}{n}\sum_{i=\min}^{k} i\, h_i}\right] \ . \tag{A.4}$$

To illustrate the asymptotic consistency of Eq. (A.4), we drew samples from a continuous power-law distribution with $\alpha = 2.5$ and $b_{\min} = 1$ and measured the mean-squared error (MSE) as a function of sample size n. The estimator's convergence rate, however, depends on the binning scheme. To illustrate this dependency, we used the continuous MLE given by [6] with the raw observations and compared its convergence against that of our binned MLE applied to binned version of the same data, with the $c = 3$ (powers of three) and $c = 5$ (powers of five) logarithmic scheme. Figure 8 shows the results; as we expect for maximum likelihood, the MSE's convergence is $O(1/n)$, but with a constant that depends on the amount of information lost from binning: the more coarse the binning (larger c), the greater the statistical uncertainty at the same sample size.

A.2. Bounding the convergence rate. The Cramér-Rao bound [9, 28] implies that the variance of our estimator is bounded by the inverse Fisher information,

$$\mathrm{Var}(\hat{\alpha}) \geq 1\,/\mathcal{I}(\hat{\alpha}) \quad . \tag{A.5}$$

The Fisher information [39] is given by the curvature of the log-likelihood function around the maximum likelihood estimate; as with the MLE, there is no closed form expression for this function except in the case of a logarithmic binning scheme (Eq. (A.3)):

$$\begin{aligned} \mathcal{I}(\hat{\alpha}) &= -\mathbf{E}\left[\frac{\partial^2}{\partial\alpha^2}\mathcal{L}(X;\hat{\alpha})\right] \\ &= n\,c^{1+\alpha}\left(\frac{\ln c}{c - c^{\alpha}}\right)^2 \quad . \end{aligned} \tag{A.6}$$

Combining Eqs. (A.5) and (A.6), taking the square root and n as sufficiently large, we obtain a closed-form expression for the standard error of $\hat{\alpha}$:

$$\hat{\sigma} = \frac{c^{\hat{\alpha}} - c}{c^{(1+\hat{\alpha})/2}\ln c\sqrt{n}} \quad . \tag{A.7}$$

(Note: when n is small, e.g., $n \lesssim 50$ for $c = 2$, σ becomes positively biased.)

We may now show analytically exactly how much information is lost by different binning schemes. Suppose we have a sample n_1 and binning scheme c_1. Given a choice of α, how much larger a sample do we need in order to achieve the same statistical certainty in $\hat{\alpha}$ using a coarser scheme $c_2 > c_1$? Assuming n_1, c_1, and c_2 are known, we solve for the n_2 that yields equal statistical uncertainty for the two settings:

$$\begin{aligned} \left|\frac{1}{\mathcal{I}(\alpha)}\right|_{n_1,c_1} &= \left|\frac{1}{\mathcal{I}(\alpha)}\right|_{n_2,c_2} \\ n_1^{-1}{c_1}^{-\alpha-1}\left(\frac{c_1 - {c_1}^{\alpha}}{\ln c_1}\right)^2 &= n_2^{-1}{c_2}^{-\alpha-1}\left(\frac{c_2 - {c_2}^{\alpha}}{\ln c_2}\right)^2 \end{aligned}$$

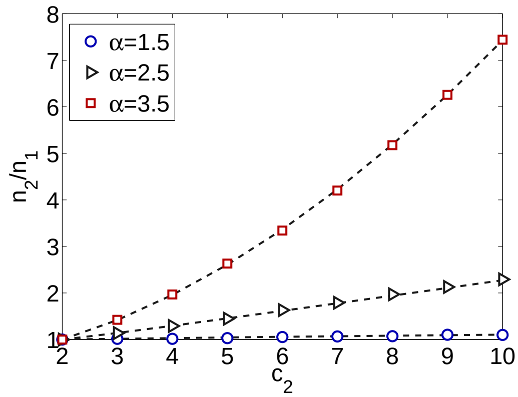

FIG 9. *The size of a data set required to achieve the same statistical certainty in α (constant MSE) when using a coarser binning scheme c_2, for several choices of α. The dashed lines indicate the values obtained analytically using Equation* (A.8).

Solving for n_2 yields

$$n_2 = \left(\left(\frac{c_1}{c_2} \right)^{1+\alpha} \left(\frac{\ln c_1}{\ln c_2} \right)^2 \left(\frac{c_2 - {c_2}^{\alpha}}{c_1 - {c_1}^{\alpha}} \right)^2 \right) n_1 \ , \tag{A.8}$$

in which the required sample size n_2 is some constant multiple of n_1, which is a complicated function of α and the two binning schemes.

Figure 9 illustrates how this constant varies with the coarseness of the second binning scheme c_2. For concreteness, we fix $c_1 = 2$ and show the constant's behavior for several choices of α and for schemes $c_2 \geq 2$. As expected, increasing the coarseness of the binning scheme decreases the information available for estimation, and the required sample size increases. Information loss also arises from variation in α. As α increases, the variance of the generating distribution decreases, and a given sample size will span fewer bins. The fundamental source of information loss for estimation is the loss of bins, i.e., the commingling of observations that are distinct, which may arise either from coarsening the binning scheme or from decreasing the variance of the generating distribution.

The information-loss effect is sufficiently strong that a powers-of-10 binning scheme can require nearly eight times as large a sample to obtain the same statistical accuracy in α, when $\alpha > 3$. Thus, if the option is available during the experimental design phase of a study, as fine a grained binning scheme as is possible should be used in collecting the data in order to maximize subsequent statistical accuracy.

APPENDIX B: LIKELIHOOD RATIO TEST

Let $B = (b_1, \ldots, b_k)$ be a fixed binning scheme, $H = (h_1, \ldots, h_k)$ be the observed bin counts within those bins, and p_1 and p_2 be the pdfs for two candidate distributions. The likelihoods of our bin counts under these distributions is

$$L_1 = \prod_{j=1}^{n} p_1(x_j) = \prod_{i=1}^{k} p_1(b_i)^{h_i}, \qquad L_2 = \prod_{j=1}^{n} p_2(x_j) = \prod_{i=1}^{k} p_1(b_i)^{h_i} \;, \tag{B.1}$$

where $p(b_i)$ is the probability that a non-binned value x_j is in the i^{th} bin, i.e., $\Pr(b_i < x_j < b_{i+1})$. Other than this slight change in the way we define our probability models, the likelihood ratio test has the usual structure.[6]

The likelihood ratio test statistic is the ratio of the likelihoods R, or equivalently, its logarithm $\mathcal{R}$:

$$\begin{aligned} \ln R &= \ln(L_1/L_2) \\ \mathcal{R} &= \sum_{j=1}^{n} \left[\ln p_1(x_j) - \ln p_2(x_j)\right] = \sum_{j=1}^{n} \left[\ell_j^{(1)} - \ell_j^{(2)}\right] \\ &= \sum_{i=1}^{k} \left[h_i \ell_i^{(1)} - h_i \ell_i^{(2)}\right] \;, \end{aligned} \tag{B.2}$$

where $\ell_i^{(z)} = \ln p_z(b_i)$ is the log-likelihood of a single observation being in bin b_i under distribution z.

By assumption, the raw observations x_j are independent, and so too are the differences $\ell_i^{(1)} - \ell_i^2$. Thus, by the central limit theorem, their sum $\mathcal{R}$ is normally distributed in limit of large n, with expected variance $n\,\sigma^2$, where σ^2 is the variance of a single term. In practice, we don't know the expected variance for a single term, but it may be approximated by the variance of the data:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{k} \left[h_i \left(\ell_i^{(1)} - \ell_i^{(2)}\right) - \left(\bar{\ell}^{(1)} - \bar{\ell}^{(2)}\right)\right]^2 \;, \tag{B.3}$$

where

$$\bar{\ell}^{(1)} = \frac{1}{n} \sum_{i=1}^{k} h_i \ell_i^{(1)}, \qquad \bar{\ell}^{(2)} = \frac{1}{n} \sum_{i=1}^{k} h_i \ell_i^{(2)} \;. \tag{B.4}$$

[6]There are some technical subtleties in the rigorous proof of our results here. However, for the distributions we consider, Vuong [37] has shown that our construction holds provided that p_1 and p_2 come from distinct, non-nested families of distributions and the estimation is done by maximizing the likelihood within each family.

One advantage of the likelihood ratio test is its ability to refuse to choose one model over the other, which occurs when the true expected value of the log-likelihood is in fact zero. In this case, the sign of $\mathcal{R}$ is due only to fluctuations and should not be trusted to indicate which model is preferred. The probability that the observed $\mathcal{R}$ has magnitude at least as large as the observed value $|\mathcal{R}|$ is

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}}\left[\int_{-\infty}^{-|\mathcal{R}|} \mathrm{e}^{-t^2/2n\sigma^2}\,\mathrm{d}t + \int_{|\mathcal{R}|}^{\infty} \mathrm{e}^{-t^2/2n\sigma^2}\,\mathrm{d}t\right]$$

$$= \mathrm{erfc}(|\mathcal{R}|/\sqrt{2n}\sigma) \ , \tag{B.5}$$

where σ is given by Eq. (B.3) and erfc is the complementary error function, defined as

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t \ . \tag{B.6}$$

Eq. (B.5) thus estimates the probability that we measured a particular value of $\mathcal{R}$ when in fact its true value is zero. A large p-value (say, $p > 0.1$) indicates that the sign of $\mathcal{R}$ is ambiguous and that our test is unable to identify which model is the better fit to the data. A low p-value (say, $p < 0.1$) implies that the observed value $\mathcal{R}$ is unlikely to be due to chance alone, and thus its sign indicates which model is a better fit.

This construction changes when our hypotheses are nested, as in the case of the power law and power law with exponential cutoff. If the true generating distribution lies in the smaller family, e.g., the power law, then fits of both families converge to the true distribution as n becomes large, $|\mathcal{R}| \to 0$ and so does σ. As a result, the p-value given in Eq. (B.5) tends to 0/0 and its distribution does not obey the simple central limit theorem. A more careful analysis shows that, in this situation, $\mathcal{R}$ adopts a χ^2-distribution as n becomes large [41], which allows us to correctly calculate a p-value. If this p-value is small, the smaller family can be ruled out. Otherwise, the best interpretation is that there is no statistical evidence that the larger family is needed to fit the data, although neither can be ruled out. For a detailed discussion, see [37].

Finally, we remind the reader that the results of a likelihood ratio test (indeed, any rigorous model comparison technique) do not indicate that the favored model is itself a good fit to the data, only that it is a less terrible fit than the unfavored model. For example, when comparing the power-law and exponential distributions for heavy-tailed data, the power law will typically be favored even when the data are not power-law distributed.

REFERENCES

[1] B. C. Arnold. *Pareto Distributions*. International Cooperative Publishing House, Fairland, MD, 1983.
[2] V. Asal and R. K. Rethemeyer. The nature of the beast: Organizational structures and the lethality of terrorist attacks. *The Journal of Politics*, 70(2):437–449, 2008.
[3] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and Asymptotics*. Chapman and Hall, London, 1995.
[4] L. Breiman, C. J. Stone, and C. Kooperberg. Robust confidence bounds for extreme upper quantiles. *J. Statistical Computation and Simulation*, 37:127–149, 1990.
[5] A. Chapuis and T. Tetzlaff. The variability of tidewater-glacier calving: origin of event-size and interval distributions. e-print, arXiv:1205.1640, 2012.
[6] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.
[7] A. Clauset and R. Woodard. Estimating the historical and future probabilities of large terrorist events. Preprint, 2012.
[8] A. Clauset, M. Young, and K. S. Gleditsch. On the frequency of severe terrorist events. *Journal of Conflict Resolution*, 51(1):58–87, 2007.
[9] H. Cramér. A contribution to the theory of statistical estimation. *Skand. Ȧktuaries Tidskrift*, 29:458–463, 1946.
[10] J. Danielsson, L. de Hann, L. Peng, and C. G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76:226–248, 2001.
[11] A. L. M. Dekkers and L. de Hann. Optimal choice of sample fraction in extreme-value estimation. *Journal of Multivariate Analysis*, 47:173–195, 1993.
[12] H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75:149–172, 1998.
[13] X. Gabaix. Power laws in economics and finance. *Annual Review of Economics*, 1:255–293, 2009.
[14] K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási. The human disease network. *Proc. Nat. Acad. Sci. USA*, 104(21):8685–8690, May 2007.
[15] P. D. Grünwald. *The Minimum Length Description Principle*. MIT Press, Cambridge, MA, 2007.
[16] M. S. Hancock and J. H. Jones. Likelihood-based inference for stochastic models of sexual network evolution. *Theoretical Population Biology*, 65:413–422, 2004.
[17] Heritage Provider Network. *Health Heritage Prize Data Files, HHP_release3*. Available at http://bit.ly/wG8Psl, 2012.
[18] Y. Ijiri and H. A. Simon. *Skew distributions and the sizes of business firms*. North-Holland, Amsterdam, 1977.
[19] B. Jarvinen, C. Neumann, and M. A. S. Davis. *NHC Data Archive*. National Hurricane Center, Available at http://1.usa.gov/cCcwTg, 2012.
[20] R. E. Kass and A. E. Raftery. Bayes factors. *J. Amer. Statist. Assoc.*, 90:773–795, 1994.
[21] M. Mitzenmacher. A brief history of generative models of power law and lognormal distributions. *Internet Math.*, 1:226–251, 2004.
[22] M .Mitzenmacher. The future of power law research. *Internet Math.*, 2:525–534, 2006.
[23] M. E. J. Newman. Power laws, Pareto distributions and Zipfs's law. *Contemporary Physics*, 46:323–351, 2005.

[24] P. T. Nicholls. Estimation of Zipf parameters. *J. Am. Soc. Information Sci.*, 40:443–445, 1989.
[25] Orphanet Report Series, Rare Diseases collection. *Prevalence of rare diseases: Bibliographic data*. Available at http://bit.ly/MezSZ6, November 2011.
[26] J. Persing and M. T. Montgomery. Hurricane superintensity. *J. Atmospheric Sciences*, 60:2349–2371, 2003.
[27] 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, 2nd edition, 1992.
[28] C. R. Rao. Minimum variance and the estimation of several parameters. *Proc. Cambridge Phil. Soc.*, 43:280–283, 1946.
[29] W. J. Reed and B. D. Hughes. From gene families and genera to income and Internet file sizes: Why power laws are so common in nature. *Physical Review E*, 66:067103, 2002.
[30] R. D. Reiss and M. Thomas. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser, Basel, Switzerland, 2007.
[31] Lewis F. Richardson. *Statistics of Deadly Quarrels*. The Boxwood Press, Pittsburgh, 1960.
[32] K. Shinokazi, K. Yoda, K. Hozumi, and T. Kira. A quantitative analysis of plant form–The Pipe Model Theory II: Further evidence of the theory and its application in forest ecology. *Japanese Journal of Ecology*, 14(4):133–139, 1964.
[33] D. Sornette. *Critical Phenomena in Natural Sciences*. Springer, Berlin, 2nd edition, 2006.
[34] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society Series B*, 36:111–133, 1974.
[35] Storm Prediction Center. *Severe Weather Database Files (1950–2011)*. Available at http://1.usa.gov/Lj7cC9, 2011.
[36] M. P. H. Stumpf and M. A. Porter. Critical truths about power laws. *Science*, 335:665–666, 2012.
[37] Q. H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.
[38] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer-Verlag, Berlin, 2003.
[39] L. Wasserman. *All of non-parametric statistics*. Springer, Berlin, 2006.
[40] G. B. West, B. J. Enquist, and J. H. Brown. A general quantitative theory of forest structure and dynamics. *Proc. Nat. Acad. Sci. USA*, 106(17):7040–7045, 2009.
[41] S. S. Wilks. The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, 9:60–62, 1938.
[42] World Glacier Monitoring Service and National Snow and Ice Data Center. *World Glacier Inventory*. Available at http://bit.ly/MhLdt6, 2012.
[43] K. Yamamoto and S. Kobayashi. Analysis of crown structure based on the pipe model theory. *Journal of the Japanese Forestry Society*, 75:445–448, 1993.

Department of Computer Science
University of Colorado, Boulder
E-mail: yogesh.virkar@colorado.edu

Department of Computer Science and
the BioFrontiers Institute,
University of Colorado, Boulder, and
Santa Fe Institute
E-mail: aaron.clauset@colorado.edu