SIAM REVIEW
Vol. 51, No. 4, pp. 661–703

Power-Law Distributions in Empirical Data*

Aaron Clauset[†]
Cosma Rohilla Shalizi[‡]
M. E. J. Newman[§]

Abstract. Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena. Unfortunately, the detection and characterization of power laws is complicated by the large fluctuations that occur in the tail of the distribution—the part of the distribution representing large but rare events—and by the difficulty of identifying the range over which power-law behavior holds. Commonly used methods for analyzing power-law data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for power-law distributions, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a power law at all. Here we present a principled statistical framework for discerning and quantifying power-law behavior in empirical data. Our approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov–Smirnov (KS) statistic and likelihood ratios. We evaluate the effectiveness of the approach with tests on synthetic data and give critical comparisons to previous approaches. We also apply the proposed methods to twenty-four real-world data sets from a range of different disciplines, each of which has been conjectured to follow a power-law distribution. In some cases we find these conjectures to be consistent with the data, while in others the power law is ruled out.

Key words. power-law distributions, Pareto, Zipf, maximum likelihood, heavy-tailed distributions, likelihood ratio test, model selection

AMS subject classifications. 62-07, 62P99, 65C05, 62F99

DOI. 10.1137/070710111

1. Introduction.

Many empirical quantities cluster around a typical value. The speeds of cars on a highway, the weights of apples in a store, air pressure, sea level, the temperature in New York at noon on a midsummer's day: all of these things vary somewhat, but their distributions place a negligible amount of probability far from the typical value, making the typical value representative of most observations. For instance, it is a useful statement to say that an adult male American is about 180cm tall because no one deviates very far from this height. Even the largest deviations, which are exceptionally rare, are still only about a factor of two from the mean in

*Received by the editors December 2, 2007; accepted for publication (in revised form) February 2, 2009; published electronically November 6, 2009. This work was supported in part by the Santa Fe Institute (AC) and by grants from the James S. McDonnell Foundation (CRS and MEJN) and the National Science Foundation (MEJN).

http://www.siam.org/journals/sirev/51-4/71011.html

[†] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, and Department of Computer Science, University of New Mexico, Albuquerque, NM 87131.

[‡] Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

[§] Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109.

either direction and hence the distribution can be well characterized by quoting just its mean and standard deviation.

Not all distributions fit this pattern, however, and while those that do not are often considered problematic or defective for just that reason, they are at the same time some of the most interesting of all scientific observations. The fact that they cannot be characterized as simply as other measurements is often a sign of complex underlying processes that merit further study.

Among such distributions, the *power law* has attracted particular attention over the years for its mathematical properties, which sometimes lead to surprising physical consequences, and for its appearance in a diverse range of natural and man-made phenomena. The populations of cities, the intensities of earthquakes, and the sizes of power outages, for example, are all thought to follow power-law distributions. Quantities such as these are not well characterized by their typical or average values. For instance, according to the 2000 U.S. Census, the average population of a city, town, or village in the United States is 8226. But this statement is not a useful one for most purposes because a significant fraction of the total population lives in cities (New York, Los Angeles, etc.) whose populations are larger by several orders of magnitude. Extensive discussions of this and other properties of power laws can be found in the reviews by Mitzenmacher [40], Newman [43], and Sornette [55], and references therein.

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

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

where α is a constant parameter of the distribution known as the *exponent* or *scaling parameter*. The scaling parameter typically lies in the range $2 < \alpha < 3$, although there are occasional exceptions.

In practice, few empirical phenomena obey power laws for all values of x. More often the power law applies only for values greater than some minimum $x_{\min}$. In such cases we say that the *tail* of the distribution follows a power law.

In this article we address a recurring issue in the scientific literature, the question of how to recognize a power law when we see one. In practice, we can rarely, if ever, be certain that an observed quantity is drawn from a power-law distribution. The most we can say is that our observations are consistent with the hypothesis that x is drawn from a distribution of the form of (1.1). In some cases we may also be able to rule out some other competing hypotheses. In this paper we describe in detail a set of statistical techniques that allow one to reach conclusions like these, as well as methods for calculating the parameters of power laws when we find them. Many of the methods we describe have been discussed previously; our goal here is to bring them together to create a complete procedure for the analysis of power-law data. A short description summarizing this procedure is given in Box 1. Software implementing it is also available online.[1]

Practicing what we preach, we also apply our methods to a large number of data sets describing observations of real-world phenomena that have at one time or another been claimed to follow power laws. In the process, we demonstrate that several of them cannot reasonably be considered to follow power laws, while for others the power-law hypothesis appears to be a good one, or at least is not firmly ruled out.

[1] See http://www.santafe.edu/˜aaronc/powerlaws/.

Box 1: Recipe for analyzing power-law distributed data

This paper contains much technical detail. In broad outline, however, the recipe we propose for the analysis of power-law data is straightforward and goes as follows.

1. Estimate the parameters $x_{\min}$ and α of the power-law model using the methods described in section 3.
2. Calculate the goodness-of-fit between the data and the power law using the method described in section 4. If the resulting p-value is greater than 0.1, the power law is a plausible hypothesis for the data, otherwise it is rejected.
3. Compare the power law with alternative hypotheses via a likelihood ratio test, as described in section 5. For each alternative, if the calculated likelihood ratio is significantly different from zero, then its sign indicates whether or not the alternative is favored over the power-law model.

Step 3, the likelihood ratio test for alternative hypotheses, could in principle be replaced with any of several other established and statistically principled approaches for model comparison, such as a fully Bayesian approach [31], a cross-validation approach [58], or a minimum description length approach [20], although these methods are not described here.

2. Definitions.

We begin our discussion of the analysis of power-law distributed data with some brief definitions of the basic quantities involved.

Power-law distributions come in two basic flavors: continuous distributions governing continuous real numbers and discrete distributions where the quantity of interest can take only a discrete set of values, typically positive integers.

Let x represent the quantity in whose distribution we are interested. A continuous power-law distribution is one described by a probability density $p(x)$ such that

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

where X is the observed value and C is a normalization constant. Clearly this density diverges as $x \to 0$ so (2.1) cannot hold for all $x \ge 0$; there must be some lower bound to the power-law behavior. We will denote this bound by $x_{\min}$. Then, provided $\alpha > 1$, it is straightforward to calculate the normalizing constant and we find that

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

In the discrete case, x can take only a discrete set of values. In this paper we consider only the case of integer values with a probability distribution of the form

$$p(x) = \Pr(X = x) = Cx^{-\alpha}\,. \tag{2.3}$$

Again this distribution diverges at zero, so there must be a lower bound $x_{\min} > 0$ on the power-law behavior. Calculating the normalizing constant, we then find that

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}\,, \tag{2.4}$$

Table 1 *Definition of the power-law distribution and several other common statistical distributions. For each distribution 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 or $\sum_{x=x_{\min}}^{\infty} Cf(x) = 1$ for the discrete case.*

	Name	Distribution $p(x) = Cf(x)$ $f(x)$	 C
Continuous	Power law	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$
	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	$\frac{1}{x}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right]$	$\sqrt{\frac{2}{\pi\sigma^2}}\left[\mathrm{erfc}\left(\frac{\ln x_{\min}-\mu}{\sqrt{2}\sigma}\right)\right]^{-1}$
Discrete	Power law	$x^{-\alpha}$	$1/\zeta(\alpha,x_{\min})$
	Yule distribution	$\frac{\Gamma(x)}{\Gamma(x+\alpha)}$	$(\alpha-1)\frac{\Gamma(x_{\min}+\alpha-1)}{\Gamma(x_{\min})}$
	Exponential	$\mathrm{e}^{-\lambda x}$	$(1-\mathrm{e}^{-\lambda})\,\mathrm{e}^{\lambda x_{\min}}$
	Poisson	$\mu^x/x!$	$\left[\mathrm{e}^{\mu}-\sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

where

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty}(n+x_{\min})^{-\alpha} \tag{2.5}$$

is the generalized or Hurwitz zeta function. Table 1 summarizes the basic functional forms and normalization constants for these and several other distributions that will be useful.

In many cases it is useful to consider also the complementary cumulative distribution function or CDF of a power-law distributed variable, which we denote $P(x)$ and which for both continuous and discrete cases is defined to be $P(x) = \Pr(X \geq x)$. For instance, in the continuous case,

$$P(x) = \int_{x}^{\infty} p(x')\,\mathrm{d}x' = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1}. \tag{2.6}$$

In the discrete case,

$$P(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{\min})}. \tag{2.7}$$

Because formulas for continuous distributions, such as (2.2), tend to be simpler than those for discrete distributions, it is common to approximate discrete power-law behavior with its continuous counterpart for the sake of mathematical convenience. But a word of caution is in order: there are several different ways to approximate a discrete power law by a continuous one and though some of them give reasonable results, others do not. One relatively reliable method is to treat an integer power law as if the values of x were generated from a continuous power law then rounded to the nearest integer. This approach gives quite accurate results in many applications. Other approximations, however, such as truncating (rounding down) or simply assuming that the probabilities of generation of integer values in the discrete and continuous cases are proportional, give poor results and should be avoided.

Where appropriate we will discuss the use of continuous approximations for the discrete power law in the sections that follow, particularly in section 3 on the estimation of best-fit values for the scaling parameter from observational data and in Appendix D on the generation of power-law distributed random numbers.

3. Fitting Power Laws to Empirical Data.

We turn now to the first of the main goals of this paper, the correct fitting of power-law forms to empirical distributions. Studies of empirical distributions that follow power laws usually give some estimate of the scaling parameter α and occasionally also of the lower bound on the scaling region $x_{\min}$. The tool most often used for this task is the simple histogram. Taking the logarithm of both sides of (1.1), we see that the power-law distribution obeys $\ln p(x) = \alpha \ln x + \text{constant}$, implying that it follows a straight line on a doubly logarithmic plot. A common way to probe for power-law behavior, therefore, is to measure the quantity of interest x, construct a histogram representing its frequency distribution, and plot that histogram on doubly logarithmic axes. If in so doing one discovers a distribution that falls approximately on a straight line, then one can, if feeling particularly bold, assert that the distribution follows a power law, with a scaling parameter α given by the absolute slope of the straight line. Typically this slope is extracted by performing a least-squares linear regression on the logarithm of the histogram. This procedure dates back to Pareto's work on the distribution of wealth at the close of the 19th century [6].

Unfortunately, this method and other variations on the same theme generate significant systematic errors under relatively common conditions, as discussed in Appendix A, and as a consequence the results they give cannot be trusted. In this section we describe a generally accurate method for estimating the parameters of a power-law distribution. In section 4 we study the equally important question of how to determine whether a given data set really does follow a power law at all.

3.1. Estimating the Scaling Parameter.

First, let us consider the estimation of the scaling parameter α. Estimating α correctly requires, as we will see, a value for the lower bound $x_{\min}$ of power-law behavior in the data. For the moment, let us assume that this value is known. In cases where it is unknown, we can estimate it from the data as well, and we will consider methods for doing this in section 3.3.

The method of choice for fitting parametrized models such as power-law distributions to observed data is the *method of maximum likelihood*, which provably gives accurate parameter estimates in the limit of large sample size [63, 7]. Assuming that our data are drawn from a distribution that follows a power law exactly for $x \geq x_{\min}$, we can derive maximum likelihood estimators (MLEs) of the scaling parameter for both the discrete and continuous cases. Details of the derivations are given in Appendix B; here our focus is on their use.

The MLE for the continuous case is [42]

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1}, \tag{3.1}$$

where x_i, $i = 1, \ldots, n$, are the observed values of x such that $x_i \geq x_{\min}$. Here and elsewhere we use "hatted" symbols such as $\hat{\alpha}$ to denote estimates derived from data; hatless symbols denote the true values, which are often unknown in practice.

Equation (3.1) is equivalent to the well-known Hill estimator [24], which is known to be asymptotically normal [22] and consistent [37] (i.e., $\hat{\alpha} \to \alpha$ in the limit of large n). The standard error on $\hat{\alpha}$, which is derived from the width of the likelihood maximum, is

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \mathrm{O}(1/n)\,, \tag{3.2}$$

where the higher-order correction is positive; see Appendix B of this paper or any of the references [42], [43], or [66].

(We assume in these calculations that $\alpha > 1$, since distributions with $\alpha \leq 1$ are not normalizable and hence cannot occur in nature. It is possible for a probability distribution to go as $x^{-\alpha}$ with $\alpha \leq 1$ if the range of x is bounded above by some cutoff, but different MLEs are needed to fit such a distribution.)

The MLE for the case where x is a discrete integer variable is less straightforward. Reference [51] and more recently [19] treated the special case $x_{\min} = 1$, showing that the appropriate estimator for α is given by the solution to the transcendental equation

$$\frac{\zeta'(\hat{\alpha})}{\zeta(\hat{\alpha})} = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i\,. \tag{3.3}$$

When $x_{\min} > 1$, a similar equation holds, but with the zeta functions replaced by generalized zetas [6, 8, 11],

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln x_i\,, \tag{3.4}$$

where the prime denotes differentiation with respect to the first argument. In practice, evaluation of $\hat{\alpha}$ requires us to solve this equation numerically. Alternatively, one can estimate α by direct numerical maximization of the likelihood function itself, or equivalently of its logarithm (which is usually simpler):

$$\mathcal{L}(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i\,. \tag{3.5}$$

To find an estimate for the standard error on $\hat{\alpha}$ in the discrete case, we make a quadratic approximation to the log-likelihood at its maximum and take the standard deviation of the resulting Gaussian form for the likelihood as our error estimate (an approach justified by general theorems on the large-sample-size behavior of maximum likelihood estimates—see, for example, Theorem B.3 of Appendix B). The result is

$$\sigma = \frac{1}{\sqrt{n\left[\dfrac{\zeta''(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} - \left(\dfrac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})}\right)^2\right]}}\,, \tag{3.6}$$

which is straightforward to evaluate once we have $\hat{\alpha}$. Alternatively, (3.2) yields roughly similar results for reasonably large n and $x_{\min}$.

Although there is no exact closed-form expression for $\hat{\alpha}$ in the discrete case, an approximate expression can be derived using the approach mentioned in section 2 in which true power-law distributed integers are approximated as continuous reals rounded to the nearest integer. The details of the derivation are given in Appendix B.

The result is

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - \frac{1}{2}}\right]^{-1}. \tag{3.7}$$

This expression is considerably easier to evaluate than the exact discrete MLE and can be useful in cases where high accuracy is not needed. The size of the bias introduced by the approximation is discussed in Appendix B. In practice, this estimator gives quite good results; in our own experiments we have found it to give results accurate to about 1% or better provided $x_{\min} \gtrsim 6$. An estimate of the statistical error on $\hat{\alpha}$ (which is quite separate from the systematic error introduced by the approximation) can be calculated by employing (3.2) again.

Another approach taken by some authors is simply to pretend that discrete data are in fact continuous and then use the MLE for continuous data, (3.1), to calculate $\hat{\alpha}$. This approach, however, gives significantly less accurate values of $\hat{\alpha}$ than (3.7) and, given that it is no easier to implement, we see no reason to use it in any circumstances.[2]

3.2. Performance of Scaling Parameter Estimators.

To demonstrate the working of the estimators described above, we now test their ability to extract the known scaling parameters of synthetic power-law data. Note that in practical situations we usually do not know a priori, as we do in the calculations of this section, that our data are power-law distributed. In that case, our MLEs will give us no warning that our fits are wrong: they tell us only the best fit to the power-law form, not whether the power law is in fact a good model for the data. Other methods are needed to address the latter question, and are discussed in sections 4 and 5.

Using methods described in Appendix D, we have generated two sets of power-law distributed data, one continuous and one discrete, with $\alpha = 2.5$, $x_{\min} = 1$, and $n = 10\,000$ in each case. Applying our MLEs to these data we calculate that $\hat{\alpha} = 2.50(2)$ for the continuous case and $\hat{\alpha} = 2.49(2)$ for the discrete case. (Values in parentheses indicate the uncertainty in the final digit, calculated from (3.2) and (3.6).) These estimates agree well with the known true scaling parameter from which the data were generated. Figure 1 shows the distributions of the two data sets along with fits using the estimated parameters. (In this and all subsequent such plots, we show not the probability density function (PDF), but the complementary CDF $P(x)$. Generally, the visual form of the CDF is more robust than that of the PDF against fluctuations due to finite sample sizes, particularly in the tail of the distribution.)

In Table 2 we compare the results given by the MLEs to estimates of the scaling parameter made using several alternative methods based on linear regression: a straight-line fit to the slope of a log-transformed histogram, a fit to the slope of a histogram with "logarithmic bins" (bins whose width increases in proportion to x, thereby reducing fluctuations in the tail of the histogram), a fit to the slope of the CDF calculated with constant width bins, and a fit to the slope of the CDF calculated without any bins (also called a "rank-frequency plot"—see [43]). As the table shows, the MLEs give the best results, while the regression methods all give significantly biased values, except perhaps for the fits to the CDF, which produce biased estimates in the discrete case but do reasonably well in the continuous case. Moreover, in each

[2] The error involved can be shown to decay as $\mathrm{O}(x_{\min}^{-1})$, while the error on (3.7) decays much faster, as $\mathrm{O}(x_{\min}^{-2})$. In our own experiments we have found that for typical values of α we need $x_{\min} \gtrsim 100$ before (3.1) becomes accurate to about 1%, as compared to $x_{\min} \gtrsim 6$ for (3.7).

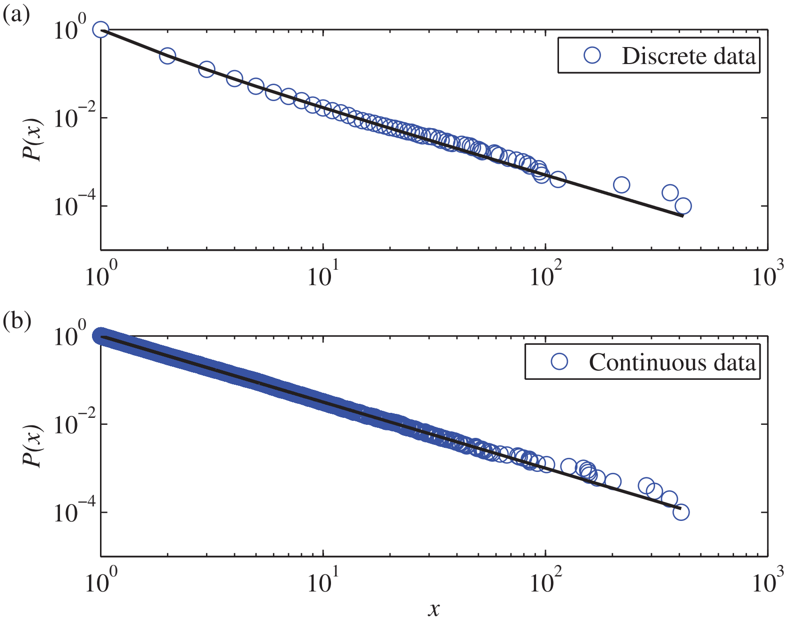

Fig. 1 *Points represent the CDFs $P(x)$ for synthetic data sets distributed according to* (a) *a discrete power law and* (b) *a continuous power law, both with $\alpha = 2.5$ and $x_{\min} = 1$. Solid lines represent best fits to the data using the methods described in the text.*

Table 2 *Estimates of the scaling parameter α using various estimators for discrete and continuous synthetic data with $\alpha = 2.5$, $x_{\min} = 1$, and $n = 10\,000$ data points. LS denotes a least-squares fit to the logarithm of the probability. For the continuous data, the PDF was computed in two different ways, using bins of constant width* 0.1 *and using up to* 500 *bins of exponentially increasing width (so-called "logarithmic binning"). The CDF was also calculated in two ways, as the cumulation of the fixed-width histogram and as a standard rank-frequency function. In applying the discrete MLE to the continuous data, the noninteger part of each measurement was discarded. Accurate estimates are shown in* ***bold****.*

Method	Notes	est. α (Discrete)	est. α (Continuous)
LS + PDF	const. width	1.5(1)	1.39(5)
LS + CDF	const. width	2.37(2)	2.480(4)
LS + PDF	log. width	1.5(1)	1.19(2)
LS + CDF	rank-freq.	2.570(6)	2.4869(3)
cont. MLE	–	4.46(3)	**2.50(2)**
disc. MLE	–	**2.49(2)**	2.19(1)

case where the estimate is biased, the corresponding error estimate gives no warning of the bias: there is nothing to alert unwary experimenters to the fact that their results are substantially incorrect. Figure 2 extends these results graphically by showing how the estimators fare as a function of the true α for a large selection of synthetic data sets with $n = 10\,000$ observations each.

Finally, we note that the MLEs are only guaranteed to be unbiased in the asymptotic limit of large sample size, $n \to \infty$. For finite data sets, biases are present but decay as $\mathrm{O}(n^{-1})$ for any choice of $x_{\min}$ (see Appendix B and Figure 10). For very small data sets, such biases can be significant but in most practical situations

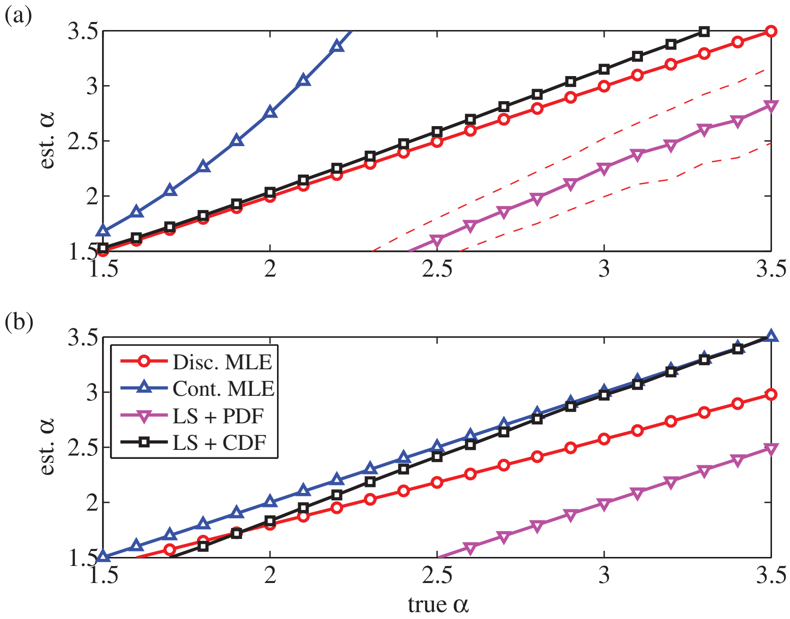

Fig. 2 *Values of the scaling parameter estimated using four of the methods of Table* 2 *(we omit the methods based on logarithmic bins for the PDF and constant width bins for the CDF) for* $n = 10\,000$ *observations drawn from* (a) *discrete and* (b) *continuous power-law distributions with* $x_{\min} = 1$. *We omit error bars where they are smaller than the symbol size. Clearly, only the discrete MLE is accurate for discrete data, and the continuous MLE for continuous data.*

they can be ignored because they are much smaller than the statistical error of the estimator, which decays as $\mathrm{O}(n^{-1/2})$. Our experience suggests that $n \gtrsim 50$ is a reasonable rule of thumb for extracting reliable parameter estimates. For the examples shown in Figure 10 this gives estimates of α accurate to about 1%. Data sets smaller than this should be treated with caution. Note, however, that there are more important reasons to treat small data sets with caution. Namely, it is difficult to rule out alternative fits to such data, even when they are truly power-law distributed, and conversely the power-law form may appear to be a good fit even when the data are drawn from a non-power-law distribution. We address these issues in sections 4 and 5.

3.3. Estimating the Lower Bound on Power-Law Behavior. As we have said above it is normally the case that empirical data, if they follow a power-law distribution at all, do so only for values of x above some lower bound $x_{\min}$. Before calculating our estimate of the scaling parameter α, therefore, we need to first discard all samples below this point so that we are left with only those for which the power-law model is valid. Thus, if we wish our estimate of α to be accurate, we will also need an accurate method for estimating $x_{\min}$. If we choose too low a value for $x_{\min}$, we will get a biased estimate of the scaling parameter since we will be attempting to fit a power-law model to non-power-law data. On the other hand, if we choose too high a value for $x_{\min}$, we are effectively throwing away legitimate data points $x_i < \hat{x}_{\min}$, which increases both the statistical error on the scaling parameter and the bias from finite size effects.

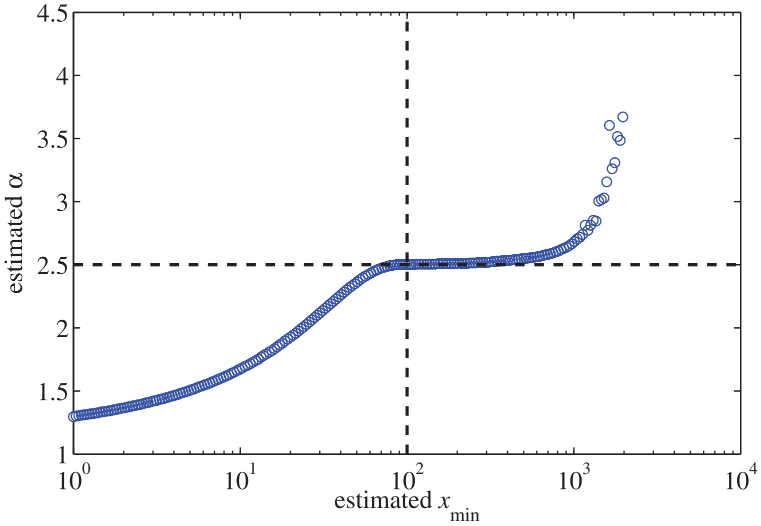

Fig. 3 *Mean of the MLE for the scaling parameter for* 5000 *samples drawn from the test distribution,* (3.10)*, with* $\alpha = 2.5$*,* $x_{\min} = 100$*, and* $n = 2500$*, plotted as a function of the value assumed for* $x_{\min}$*. Statistical errors are smaller than the data points in all cases.*

The importance of using the correct value for $x_{\min}$ is demonstrated in Figure 3, which shows the maximum likelihood value $\hat{\alpha}$ of the scaling parameter averaged over 5000 data sets of $n = 2500$ samples, each drawn from the continuous form of (3.10) with $\alpha = 2.5$, as a function of the assumed value of $x_{\min}$, where the true value is 100. As the figure shows, the MLE gives accurate answers when $x_{\min}$ is chosen exactly equal to the true value, but deviates rapidly below this point (because the distribution deviates from power law) and more slowly above (because of dwindling sample size). It would probably be acceptable in this case for $x_{\min}$ to err a little on the high side (though not too much), but estimates that are too low could have severe consequences.

The most common ways of choosing $\hat{x}_{\min}$ are either to estimate visually the point beyond which the PDF or CDF of the distribution becomes roughly straight on a log-log plot, or to plot $\hat{\alpha}$ (or a related quantity) as a function of $\hat{x}_{\min}$ and identify a point beyond which the value appears relatively stable. But these approaches are clearly subjective and can be sensitive to noise or fluctuations in the tail of the distribution—see [57] and references therein. A more objective and principled approach is desirable. Here we review two such methods, one that is specific to discrete data and is based on a so-called marginal likelihood, and one that works for either discrete or continuous data and is based on minimizing the "distance" between the power-law model and the empirical data.

The first approach, put forward by Handcock and Jones [23], uses a generalized model to represent all of the observed data, both above and below $\hat{x}_{\min}$. Above $\hat{x}_{\min}$ the data are modeled by the standard discrete power-law distribution of (2.4); below $\hat{x}_{\min}$ each of the $\hat{x}_{\min} - 1$ discrete values of x are modeled by a separate probability $p_k = \Pr(X = k)$ for $1 \leq k < \hat{x}_{\min}$ (or whatever range is appropriate for the problem at hand). The MLE for p_k is simply the fraction of observations with value k. The task then is to find the value for $\hat{x}_{\min}$ such that this model best fits the observed data. One cannot, however, fit such a model to the data directly within the maximum likelihood framework because the number of model parameters is not fixed: it is equal to $x_{\min}$.[3] In this kind of situation, one can always achieve a higher likelihood by increasing the number of parameters, thus making the model more flexible, so the maximum likelihood is always achieved for $x_{\min} \to \infty$. A standard (Bayesian) approach in such cases is instead to maximize the *marginal likelihood* (also called the *evidence*) [29, 34], i.e., the likelihood of the data given the number of model parameters, integrated over the parameters' possible values. Unfortunately, the integral cannot usually be performed analytically, but one can employ a Laplace or steepest-descent approximation in which the log-likelihood is expanded to leading (i.e., quadratic) order about its maximum and the resulting Gaussian integral is carried out to yield an expression in terms of the value at the maximum and the determinant of the appropriate Hessian matrix [60]. Schwarz [50] showed that the terms involving the Hessian can be simplified for large n yielding an approximation to the log marginal likelihood of the form

$$\ln \Pr(x|x_{\min}) \simeq \mathcal{L} - \tfrac{1}{2} x_{\min} \ln n \,, \tag{3.8}$$

where $\mathcal{L}$ is the value of the conventional log-likelihood at its maximum. This type of approximation is known as a *Bayesian information criterion* or BIC. The maximum of the BIC with respect to $x_{\min}$ then gives the estimated value $\hat{x}_{\min}$.[4]

This method works well under some circumstances, but can also present difficulties. In particular, the assumption that $x_{\min} - 1$ parameters are needed to model the data below $x_{\min}$ may be excessive: in many cases the distribution below $x_{\min}$, while not following a power law, can nonetheless be represented well by a model with a much smaller number of parameters. In this case, the BIC tends to underestimate the value of $x_{\min}$ and this could result in biases on the subsequently calculated value of the scaling parameter. More importantly, it is also unclear how the BIC (and similar methods) can be generalized to the case of continuous data, for which there is no obvious choice for how many parameters are needed to represent the empirical distribution below $x_{\min}$.

Our second approach for estimating $x_{\min}$, proposed by Clauset, Young, and Gleditsch [11], can be applied to both discrete and continuous data. The fundamental idea behind this method is simple: we choose the value of $\hat{x}_{\min}$ that makes the probability distributions of the measured data and the best-fit power-law model as similar as possible above $\hat{x}_{\min}$. In general, if we choose $\hat{x}_{\min}$ higher than the true value $x_{\min}$, then we are effectively reducing the size of our data set, which will make the probability distributions a poorer match because of statistical fluctuation. Conversely, if we choose $\hat{x}_{\min}$ smaller than the true $x_{\min}$, the distributions will differ because of the fundamental difference between the data and model by which we are describing it. In between lies our best estimate.

There are a variety of measures for quantifying the distance between two probability distributions, but for nonnormal data the commonest is the Kolmogorov–Smirnov or KS statistic [46], which is simply the maximum distance between the CDFs of the

[3] There is one parameter for each of the p_k plus the scaling parameter of the power law. The normalization constant does not count as a parameter because it is fixed once the values of the other parameters are chosen, and $x_{\min}$ does not count as a parameter because we know its value automatically once we are given a list of the other parameters—it is just the length of that list.

[4] The same procedure of reducing the likelihood by $\frac{1}{2} \ln n$ times the number of model parameters to avoid overfitting can also be justified on non-Bayesian grounds for many model selection problems.

data and the fitted model:

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \ . \tag{3.9}$$

Here $S(x)$ is the CDF of the data for the observations with value at least $x_{\min}$, and $P(x)$ is the CDF for the power-law model that best fits the data in the region $x \geq x_{\min}$. Our estimate $\hat{x}_{\min}$ is then the value of $x_{\min}$ that minimizes D.[5]

There is good reason to expect this method to produce reasonable results. Note in particular that for right-skewed data of the kind we consider here the method is especially sensitive to slight deviations of the data from the power-law model around $x_{\min}$ because most of the data, and hence most of the dynamic range of the CDF, lie in this region. In practice, as we show in the following section, the method appears to give excellent results and generally performs better than the BIC approach.

3.4. Tests of Estimates for the Lower Bound. As with our MLEs for the scaling parameter, we test our two methods for estimating $x_{\min}$ by generating synthetic data and examining the methods' ability to recover the known value of $x_{\min}$. For the tests presented here we use synthetic data drawn from a distribution with the form

$$p(x) = \begin{cases} C(x/x_{\min})^{-\alpha} & \text{for } x \geq x_{\min} \ , \\ C\mathrm{e}^{-\alpha(x/x_{\min}-1)} & \text{for } x < x_{\min} \ , \end{cases} \tag{3.10}$$

with $\alpha = 2.5$. This distribution follows a power law at $x_{\min}$ and above but an exponential below. Furthermore, it has a continuous slope at $x_{\min}$ and thus deviates only gently from the power law as we pass below this point, making for a challenging test. Figure 4a shows a family of curves from this distribution for different values of $x_{\min}$.

In Figure 4b we show the results of the application of both the BIC and KS methods for estimating $x_{\min}$ to a large collection of data sets drawn from (3.10). The plot shows the average estimated value $\hat{x}_{\min}$ as a function of the true $x_{\min}$ for the discrete case. The KS method appears to give good estimates of $x_{\min}$ in this case and performance is similar for continuous data also (not shown), although the results tend to be slightly more conservative (i.e., to yield slightly larger estimates $\hat{x}_{\min}$). The BIC method also performs reasonably, but, as the figure shows, the method displays a tendency to underestimate $x_{\min}$, as we might expect given the arguments of the previous section. Based on these observations, we recommend the KS method for estimating $x_{\min}$ for general applications.

These tests used synthetic data sets of $n = 50\,000$ observations, but good estimates of $x_{\min}$ can be extracted from significantly smaller data sets using the KS method; results are sensitive principally to the number n_{tail} of observations in the power-law part of the distribution. For both the continuous and discrete cases we find that good results can be achieved provided we have about 1000 or more observations in this part of the distribution. This figure does depend on the particular form of the non-power-law part of the distribution. In the present test, the distribution was designed specifically to make the determination of $x_{\min}$ challenging. Had we chosen a form that makes a more pronounced departure from the power law below

[5] We note in passing that this approach can easily be generalized to the problem of estimating a lower cut-off for data following other (non-power-law) types of distributions.

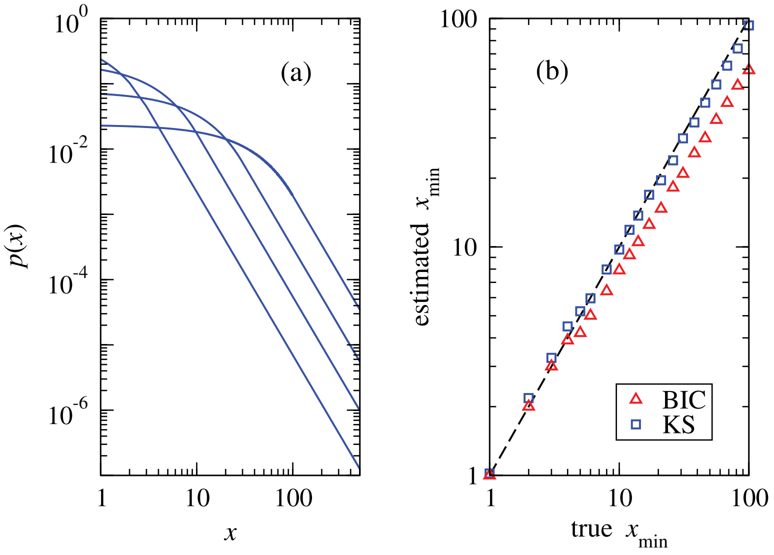

Fig. 4 (a) *Examples of the test distribution,* (3.10)*, used in the calculations described in the text, with power-law behavior for x above $x_{\min}$ but non-power-law behavior below.* (b) *The value of $x_{\min}$ estimated using the BIC and KS approaches as described in the text, plotted as a function of the true value for discrete data with $n = 50\,000$. Results are similar for continuous data.*

$x_{\min}$, then the task of estimating $\hat{x}_{\min}$, would have been easier and presumably fewer observations would have been needed to achieve results of similar quality.

For some possible distributions there is, in a sense, no true value of $x_{\min}$. The distribution $p(x) = C(x + k)^{-\alpha}$ follows a power law in the limit of large x, but there is no value of $x_{\min}$ above which it follows a power law exactly. Nonetheless, in cases such as this, we would like our method to return an $\hat{x}_{\min}$ such that when we subsequently calculate a best-fit value for α we get an accurate estimate of the true scaling parameter. In tests with such distributions we find that the KS method yields estimates of α that appear to be asymptotically consistent, meaning that $\hat{\alpha} \to \alpha$ as $n \to \infty$. Thus again the method appears to work well, although it remains an open question whether one can derive rigorous performance guarantees.

Variations on the KS method are possible that use some other goodness-of-fit measure that may perform better than the KS statistic under certain circumstances. The KS statistic is, for instance, known to be relatively insensitive to differences between distributions at the extreme limits of the range of x because in these limits the CDFs necessarily tend to zero and one. It can be reweighted to avoid this problem and be uniformly sensitive across the range [46]; the appropriate reweighting is

$$D^* = \max_{x \geq \hat{x}_{\min}} \frac{|S(x) - P(x)|}{\sqrt{P(x)(1 - P(x))}}. \tag{3.11}$$

In addition, a number of other goodness-of-fit statistics have been proposed and are in common use, such as the Kuiper and Anderson–Darling statistics [13]. We have performed tests with each of these alternative statistics and find that results for the reweighted KS and Kuiper statistics are very similar to those for the standard KS statistic. The Anderson–Darling statistic, on the other hand, we find to be highly conservative in this application, giving estimates $\hat{x}_{\min}$ that are too large by an order of magnitude or more. When there are many samples in the tail of the distribution, this degree of conservatism may be acceptable, but in most cases the reduction in the number of tail observations greatly increases the statistical error on our MLE for the scaling parameter and also reduces our ability to validate the power-law model.

Finally, as with our estimate of the scaling parameter, we would like to quantify the uncertainty in our estimate for $x_{\min}$. One way to do this is to make use of a nonparametric "bootstrap" method [16]. Given our n measurements, we generate a synthetic data set with a similar distribution to the original by drawing a new sequence of points x_i, $i = 1, \dots, n$, uniformly at random from the original data (with replacement). Using either method described above, we then estimate $x_{\min}$ and α for this surrogate data set. By taking the standard deviation of these estimates over a large number of repetitions of this process (say, 1000), we can derive principled estimates of our uncertainty in the original estimated parameters.

3.5. Other Techniques.

We would be remiss should we fail to mention some of the other techniques in use for the analysis of power-law distributions, particularly those developed within the statistics and finance communities, where the study of these distributions has, perhaps, the longest history. We give only a brief summary of this material here; readers interested in pursuing the topic further are encouraged to consult the books by Adler, Feldman, and Taqqu [4] and Resnick [48] for a more thorough explanation.[6]

In the statistical literature, researchers often consider a family of distributions of the form

$$p(x) \propto L(x)\, x^{-\alpha}\,, \tag{3.12}$$

where $L(x)$ is some slowly varying function, so that, in the limit of large x, $L(cx)/L(x) \to 1$ for any $c > 0$. An important issue in this case—as in the calculations presented in this paper—is finding the point $x_{\min}$ at which the $x^{-\alpha}$ can be considered to dominate over the nonasymptotic behavior of the function $L(x)$, a task that can be tricky if the data span only a limited dynamic range or if the non-power-law behavior $|L(x)-L(\infty)|$ decays only a little faster than $x^{-\alpha}$. In such cases, a visual approach—plotting an estimate $\hat{\alpha}$ of the scaling parameter as a function of $x_{\min}$ (called a Hill plot) and choosing for $\hat{x}_{\min}$ the value beyond which $\hat{\alpha}$ appears stable—is a common technique. Plotting other statistics, however, can often yield better results—see, for example, [33] and [57]. An alternative approach, quite common in the quantitative finance literature, is simply to limit the analysis to the largest observed samples only, such as the largest $\sqrt{n}$ or $\frac{1}{10}n$ observations [17].

The methods described in section 3.3, however, offer several advantages over these techniques. In particular, the KS method of section 3.3 gives estimates of $x_{\min}$ as least as good while being simple to implement and having low enough computational costs that it can be effectively used as a foundation for further analyses such as the calculation of p-values in section 4. And, perhaps more importantly, because the KS method removes the non-power-law portion of the data entirely from the estimation

[6] Another related area of study is "extreme value theory," which concerns itself with the distribution of the largest or smallest values generated by probability distributions, values that assume some importance in studies of, for instance, earthquakes, other natural disasters, and the risks thereof; see [14].