Power-law distributions in empirical data

Aaron Clauset,[1,2] Cosma Rohilla Shalizi,[3] and M. E. J. Newman[4]
[1]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*
[2]*Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA*
[3]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[4]*Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109, USA*

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 empirical detection and characterization of power laws is made difficult by the large fluctuations that occur in the tail of the distribution. In particular, standard methods such as least-squares fitting are known to produce systematically biased estimates of parameters for power-law distributions and should not be used in most circumstances. Here we describe statistical techniques for making accurate parameter estimates for power-law data, based on maximum likelihood methods and the Kolmogorov-Smirnov statistic. We also show how to tell whether the data follow a power-law distribution at all, defining quantitative measures that indicate when the power law is a reasonable fit to the data and when it is not. We demonstrate these methods by applying them to twenty-four real-world data sets from a range of different disciplines. Each of the data sets has been conjectured previously 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.

PACS numbers: 02.50.Tt,02.50.Ng,89.75.Da
Keywords: Power-law distributions; Pareto; Zipf; maximum likelihood; heavy-tailed distributions; likelihood ratio test; model selection

I. INTRODUCTION

Scientists have learned many things from observation of the statistical distributions of experimental quantities such as the lifetimes of excited atomic or particle states, populations of animals, plants, or bacteria, prices on the stock market, or the arrival times of messages sent across the Internet. Many, perhaps most, such quantities have distributions that are tightly clustered around their average values. That is, these distributions place a trivial amount of probability far from the mean and hence the mean is representative of most observations. For instance, it is a useful statement to say that most adult male Americans are about 180cm tall, because no one deviates very far from this average figure. Even the largest deviations, which are exceptionally rare, are still only about a factor of two from the mean in either direction and are well characterized by quoting a simple standard deviation.

Not all distributions fit this pattern, however, and while those that do not are often considered problematic or defective because they are not well characterized by their mean and standard deviation, 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 interest 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 sizes of solar flares, the populations of cities, and the intensities of earthquakes, for example, are all quantities whose distributions are thought to follow power laws. Quantities such as these are not well characterized by their averages. For instance, according to the 2000 US 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 population differs from the mean by several orders of magnitude. Extensive discussions of this and other properties of power laws can be found in the reviews by Mitzenmacher (2004) and Newman (2005), and references therein.

Power laws are the focus of this article. Specifically, we address a thorny and recurring issue in the scientific literature, the question of how to recognize a power law when we see one. A quantity x obeys a power law if it is drawn from a probability distribution

$$p(x) \propto x^{-\alpha}, \tag{1}$$

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

In practice, we rarely, if ever, know for certain that an observed quantity is drawn from a power-law distribution. Instead, the best we can typically do is to say that our observations are consistent with a model of the world in which x is drawn from a distribution of the form Eq. (1). In this paper we explain how one reaches conclusions of this kind in a reliable fashion. 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 by any stretch of the imagination 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.

II. FUNDAMENTAL PROPERTIES OF POWER LAWS

Before turning to our main topic of discussion, we first consider some fundamental mathematical issues that will be important for what follows. Further details on the mathematics of power laws can be found in Mitzenmacher (2004) and Newman (2005).

A. Continuous and discrete power-law behavior

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, normally positive integers.

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

$$p(x)\,\mathrm{d}x = \Pr(x \le X < x + \mathrm{d}x) = Cx^{-\alpha}\,\mathrm{d}x, \tag{2}$$

where X is the observed value and C is a normalization constant. Clearly this density diverges as $x \to 0$ so Eq. (2) 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

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}}\right)^{-\alpha}. \tag{3}$$

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

$$p(x) = \Pr(X = x) = Cx^{-\alpha}. \tag{4}$$

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

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}, \tag{5}$$

where

$$\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha} \tag{6}$$

is the generalized or Hurwitz zeta function.

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 $P(x) = \Pr(X \ge x)$. For instance, in the continuous case

$$P(x) = \int_x^{\infty} p(x')\,\mathrm{d}x' = \left(\frac{x}{x_{\min}}\right)^{-\alpha+1}. \tag{7}$$

In the discrete case

$$P(x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{\min})}. \tag{8}$$

As these results show, formulas for continuous power laws tend to be simpler than those for discrete power laws, with the latter often involving special functions. As a result it is common in many applications 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 a number of different ways to approximate a discrete power law by a continuous one and though some of them give reasonable results, others do not and should be avoided. 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 II.B on the generation of power-law distributed random numbers and Section III on the extraction of best-fit values for the scaling parameter from observational data.

B. Generating power-law distributed random numbers

It is often the case in statistical studies of probability distributions that we wish to generate random numbers with a given distribution. For instance, in later sections of this paper we will use uncorrelated random numbers drawn from power-law distributions to test how well our fitting procedures can estimate parameters such as α and $x_{\min}$. How should we generate such numbers? There are a variety of possible methods, but perhaps the simplest and most elegant is the *transformation method* (Press *et al.*, 1992). The method can be applied to both continuous and discrete distributions; we describe both variants in turn in this section.

Suppose $p(x)$ is a continuous probability density from which we wish to draw random reals $x \ge x_{\min}$. Typically

we will have a source of random reals r uniformly distributed in the interval $0 \le r < 1$, generated by any of a large variety of standard pseudo-random number generators. The probability densities $p(x)$ and $p(r)$ are related by

$$p(x) = p(r)\frac{\mathrm{d}r}{\mathrm{d}x} = \frac{\mathrm{d}r}{\mathrm{d}x}, \tag{9}$$

where the second equality follows because $p(r) = 1$ over the interval from 0 to 1. Integrating both sides with respect to x, we then get

$$P(x) = \int_x^\infty p(x')\,\mathrm{d}x' = \int_r^1 \mathrm{d}r' = 1 - r, \tag{10}$$

or equivalently

$$x = P^{-1}(1 - r), \tag{11}$$

where P^{-1} indicates the functional inverse of the cumulative probability distribution P. For the case of the power law, $P(x)$ is given by Eq. (7) and we find that

$$x = x_{\min}(1 - r)^{-1/(\alpha-1)}, \tag{12}$$

which can be implemented in straightforward fashion in most computer languages.

For a discrete power law the equivalent of Eq. (10) is

$$P(x) = \sum_{x'=x}^{\infty} p(x') = 1 - r. \tag{13}$$

Unfortunately, $P(x)$ is given by Eq. (8), which cannot be inverted in closed form, so we cannot write a direct expression equivalent to Eq. (12) for the discrete case. Instead, we typically solve Eq. (13) numerically by a combination of "doubling up" and binary search (Press *et al.*, 1992). That is, for a given random number r, we first bracket a solution x to the equation by the following steps:

$x_2 \leftarrow x_{\min}$
repeat
  $x_1 \leftarrow x_2$
  $x_2 \leftarrow 2x_1$
until $P(x_2) < 1 - r$

Then we pinpoint the solution within the range x_1 to x_2 by binary search. We need only continue the binary search until the value of x is narrowed down to $k \le x < k+1$ for some integer k: then we discard the integer part and the result is a power-law distributed random integer. The generalized zeta functions needed to evaluate $P(x)$ from Eq. (8) are typically calculated using special functions from standard scientific libraries. These functions can be slow, however, so for cases where speed is important, such as cases where we wish to generate very many random numbers, it may be worthwhile to store the first few thousand values of the zeta function in an array ahead of time to avoid recalculating them frequently. Only the values for smaller x are worth precalculating in this fashion, however, since those in the tail are needed only rarely.

If great accuracy is not needed it is also possible, as discussed in the previous section, to approximate the discrete power law by a continuous one. The approximation has to be done in the right way, however, if we are to get good results. Specifically, to generate integers $x \ge x_{\min}$ with an approximate power-law distribution, we first generate continuous power-law distributed reals $y \ge x_{\min} - \frac{1}{2}$ and then round off to the nearest integer $x = \lfloor y + \frac{1}{2} \rfloor$. Employing Eq. (12), this then gives

$$x = \Big\lfloor \big(x_{\min} - \tfrac{1}{2}\big)\big(1 - r\big)^{-1/(1-\alpha)} + \tfrac{1}{2} \Big\rfloor. \tag{14}$$

The approximation involved in this approach is largest for the smallest value of x, which is by definition $x_{\min}$. For this value the difference between the true power-law distribution, Eq. (5), and the approximation is given by

$$\Delta p = 1 - \left(\frac{x_{\min} + \frac{1}{2}}{x_{\min} - \frac{1}{2}}\right)^{-\alpha+1} - \frac{x_{\min}}{\zeta(\alpha, x_{\min})}. \tag{15}$$

For instance, when $\alpha = 2.5$, this difference corresponds to an error of more than 8% on the probability $p(x)$ for $x_{\min} = 1$, but the error diminishes quickly to less than 1% for $x_{\min} = 5$, and less than 0.2% for $x_{\min} = 10$. Thus the approximation is in practice a reasonably good one for quite modest values of $x_{\min}$. (Almost all of the data sets considered in Section V, for example, have $x_{\min} > 5$.) For very small values of $x_{\min}$ the true discrete generator should still be used unless large errors can be tolerated. Other approximate approaches for generating integers, such as rounding down (truncating) the value of y, give substantially poorer results and should not be used.

As an example of these techniques, consider continuous and discrete power laws having $\alpha = 2.5$ and $x_{\min} = 5$. Table I gives the cumulative density functions for these two distributions, evaluated at integer values of x, along with the corresponding cumulative density functions for three sets of 100 000 random numbers generated using the methods described here. As the table shows, the agreement between the exact and generated CDFs is good in each case, although there are small differences because of statistical fluctuations. For numbers generated using the continuous approximation to the discrete distribution the errors are somewhat larger than for the exact generators, but still small enough for many practical applications.

III. 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 data. Studies of empirical data 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

	continuous		discrete		
x	theory	generated	theory	generated	approx.
5	1.000	1.000	1.000	1.000	1.000
6	0.761	0.761	0.742	0.740	0.738
7	0.604	0.603	0.578	0.578	0.573
8	0.494	0.493	0.467	0.466	0.463
9	0.414	0.413	0.387	0.385	0.384
10	0.354	0.352	0.328	0.325	0.325
15	0.192	0.192	0.174	0.172	0.173
20	0.125	0.124	0.112	0.110	0.110
50	0.032	0.032	0.028	0.027	0.027
100	0.011	0.011	0.010	0.010	0.009

TABLE I CDFs of discrete and continuous power-law distributions with $x_{\min} = 1$ and $\alpha = 2.5$. The second and fourth columns show the theoretical values of the CDFs for the two distributions, while the third and fifth columns show the CDFs for sets of 100 000 random numbers generated from the same distributions using the transformation technique described in the text. The final column shows the CDF for 100 000 numbers generated using the continuous approximation to the discrete distribution.

task is the simple histogram. Taking logs of both sides of Eq. (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. One 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 approximately falls on a straight line, then one can, if one is 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.

Unfortunately, this method and other variations on the same theme show significant biases under relatively common conditions, as discussed in Appendix A. As a consequence, the results they return are often incorrect, sometimes substantially so, and should not be trusted. In this section we describe some alternative methods for estimating the parameters of a power-law distribution that are generally accurate. In Section IV we study the equally important question of how to determine whether a given data set really does follow a power law at all.

A. Estimating the scaling parameter

First, let us consider the estimation of the scaling parameter α. Estimating α correctly requires a value for the lower bound $x_{\min}$ of power-law behavior in the data. Let us assume for the moment 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 shortly.

The method of choice for fitting parameterized models such as power-law distributions to observed data is the *method of maximum likelihood*, which provably gives accurate (asymptotically normal) parameter estimates in the limit of large sample size (Barndorff-Nielsen and Cox, 1995; Wasserman, 2003). 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

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1} \tag{16}$$

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 (16) is equivalent to the well-known Hill estimator (Hill, 1975), which is known to be asymptotically normal (Hall, 1982) and consistent (Mason, 1982) (i.e., $\hat{\alpha} \to \alpha$ in the limits of large n, $x_{\min}$, and $n/x_{\min}$). The standard error on $\hat{\alpha}$, which is derived from the width of the likelihood maximum, is

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + \mathrm{O}(1/n), \tag{17}$$

where the higher-order correction is positive; see Appendix B of this paper, Wheatland (2004), or Newman (2005).

(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 estimators are needed to fit such a distribution.)

The MLE for the case where x is a discrete integer variable is less straightforward. Seal (1952) and more recently Goldstein *et al.* (2004) treated the situation of $x_{\min} = 1$, showing that the appropriate estimator for α is given by the solution to the transcendental equation

$$\frac{\zeta'(\hat{\alpha})}{\zeta(\hat{\alpha})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i. \tag{18}$$

When $x_{\min} > 1$, a similar equation holds, but with the zeta functions replaced by generalized zetas:

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i, \tag{19}$$

where the prime denotes differentiation with respect to the first argument. In practice, evaluation of $\hat{\alpha}$ requires us to solve this equation numerically, for instance using

	name	distribution $p(x) = Cf(x)$ $f(x)$	C
continuous	power law	$x^{-\alpha}$	$(\alpha-1)x_{\min}^{\alpha-1}$
	power law with cutoff	$x^{-\alpha}\mathrm{e}^{-\lambda x}$	$\frac{\lambda^{\alpha-1}}{\Gamma(1-\alpha,\lambda x_{\min})}$
	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	$\mu^x/x!$	$\left[\mathrm{e}^\mu - \sum_{k=0}^{x_{\min}-1}\frac{\mu^k}{k!}\right]^{-1}$

TABLE II 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	random numbers
power law	$x = x_{\min}(1-r)^{-1/(\alpha-1)}$
exponential	$x = x_{\min} - \frac{1}{\lambda}\ln(1-r)$
stretched exponential	$x = \left[x_{\min}^\beta - \frac{1}{\lambda}\ln(1-r)\right]^{1/\beta}$
log-normal	$x_1=\exp(\rho\sin\theta),\ x_2=\exp(\rho\cos\theta)$ $\rho=\sqrt{-2\sigma^2\ln(1-r_1)},\ \theta=2\pi r_2$
power law with cutoff	see caption

TABLE III Formulas for generating random numbers x drawn from continuous distributions, given a source of uniform random numbers r in the range $0 \le r < 1$. Note that for the case of the log-normal, we know of no closed-form expression for generating a *single* random number, but the expressions given will generate two independent log-normally distributed random numbers x_1, x_2, given two uniform numbers r_1, r_2 as input. For the case of the power law with cutoff, there is also no closed-form expression, but one can generate an exponentially distributed random number using the formula above and then accept or reject it with probability p or $1-p$ respectively, where $p = (x/x_{\min})^{-\alpha}$. Repeating the process until a number is accepted then gives the appropriate distribution for x.

binary search again. Alternatively, one can estimate α by direct numerical maximization of the likelihood function itself, or equivalently of its logarithm (which is usually simpler):

$$\mathcal{L}(\alpha) = -n\ln\zeta(\alpha, x_{\min}) - \alpha\sum_{i=1}^{n}\ln x_i. \tag{20}$$

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 approximation to the likelihood as our error estimate. The result is

$$\sigma = \frac{1}{\sqrt{n\left[\dfrac{\zeta''(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} - \left(\dfrac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})}\right)^2\right]}}, \tag{21}$$

which is straightforward to evaluate once we have $\hat{\alpha}$. Alternatively, Eq. (17) 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 a variant of the idea introduced in Section II.B 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

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n}\ln\frac{x_i}{x_{\min}-\frac{1}{2}}\right]^{-1}. \tag{22}$$

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 the next section, where we show that in practice this estimator gives

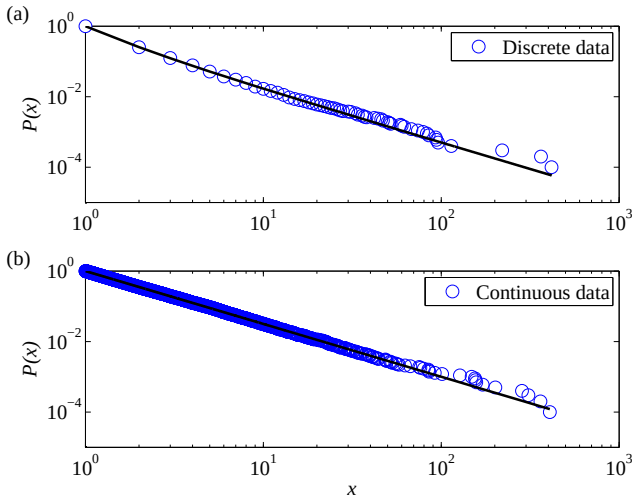

FIG. 1 (color online) Points represent the cumulative density functions $P(x)$ for synthetic data sets distributed according to (a) a discrete power law and (b) a continuous power law, both with $\alpha = 2.5$ and $x_{\min} = 1$. Solid lines represent best fits to the data using the methods described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

quite good results provided $x_{\min} \gtrsim 6$. An estimate of the statistical error on $\hat{\alpha}$ (which is quite separate from the systematic bias of the approximation) can be calculated by employing Eq. (17) 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, Eq. (16), to calculate $\hat{\alpha}$. This approach, however, gives significantly more biased values of $\hat{\alpha}$ than Eq. (22) and, given that it is no easier to implement, we see no reason to use it in any circumstances.[1]

B. Tests 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 here, 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, which are discussed in Section IV.

[1] The error involved can be shown to decay as $O(x_{\min}^{-1})$, while the error on Eq. (22) decays much faster, as $O(x_{\min}^{-2})$. In our own experiments we have found that for typical values of α we need $x_{\min} \gtrsim 100$ before Eq. (16) becomes accurate to about 1%, as compared to $x_{\min} \gtrsim 6$ for Eq. (22).

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)

TABLE IV 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 regression on the log-transformed densities. For the continuous data, the probability density function (PDF) was computed in two different ways, using bins of constant width 0.1 and using up to 500 bins of logarithmic width. The cumulative density function (CDF) was also calculated in two ways, as the cumulation of the fixed-width histogram and as a standard rank-frequency distribution. In applying the discrete MLE to the continuous data, the non-integer part of each measurement was discarded. Accurate estimates are shown in boldface.

Using the methods described in Section II.B 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 Eqs. (17) and (21).) These estimates agree well with the known true scaling parameter from which the data were generated. Figure 1 shows the actual distributions along with fits using the estimated parameters. (In this and all subsequent such plots, we show not the probability density function but the complementary cumulative density function $P(x)$. 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 IV we compare the results given by the MLEs to estimates of the scaling parameter made using several competing 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 Newman (2005)). 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 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. Fig-

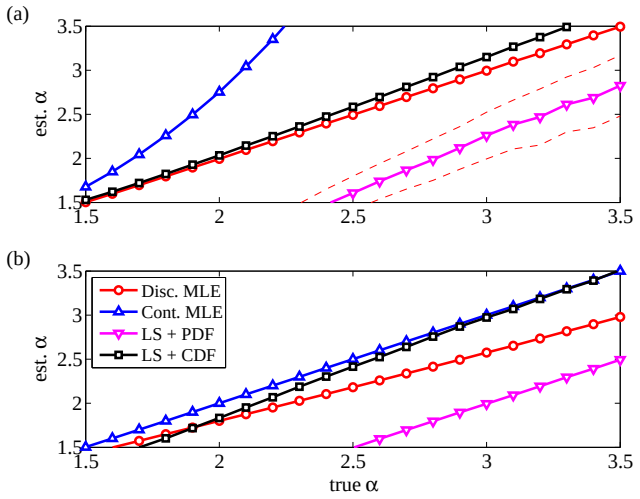

FIG. 2 (color online) Values of the scaling parameter estimated using four of the methods of Table IV (we omit the methods based on logarithmic bins for the PDF and constant width bins for the CDF) for $n = 10\,000$ observations drawn from (a) discrete and (b) continuous power law distributions with $x_{\min} = 1$. 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.

ure 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.

Figure 3 shows separately the performance of the approximate MLE for the discrete case, Eq. (22), as a function of $x_{\min}$. As shown in Appendix B, the bias in the estimator decays as $O(x_{\min}^{-2})$ and in practice falls below 1% when $x_{\min} \gtrsim 6$ for typical values of α. Many real-world data sets have $x_{\min}$ at least this large (see Table V) and hence the approximate MLE is a very practical alternative to the more cumbersome exact estimator in many cases.

Finally, the maximum likelihood estimators 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 $O(n^{-1})$ for any choice of $x_{\min}$—see Fig. 4 and Appendix B. For very small data sets, such biases can be significant but in most practical situations they can be ignored because they are much smaller than the statistical error on the estimator, which decays as $O(n^{-1/2})$. Our experience suggests that $n \gtrsim 50$ is a reasonable rule of thumb for extracting reliable parameter estimates. For the examples shown in Fig. 4 this gives estimates of α accurate to about 1% again. Data sets smaller than this should be treated with caution. Note, however, that there is another reason to treat small data sets with caution, which is typically more important, namely that it is difficult with such data to rule out

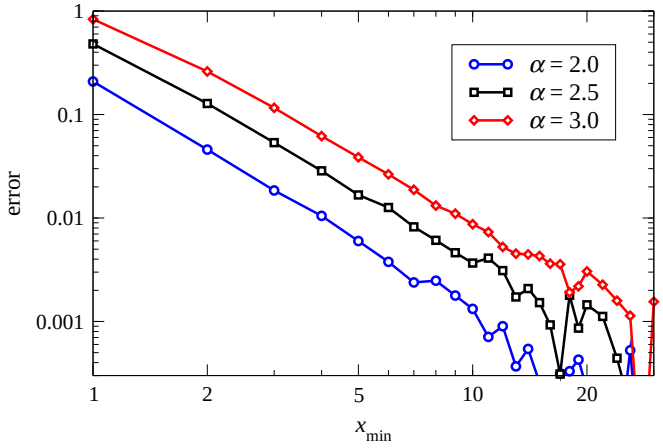

FIG. 3 (color online) The error on the estimated scaling parameter for discrete data that arises from using the approximate MLE, Eq. (22), for $\alpha = 2$, 2.5, and 3, as a function of $x_{\min}$. The average error decays as $O(x_{\min}^{-2})$ and becomes smaller than 1% of the value of α when $x_{\min} \gtrsim 6$.

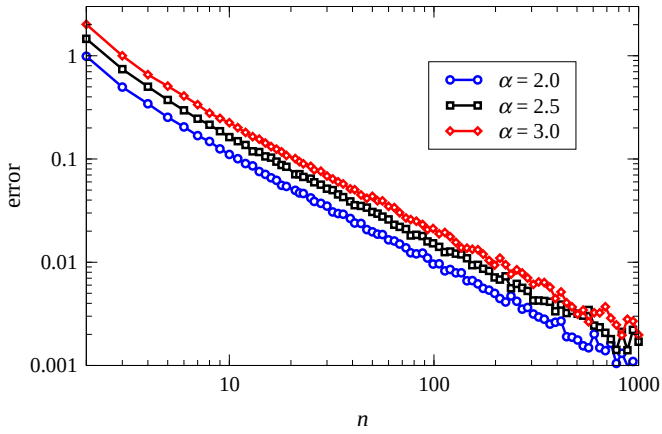

FIG. 4 (color online) The error on the estimated scaling parameter from sample size effects for continuous data (similar results hold for the discrete case), for $\alpha = 2$, 2.5, and 3, as a function of sample size. The average error decays as $O(n^{-1})$ and becomes smaller than 1% of the value of α when $n \gtrsim 50$.

alternative forms for the distribution. That is, for small data sets 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 this issue in Section IV.

C. Estimating the lower bound on power-law behavior

We now turn to the problem of estimating the lower limit $x_{\min}$ on the scaling behavior from data. This issue is important in the typical case where there is some non-power-law behavior at the lower end of the distribution of x. In such cases, we need a reliable method for estimating where power-law behavior starts: without it,

we cannot make a reliable estimate of the scaling parameter. 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. Our goal is to find a good compromise between these cases.

Traditionally, $\hat{x}_{\min}$ has been chosen either by visually identifying a point beyond which the PDF or CDF of the distribution becomes roughly straight on a log-log plot, or by plotting $\hat{\alpha}$ (or a related quantity) as a function of $\hat{x}_{\min}$ and identifying a point beyond which $\hat{\alpha}$ appears relatively stable. But these approaches are clearly subjective and can be sensitive to noise or fluctuations in the tail of the distribution—see Stoev *et al.* (2006) and references therein. A more objective and principled approach is desirable.

One approach that is appropriate for the discrete case has been described by Handcock and Jones (2004) who proposed a general model for the data both above and below $x_{\min}$ in which points above follow the normal power-law distribution and those below have a distribution parametrized by a separate probability $p_k = \Pr(X = k)$ for each possible integer value k. They then look for the best fit of this model to the observed data, allowing $x_{\min}$, as well as all the model parameters, to vary. One cannot, however, fit such a model to the data directly within the maximum likelihood framework because the number of parameters in the model is not fixed: it is equal to $x_{\min} + 1$.[2] One can always achieve higher values of the likelihood by increasing the number of parameters, thus making the model more flexible, so the maximum likelihood would always be achieved for $x_{\min} \to \infty$. The standard approach in such cases is instead to maximize the *marginal likelihood*, i.e., the likelihood of the data given the number of model parameters, but with the model parameters themselves integrated out. 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 carried out to yield an expression in terms of the value at the maximum and the determinant of the appropriate Hessian matrix. Usually we don't know the Hessian, but Schwarz (1978) has pointed out that for large n the terms involving it become negligible anyway and, dropping these terms, one derives an approximation for the log marginal likelihood of the form

$$\ln \Pr(x|x_{\min}) \simeq \mathcal{L} - \tfrac{1}{2}(x_{\min} + 1) \ln n, \tag{23}$$

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 $\hat{x}_{\min}$.

This method works well under some circumstances, but can also present difficulties. In particular, the assumption that $x_{\min}$ 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 should be generalized to the case of continuous data, for which there is no obvious choice for how many parameters are needed to represent the distribution below $x_{\min}$.

Here we present an alternative method for selecting $x_{\min}$ in both discrete and continuous cases. The fundamental idea behind the method is very simple: we choose the value $\hat{x}_{\min}$ 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 ideal value.

There are a variety of measures for quantifying the distance between two probability distributions, but for non-normal data the commonest is the Kolmogorov-Smirnov or KS statistic (Press *et al.*, 1992), which is simply the maximum distance between the CDFs of the data and the fitted model:

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)| \,. \tag{24}$$

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.

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.

[2] 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.

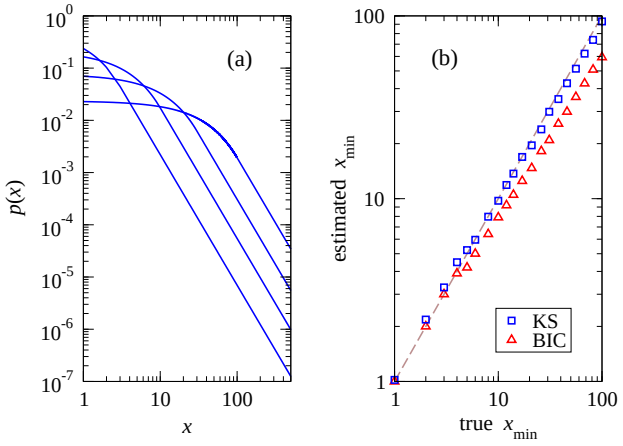

FIG. 5 (color online) (a) Examples of the test distribution, Eq. (25), used in the calculations described in the text, with power-law behavior for x above $x_{\min}$ but non-power-law behavior below. (b) Value of $x_{\min}$ estimated using the KS statistic as described in the text and using the Bayesian information criterion approach of Handcock and Jones (2004), as a function of the true value, for discrete data with $n = 50\,000$. Results are similar for continuous data.

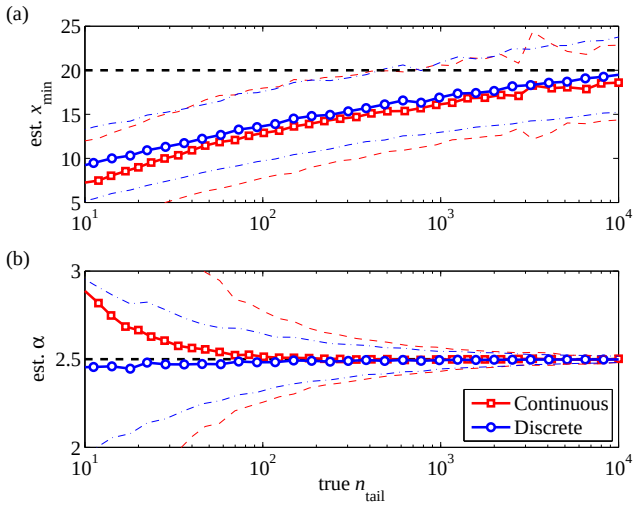

FIG. 6 (color online) Convergence of estimates for (a) $x_{\min}$ and (b) α as a function of the number n_{tail} of observations in the power-law region, for both continuous and discrete variables. (Standard deviations of the estimates are shown as the dashed lines.) The large deviations for the continuous estimator and small n_{tail} are due to finite-size sample effects. In general, we find that our parameter estimates appear to be asymptotically consistent, i.e., as $n_{\text{tail}} \to \infty$, both $\hat{x}_{\min} \to x_{\min}$ and $\hat{\alpha} \to \alpha$.

D. Tests of estimates for the lower bound

As with our MLEs for the scaling parameter, we can test our method for estimating $x_{\min}$ by generating synthetic data and examining the method's 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

$$p(x) = \begin{cases} C(x/x_{\min})^{-\alpha} & \text{for } x \geq x_{\min}, \\ C\mathrm{e}^{-\alpha(x/x_{\min}-1)} & \text{for } x < x_{\min}, \end{cases} \tag{25}$$

with, in our case, $\alpha = 2.5$. This distribution follows a power law above $x_{\min}$ but an exponential below. Furthermore, it has both a continuous value and a continuous slope at $x_{\min}$ and thus deviates only gently from the power law as we pass this point, making for a challenging test of our method. Figure 5a shows a family of curves from this distribution for different values of $x_{\min}$.

In Fig. 5b we show the results of the application of our method for estimating $x_{\min}$ to a large collection of data sets drawn from this distribution. The plot shows the average estimated value $\hat{x}_{\min}$ as a function of the true $x_{\min}$ for the discrete case using the KS statistic. Results are similar for the continuous case, although they tend to be slightly more conservative (i.e., to yield slightly larger estimates $\hat{x}_{\min}$). We also show in the same figure estimates of $x_{\min}$ made using the BIC method, which also performs acceptably, but displays a tendency to underestimate $x_{\min}$, as we might expect based on the arguments of the previous section. At least for these data, therefore, the method described in this paper appears to give better results.

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 (Fig. 6). The method is sensitive principally to the number 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 pose a challenge to the method. Had we chosen a form that makes a more pronounced departure from the power law below $x_{\min}$ then the task of estimating $\hat{x}_{\min}$ would be easier and presumably fewer observations would be needed to achieve results of similar quality.

Another slightly different class of distributions to test our method against would be those that only tend to a power law asymptotically, such as the shifted power law $p(x) = C(x + k)^{-\alpha}$ with constant k or Student's t-distribution with $\alpha - 1$ degrees of freedom. For distributions such as these there is, in a sense, no correct value of $x_{\min}$. Nonetheless, we would like our method to choose 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 (not shown) we find that our estimates of α appear to be asymptotically consistent, i.e., $\hat{\alpha} \to \alpha$ as $n \to \infty$. Thus our estimator for $x_{\min}$ seems to work well, although we

do not have rigorous guarantees of its performance in this situation.

Variations on the method described here are possible. We could use some other goodness-of-fit measure on place of the KS statistic. For instance, the KS statistic is 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 (Press *et al.*, 1992); the appropriate reweighting is

$$D^* = \max_{x \geq \hat{x}_{\min}} \frac{|S(x) - P(x)|}{\sqrt{P(x)(1-P(x))}}. \tag{26}$$

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 (D'Agostino and Stephens, 1986). 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 (Efron and Tibshirani, 1993). 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, $i = 1 \ldots n$ uniformly at random from the original data. Using the 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, we can derive principled estimates of our uncertainty in the original estimated parameters.

In summary, it is possible to make reliable estimates of $x_{\min}$, and thus also α, by minimizing the standard KS statistic. Some other statistics (Kuiper, reweighted KS) give results of comparable quality but not better. We do not recommend using the Anderson-Darling statistic in most cases.

IV. TESTING THE POWER-LAW HYPOTHESIS

The tools described in the previous sections allow us to fit a power-law distribution to a given data set and provide good estimates of the parameters α and $x_{\min}$. They tell us nothing, however, about whether the data are *well* fitted by the power law. In particular, data that are actually generated from a different distribution—an exponential, say, or a log-normal—can always be fit to a power-law model, but the fit may be very poor and in any case tells us nothing if the model itself is wrong. In practice, therefore, when considering a data set that may be derived from a power-law distribution, our challenge is to decide not only what the best parameter choices are but also whether the power-law distribution is even a reasonable hypothesis to begin with.

Many previous empirical studies of ostensibly power-law distributed data have not attempted to test the power-law hypothesis quantitatively. Instead, they typically rely on qualitative appraisals of the data, based for instance on visualizations. But these can be deceptive and can lead to claims of power-law behavior that do not hold up under closer scrutiny. Consider, for example, Fig. 7a, which shows the CDFs of three test data sets drawn from a power-law distribution with $\alpha = 2.5$, a log-normal distribution with $\mu = 0.3$ and $\sigma = 2.0$, and an exponential distribution with exponential parameter $\lambda = 0.125$. In each case the distributions have a lower cut-off of $x_{\min} = 15$. Because each of these distributions looks roughly straight on the log-log plot used in the figure, one might, upon cursory inspection, judge all three to follow power laws, albeit with different scaling parameters. This would, however, be an erroneous judgment—being roughly straight on a log-log plot is a necessary but not sufficient condition for power-law behavior.

In this section we describe quantitative methods for testing the hypothesis that a given data set is drawn from a power-law distribution for $x \geq x_{\min}$. The approach we recommend has two parts. The first, described in Section IV.A, focuses on the question of whether the data we observe could plausibly have been drawn from a power-law distribution. The second (Sections IV.B and IV.C) focuses on whether there exist other competing distributions that fit the data as well or better. Together, the techniques we describe can be used to provide objective evidence for or against the power-law hypothesis.

A. Goodness-of-fit tests

Given an observed data set and a power-law distribution from which, it is hypothesized, the data are drawn, we want to know whether that hypothesis is a likely one given the data. That is, could the data we see have plausibly been drawn from the specified power-law distribution? If the answer to this question is no, then we are wasting our time: the power law is the wrong model for our data. Questions of this type can be answered using goodness-of-fit tests that compare the observed data to the hypothesized distribution. Many such tests have been proposed, but one of the simplest, and the one we apply here, is based on the Kolmogorov-Smirnov statistic, which we encountered in Section III.C.

As we have seen, we can quantify how closely a hypothesized distribution resembles the actual distribution of an

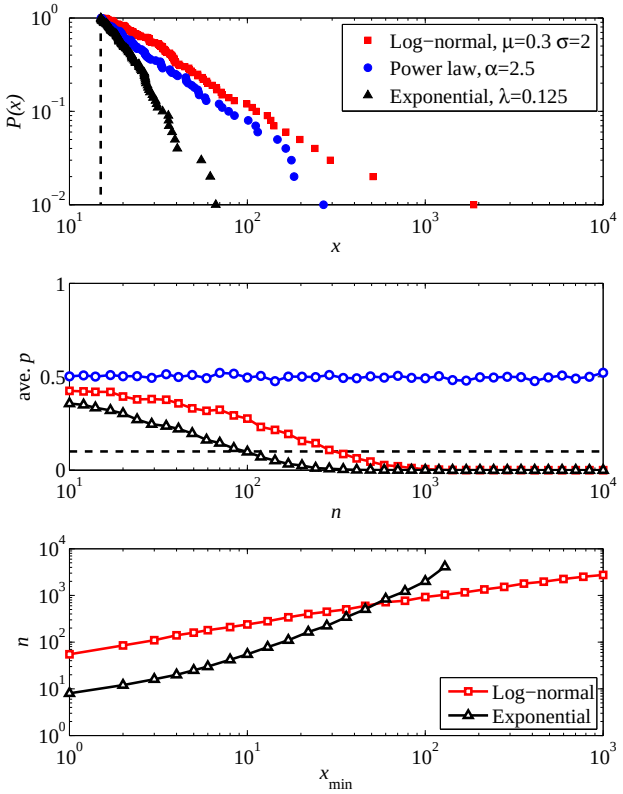

FIG. 7 (color online) (a) The CDF $P(x)$ for small samples ($n = 100$) from three continuous distributions, a log-normal with $\mu = 0.3$ and $\sigma = 2$, a power law with $\alpha = 2.5$, and an exponential with $\lambda = 0.125$, all with $x_{\min} = 15$. (Definitions of the parameters are given in Table II.) (b) The average p-value relative to the maximum likelihood power-law model for samples from the same three distributions, as a function of n. (c) The average number of observations n required to make the p-value less than 0.1 for the log-normal and exponential distributions, as a function of $x_{\min}$.

observed set of samples by calculating the Kolmogorov-Smirnov (KS) statistic, Eq. (24). The calculation returns a single number that is smaller for hypothesized distributions that are a better fit to the data. Our approach in this section is to calculate this number for the observed data set and the best-fit power-law distribution computed as described in Section III. Then, if this value is suitably small we can say that the power law is a plausible fit to the data; if the value is too large the power-law model can be ruled out. The crucial question we need to answer, of course, is how large is too large?

The standard way to answer this question is to compute a *p-value*. The p-value quantifies the probability that our data were drawn from the hypothesized distribution, based on the observed goodness of fit. It is defined to be the probability that a data set of the same size that is truly drawn from the hypothesized distribution would have goodness of fit D as bad or worse than the observed value. In essence, it tells you how likely it is that you saw results as bad as you did if the data really are power-law distributed. If the p-value is much less than 1, then it is unlikely that the data are drawn from a power law. If it is closer to 1 then the data *may* be drawn from a power law, but it cannot be guaranteed. This last point is an important one: the goodness-of-fit test and the accompanying p-value are a tool only for *ruling out* models, not for ruling them in. They can tell us when a model such as the power law is probably wrong, but they cannot tell us when it is right. The best we can do by way of confirming the power-law model, in a strictly statistical sense, is to say that it is not ruled out by the observed data.

One of the nice features of the KS statistic is that *its* distribution is known for data sets truly drawn from any given distribution. This allows one to write down an explicit expression in the limit of large n for the p-value as a function of D—see for example Press *et al.* (1992). Unfortunately, this calculation assumes that you know in advance the hypothesized distribution for the data. In our case, however, the distribution is not known. Instead it is determined by fitting to the very same data whose goodness-of-fit we wish to test. This introduces a correlation between the distribution and the data that makes the standard formula for the p-value incorrect (Goldstein *et al.*, 2004).

The problem is that, even if a data set is drawn from a perfect power-law distribution, the fit between that data set and the true distribution will on average be poorer than the fit to the best-fit distribution, because of statistical fluctuations. This means we cannot treat the best-fit distribution as if it were the true distribution; if we did we would find an apparent fit that was misleadingly good, and hence calculate too large a p-value.

In cases such as this, there is no known formula for calculating the p-value, but we can still calculate it numerically by the following Monte Carlo procedure. We generate a large number of synthetic data sets drawn from the power-law distribution that best fits the observed data, fit each one individually to the power-law model using the methods of Section III, calculate the KS statistic for each one *relative to its own best-fit model*, and then simply count what fraction of the time the resulting statistic is larger than the value D observed for the true data. This fraction is our p-value.

Note crucially that for each synthetic data set we compute the KS statistic relative to the best-fit power law for that data set, not relative to the original distribution from which the data set was drawn. In this way we ensure that we are comparing apples to apples: we are performing for each synthetic data set the same calculation that we performed for the real data set.

In the present case we need to create synthetic data that have a distribution similar to the empirical data below $x_{\min}$ but that follow the fitted power law above $x_{\min}$.

To do this, we make use of a semi-parametric approach. Suppose that our observed data set has n_{tail} observations $x \geq x_{\min}$ and n observations in total. We generate a new data set with n observations as follows. With probability n_{tail}/n we generate a random number x_i drawn from a power law with scaling parameter $\hat{\alpha}$ and $x \geq x_{\min}$. Otherwise, with probability $1-n_{\text{tail}}/n$, we select one element uniformly at random from among the elements of the observed data set that have $x < x_{\min}$ and set x_i equal to that element. Repeating the process for all $i = 1 \ldots n$ we generate a complete synthetic data set that indeed follows a perfect power law above $x_{\min}$ but has the same (non-power-law) distribution as the observed data below.

It is now quite straightforward to test the hypothesis that an observed data set is drawn from a power-law distribution. The steps are as follows:

1. Determine the best fit of the power law to the data, estimating both the scaling parameter α and the cutoff parameter $x_{\min}$ using the methods of Section III.

2. Calculate the KS statistic for the goodness-of-fit of the best-fit power law to the data.

3. Generate a large number of synthetic data sets using the procedure above, fit each according to the methods of Section III, and calculate the KS statistic for each fit.

4. Calculate the p-value as the fraction of the KS statistics for the synthetic data sets whose value exceeds the KS statistic for the real data.

5. If the p-value is sufficiently small the power-law distribution can be ruled out.

An obvious question to ask is what constitutes a "large number" of synthetic data sets. Based on an analysis of the expected worst-case performance of the method, a good rule of thumb turns out to be the following: if we wish our p-values to be accurate to within about ϵ of the true value, then we should generate at least $\frac{1}{4}\epsilon^{-2}$ synthetic data sets. Thus if, for example, we wish our p-value to be accurate to about 2 decimal digits, we would choose $\epsilon = 0.01$, which implies we should generate about 2500 synthetic sets. For the example calculations described in Section V we used numbers of this order, ranging from 1000 to 10 000 depending on the particular application.

We also need to decide what value of p should be considered sufficiently small to rule out the power-law hypothesis. In our calculations we have made the relatively conservative choice that the power law is ruled out if $p \leq 0.1$: that is, it is ruled out if there is a probability of 1 in 10 or less that we would merely by chance get data that agree this poorly with the model. (Many authors use the more lenient rule $p \leq 0.05$, but we feel this would let through some candidate distributions that have only a very small chance of really following a power law. Of course, in practice, the particular rule adopted must depend on the judgment of the investigator in the particular circumstances at hand.[3])

It is important to appreciate, as discussed above, that a large p-value does not necessarily mean the power law is the correct distribution for the data. There are (at least) two reasons for this. First, there may be other distributions that match the data equally well or better over the range of x observed. Other tests are needed to rule out such alternatives, which we discuss in Sections IV.B and IV.C.

Second, the statistical variation of the KS statistic becomes smaller as n becomes large. This means that the p-value becomes a more reliable test as n becomes large, but for small n it is possible to get quite high p-values purely by chance even when the power law is the wrong model for the data. This is not a deficiency of the method; it reflects the fact that it genuinely is harder to rule out the power law if we have less data. For this reason, the p-value should be treated with caution when n is small.

As a demonstration of the approach, consider data of the type shown in Fig. 7a, drawn from continuous power-law, log-normal, and exponential distributions. In Fig. 7b we show the average p-value for data sets drawn from these three distributions with the same parameters as in panel (a), calculated for the best-fit power-law model in each case, as a function of the number n of samples in the data sets. As the figure shows, the p-values for all three distributions are well above our threshold of 0.1 when n is small: for samples this small one cannot typically distinguish between the three distributions because we simply do not have enough data to go on. As the sizes of the samples become larger however, the p-values for the two non-power-law distributions fall off and it becomes possible to say that the power-law model is a poor fit for these data sets, while remaining a good fit for the true power-law data set.

For the log-normal and exponential distributions the p-value first falls below the threshold of 0.1 when $n \simeq 300$ and $n \simeq 100$, respectively, meaning that for these distributions samples of about these sizes or larger are needed if we wish to firmly rule out the power-law hypothesis. As shown in Fig. 7c, these values of n depend quite strongly on the choice of $x_{\min}$. They also depend more weakly on the other parameters of the distributions.

---

[3] Some readers may be familiar with the use of p-values to confirm (rather than rule out) hypotheses for experimental data. In the latter context, one quotes a p-value for a "null" model, a model *other* than the model the experiment is attempting to verify. Normally one then considers low values of p to be good, since they indicate that the null hypothesis is unlikely to be correct. Here, by contrast, we use the p-value as a measure of the hypothesis we are trying to verify, and hence high values, not low, are "good." For a general discussion of the interpretation of p-values, see Mayo and Cox (2006).

B. Goodness-of-fit tests for competing distributions

As discussed above, one of the problems that arises in attempting to validate the power-law hypothesis is that, even though our data may fit a power law quite well, there is still the possibility that another distribution, such as an exponential or a log-normal, might also give a good fit over the range of x covered by the data. We can use the approach described in the previous section to address this problem as well: we simply calculate a p-value for a fit to the competing distribution.

Suppose, for instance, that we believe our data might follow either a power-law or an exponential distribution. If we discover that the p-value for the power law is reasonably large (say larger than 0.1) then the power law is not ruled out. To strengthen our case for the power law we would like to rule out the competing exponential distribution if possible. To do this, we would find the best-fit exponential distribution, using the equivalent for exponentials of the methods of Section III, and the corresponding KS statistic, then repeat the calculation for a large number of synthetic data sets and hence calculate a p-value. If the p-value is sufficiently small, we can rule out the exponential as a model for our data.

By combining p-value calculations with respect to the power law and several plausible competing distributions, we can in this way make a good case for or against the power-law form for our data. In particular, if the p-value for the power law is high, while those for competing distributions are small, then the competition is ruled out and, although we cannot guarantee that the power law is correct, the case in its favor is strengthened.

It is worth emphasizing that we cannot of course compare the power-law fit of our data with fits to every competing the distribution, of which there are an infinite number. Furthermore, it will always be possible to find a distribution that fits the data better than the power law if we define a family of curves with a sufficiently large number of parameters. Fitting the statistical distribution of data should therefore be approached using a combination of statistical techniques like those described here and physical intuition about what constitutes a reasonable model for the data. Statistical tests can be used to rule out specific hypotheses, but it is up to the researcher to decide what a reasonable hypothesis is in the first place.

C. Direct comparison of models

Sometimes we would like to ask not whether a specific model or models are ruled out as a description of the data, but which, if any, of two models is the better fit to the data. For instance, one might like to know whether a given data set is better fit by a power law or an exponential. This question can be answered using the methods of the previous section, but there are other more direct approaches too. In this section we describe one such approach, the *likelihood ratio test*, which is typically considerably easier to implement than KS tests against competing distributions. The disadvantage of the likelihood ratio test is that it cannot tell us when both of our two distributions are poor fits to the data; it tells us only which (if either) is the least bad. In the same circumstances the methods of the previous section would rule out both distributions. On the other hand, if we already know, for example from performing a KS test against the power-law model as in Section IV.A, that the power-law model is not ruled out by the data, then there is no danger that both models are poor fits and it is safe to use the likelihood ratio test. In these circumstances this test can give us exactly the information we need without demanding a great deal of work.

The basic idea behind the likelihood ratio test is to compute the likelihood of our data in two competing distributions. The one with the higher likelihood is then the better fit. Equivalently one can calculate the ratio R of the two likelihoods, and the winning distribution is indicated by whether this likelihood ratio is greater than or less than unity. In fact, more commonly we use the logarithm, $\mathcal{R}$, of the ratio, which is positive or negative depending on which candidate distribution is the winner.

The simple sign of the log likelihood ratio, however, does not on its own tell us definitively if one distribution is better than the other; the log likelihood ratio, like other quantities, is subject to statistical fluctuation. If its true value, meaning its mean value over many independent data sets drawn from the same distribution, is close to zero, then the fluctuations can easily change the sign of the ratio and hence the results of the test cannot be trusted. In order to make a firm choice between distributions we need a log ratio that is sufficiently positive or negative that it could not plausibly be the result of a chance fluctuation from a true result that is close to zero.[4]

To make a quantitative judgment about whether the observed value of the log likelihood ratio is sufficiently far from zero, we need to know the size of the expected fluctuations, i.e., we need to know the standard deviation on $\mathcal{R}$. This we can estimate from our data using the following method, which was first proposed and analyzed by Vuong (1989).

Let us denote our two candidate distributions by $p_1(x)$ and $p_2(x)$. Then the likelihoods of our data set within

[4] An alternative method for choosing between distributions, the Bayesian approach described by Stouffer *et al.* (2005), is asymptotically equivalent to the likelihood ratio test under reasonable conditions. Bayesian estimation in this context is equivalent to a smoothing of the MLE, which buffers the results against fluctuations to some extent (Shalizi, 2007), but the method is incapable, itself, of saying whether the results could be due to chance (Mayo, 1996; Wasserman, 2006).

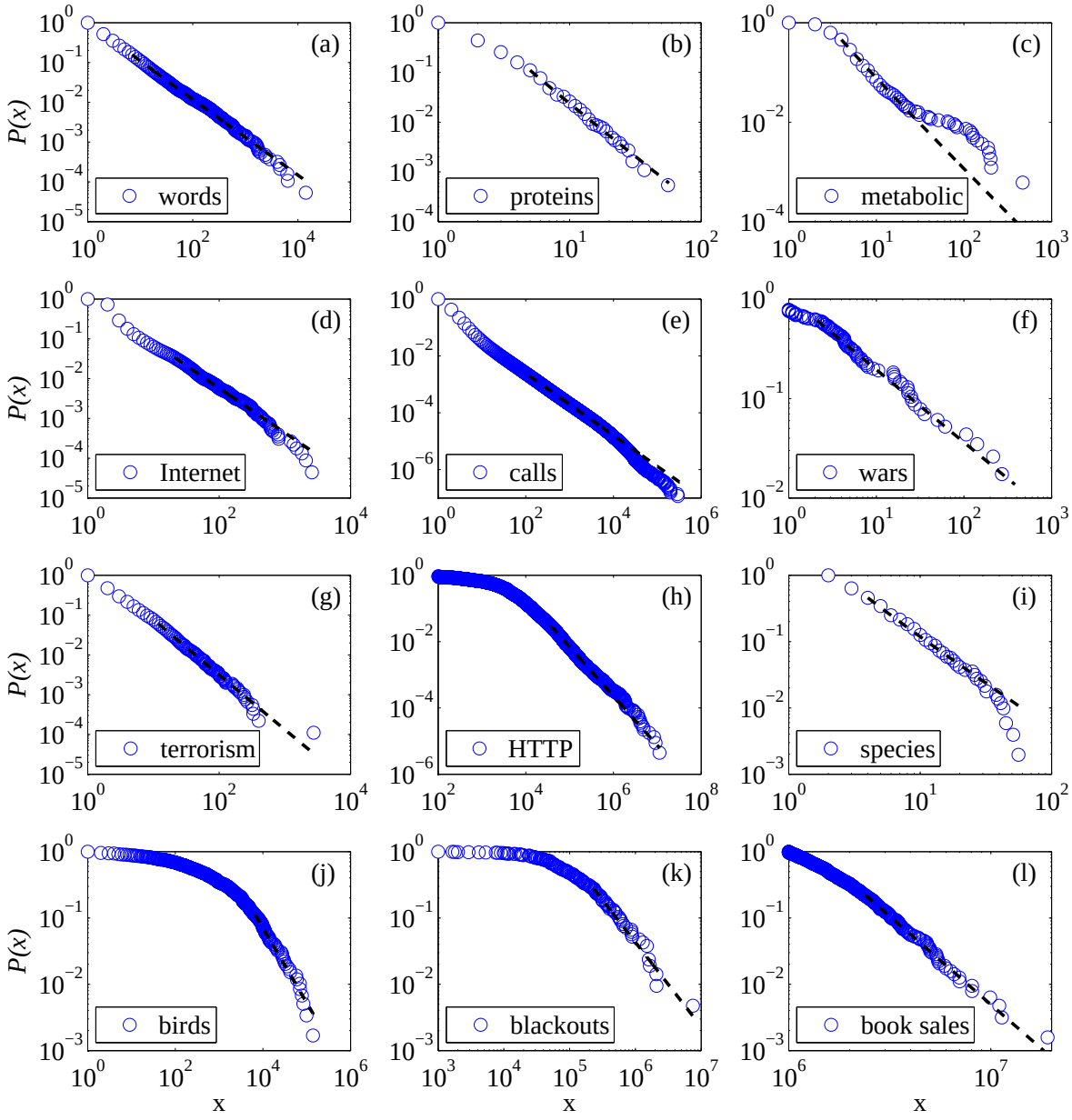

FIG. 8 (color online) The cumulative distribution functions $P(x)$ and their maximum likelihood power-law fits, for the first twelve of our twenty-four empirical data sets. (a) The frequency of occurrence of unique words in the novel *Moby Dick* by Herman Melville. (b) The degree distribution of proteins in the protein interaction network of the yeast *S. cerevisiae*. (c) The degree distribution of metabolites in the metabolic network of the bacterium *E. coli*. (d) The degree distribution of autonomous systems (groups of computers under single administrative control) on the Internet. (e) The number of calls received by US customers of the long-distance telephone carrier AT&T. (f) The intensity of wars from 1816–1980 measured as the number of battle deaths per 10 000 of the combined populations of the warring nations. (g) The severity of terrorist attacks worldwide from February 1968 to June 2006 measured as the number of deaths greater than zero. (h) The size in bytes of HTTP connections at a large research laboratory. (i) The number of species per genus in the class *Mammalia* during the late Quaternary period. (j) The frequency of sightings of bird species in the United States. (k) The number of customers affected by electrical blackouts in the United States. (l) Sales volume of bestselling books in the United States.

the two distributions are given by

$$L_1 = \prod_{i=1}^{n} p_1(x_i), \qquad L_2 = \prod_{i=1}^{n} p_2(x_i), \tag{27}$$

and the ratio of the likelihoods is

$$R = \frac{L_1}{L_2} = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_2(x_i)}. \tag{28}$$

Taking logs, the log likelihood ratio is

$$\mathcal{R} = \sum_{i=1}^{n} \bigl[\ln p_1(x_i) - \ln p_2(x_i)\bigr] = \sum_{i=1}^{n} \bigl[\ell_i^{(1)} - \ell_i^{(2)}\bigr], \tag{29}$$

where $\ell_i^{(j)} = \ln p_j(x_i)$ can be thought of as the log-likelihood for a single measurement x_i within distribution j.

But since, by hypothesis, the x_i are independent, so also are the differences $\ell_i^{(1)} - \ell_i^{(2)}$, and hence, by the central limit theorem, their sum $\mathcal{R}$ becomes normally distributed as n becomes large with expected variance $n\sigma^2$, where σ^2 is the expected variance of a single term. In practice we don't know the expected variance of a single term, but we can approximate it in the usual way by the variance of the data:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \Bigl[\bigl(\ell_i^{(1)} - \ell_i^{(2)}\bigr) - \bigl(\bar\ell^{(1)} - \bar\ell^{(2)}\bigr)\Bigr]^2, \tag{30}$$

with

$$\bar\ell^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \ell_i^{(1)}, \qquad \bar\ell^{(2)} = \frac{1}{n} \sum_{i=1}^{n} \ell_i^{(2)}. \tag{31}$$

Now suppose that the true log likelihood ratio is in fact zero or close to it, so that the observed sign of $\mathcal{R}$ is a product purely of the fluctuations and cannot be trusted as an indicator of which model is preferred. Then the probability that the measured log likelihood ratio has a magnitude as large or larger than the observed value $|\mathcal{R}|$ is given by

$$\begin{aligned} p &= \frac{1}{\sqrt{2\pi n\sigma^2}} \biggl[\int_{-\infty}^{-|\mathcal{R}|} \mathrm{e}^{-t^2/2n\sigma^2}\,\mathrm{d}t + \int_{|\mathcal{R}|}^{\infty} \mathrm{e}^{-t^2/2n\sigma^2}\,\mathrm{d}t\biggr] \\ &= \bigl|\mathrm{erfc}(\mathcal{R}/\sqrt{2n}\sigma)\bigr|, \end{aligned} \tag{32}$$

where σ is given by Eq. (30) and

$$\mathrm{erfc}(z) = 1 - \mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^{\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t \tag{33}$$

is the complementary Gaussian error function (a function widely available in scientific computing libraries and numerical analysis programs).

This probability is another example of a p-value of the type discussed in Section IV.B. It gives us an estimate of the probability that we measured a given value of $\mathcal{R}$ when the true value of $\mathcal{R}$ is close to zero (and hence is unreliable as a guide to which model is favored). If p is small (say $p < 0.1$) then our value for $\mathcal{R}$ is unlikely to be a chance result and hence its sign can probably be trusted as an indicator of which model is the better fit to the data. (It does not however mean that the model is a *good* fit, only that it is better than the alternative.) If on the other hand p is large, then the likelihood ratio test is inadequate to discriminate between the distributions in question.[5]

Vuong (1989) in fact recommends quoting the normalized log likelihood ratio $n^{-1/2}\mathcal{R}/\sqrt{\sigma}$ that appears in Eq. (32). This quantity contains, in theory, everything one needs to know about the results of the likelihood ratio test. Its sign tells us which model is favored and its value, via Eq. (32), allows us to compute p and hence test the significance of the result. In practice, however, we find it convenient to quote both the normalized ratio and an explicit value for p; although technically the latter can be computed from the former, it is helpful in making judgments about particular cases to have the actual p-value at hand. In our tests on real data in Section V we give both.

D. Nested hypotheses

In some cases the distributions we wish to compare may be *nested*, meaning that one family of distributions is a subset of the other. The power law and the power law with exponential cutoff in Table II provide an example of such nested distributions. When distributions are nested it is always the case that the larger family of distributions will provide a fit as good or better than the smaller, since every member of the smaller family is also a member of the larger. Thus, the test just described can never select the smaller of the two families, even if the data truly were drawn from the smaller family and the data are abundant. A different test is thus required for nested hypotheses.

When the true distribution lies in the smaller family of distributions, the best fits to both distributions converge to the true distribution as n becomes large. This means that the individual differences $\ell_i^{(1)} - \ell_i^{(2)}$ in Eq. (29) each converge to zero, as does their variance σ^2. Consequently the ratio $|\mathcal{R}|/\sigma$ appearing in the expression for the p-value tends to 0/0, and its distribution does not obey the simple central limit theorem argument given above. A more refined analysis, using a kind of probabilistic version of L'Hopital's rule, shows that in fact $\mathcal{R}$ adopts a chi-squared distribution as n becomes large

[5] Note that, if we are interested in confirming or denying the power-law hypothesis, then a small p-value is "good" in the likelihood ratio test—it tells us whether the test's results are trustworthy—whereas it is "bad" in the case of the KS test, where it tells us that our model is a poor fit to the data.

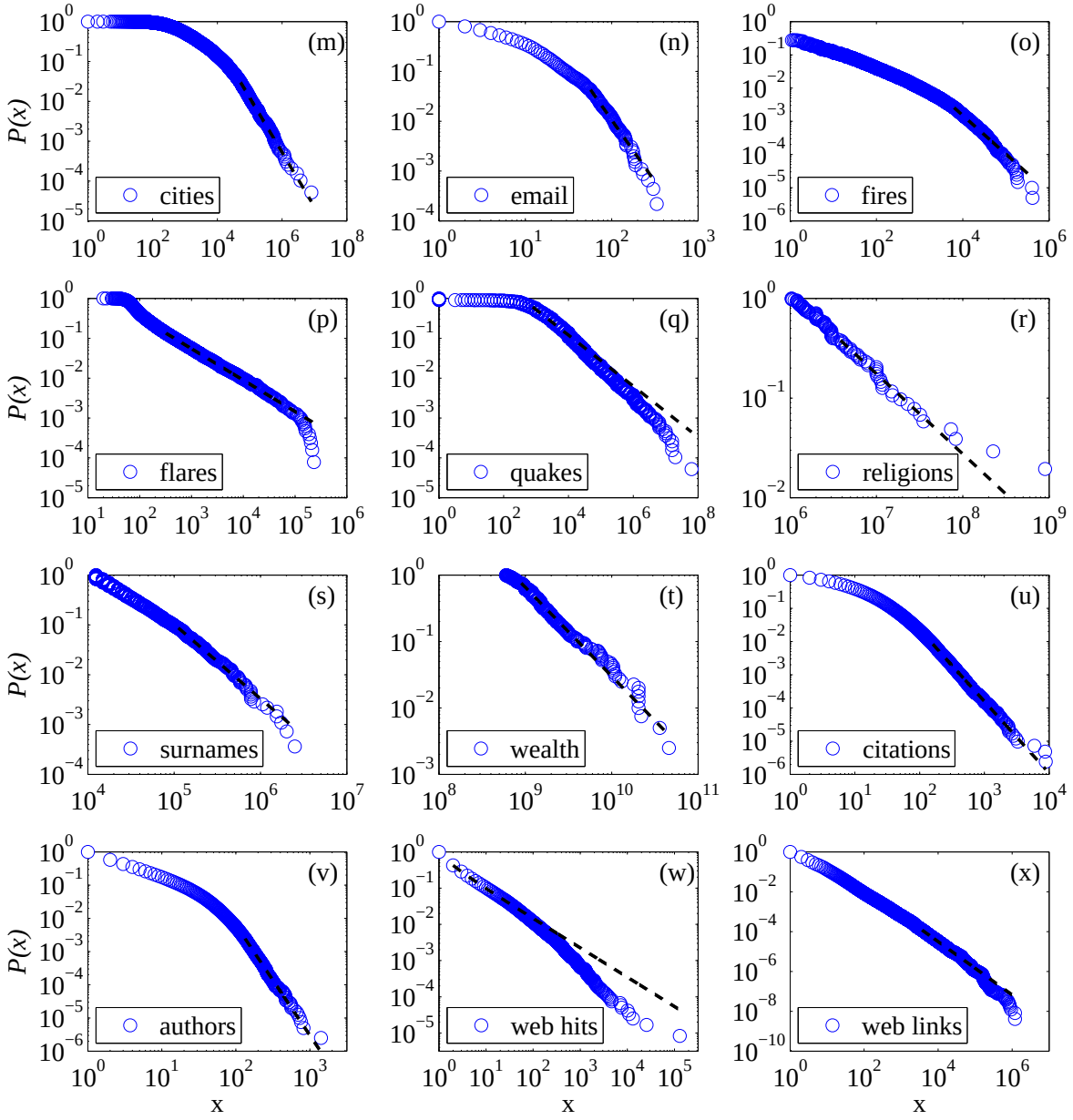

FIG. 9 (color online) The cumulative distribution functions $P(x)$ and their maximum likelihood power-law fits, for the second twelve of our twenty-four empirical data sets. (m) Population of cities in the United States. (n) Size of email address books at a university. (o) Number of acres burned in California forest fires. (p) Intensity of solar flares. (q) Intensity of earthquakes. (r) Number of adherents of religious sects. (s) Frequency of surnames in the United States. (t) Net worth (in USD) of the richest American citizens. (u) The number of citations received by published academic papers. (v) The number of authors on published academic papers. (w) The number of hits on web sites from AOL users. (x) The number of hyperlinks to web sites.

(Wilks, 1938). One can use this result to calculate a correct p-value giving the probability that the log likelihood ratio takes the observed value or worse, if the true distribution falls in the smaller family. If this p-value is small, then the smaller family can be ruled out. If not, then the best we can say is that the there is no evidence that the larger family is needed to fit to the data, although neither can it be ruled out. For a more detailed discussion of this special case see, for instance, Vuong (1989).

V. TESTS OF REAL-WORLD DATA

We are now in a position to apply the methods we have described to real-world data. In this section we examine a large number of data sets representing measurements of quantities whose distributions, it has been conjectured, follow power laws. For each data set, we fit and test the power-law hypothesis using the methods described in the previous sections. As we will see, the results indicate that some of the data sets are indeed consistent with power-law distributions, some are not, and some are marginal cases for which the power law is a possible candidate distribution, but is not strongly supported by the data.

The data sets we study are all previously published and come from a broad variety of different branches of human endeavor, including physics, astrophysics, earth sciences, biology and biochemistry, ecology, paleontology, computer and information sciences, engineering, forestry, geography, economics and finance, and the social sciences. They are as follows:

a) The frequency of occurrence of unique words in the novel *Moby Dick* by Herman Melville (Newman, 2005).

b) The degrees (i.e., numbers of distinct interaction partners) of proteins in the partially known protein-interaction network of the yeast *Saccharomyces cerevisiae* (Ito *et al.*, 2000).

c) The degrees of metabolites in the metabolic network of the bacterium *Escherichia coli* (Huss and Holme, 2006).

d) The degrees of nodes in the partially known network representation of the Internet at the level of autonomous systems for May 2006 (Holme *et al.*, 2007). (An autonomous system is a group of IP addresses on the Internet among which routing is handled internally or "autonomously," rather than using the Internet's large-scale BGP routing mechanism.)

e) The number of calls received by customers of AT&T's long distance telephone service in the US during a single day (Abello *et al.*, 1998; Aiello *et al.*, 2000).

f) The intensity of wars from 1816–1980 measured as the number of battle deaths per 10 000 of the combined populations of the warring nations (Roberts and Turcotte, 1998; Small and Singer, 1982).

g) The severity of terrorist attacks worldwide from February 1968 to June 2006, measured as the number of deaths directly resulting (Clauset *et al.*, 2007).

h) The number of bytes of data received as the result of individual web (HTTP) requests from computer users at a large research laboratory during a 24-hour period in June 1996 (Willinger and Paxson, 1998). Roughly speaking this distribution represents the size distribution of web files transmitted over the Internet.

i) The number of species per genus of mammals. This data set, compiled by Smith *et al.* (2003), is composed primarily of species alive today but also includes a subset of recently extinct species, where "recent" in this context means the last few tens of thousands of years.

j) The numbers of sightings of bird species in the North American Breeding Bird Survey for 2003.

k) The number customers affected in electrical blackouts in the United States between 1984 and 2002 (Newman, 2005).

l) The numbers of copies sold in the United States of bestselling books for the period 1895 to 1965 (Hackett, 1967).

m) Human populations of US cities in the 2000 US Census.

n) Sizes of email address books of computer users at a large university (Newman *et al.*, 2002).

o) The size in acres of wildfires occurring on US federal land between 1986 and 1996 (Newman, 2005).

p) Peak gamma-ray intensity of solar flares between 1980 and 1989 (Newman, 2005).

q) Intensity of earthquakes occurring in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake (Newman, 2005).

r) Number of adherents of religious denominations, bodies, or sects, as compiled and published on the web site adherents.com.

s) Frequency of occurrence of US family names in the 1990 US Census.

t) Aggregate net worth in US dollars of the richest individuals in the US in October 2003 (Newman, 2005).

u) The number of citations received between publication and June 1997 by scientific papers published in 1981 and listed in the Science Citation Index (Redner, 1998).

quantity	n	$\langle x \rangle$	σ	$x_{\max}$	$\hat{x}_{\min}$	$\hat{\alpha}$	n_{tail}
count of word use	18 855	11.14	148.33	14 086	7 ± 2	1.95(2)	2958 ± 987
protein interaction degree	1846	2.34	3.05	56	5 ± 2	3.1(3)	204 ± 263
metabolic degree	1641	5.68	17.81	468	4 ± 1	2.8(1)	748 ± 136
Internet degree	22 688	5.63	37.83	2583	21 ± 9	2.12(9)	770 ± 1124
telephone calls received	51 360 423	3.88	179.09	375 746	120 ± 49	2.09(1)	$102\,592 \pm 210\,147$
intensity of wars	115	15.70	49.97	382	2.1 ± 3.5	1.7(2)	70 ± 14
terrorist attack severity	9101	4.35	31.58	2749	12 ± 4	2.4(2)	547 ± 1663
HTTP size (kilobytes)	226 386	7.36	57.94	10 971	36.25 ± 22.74	2.48(5)	6794 ± 2232
species per genus	509	5.59	6.94	56	4 ± 2	2.4(2)	233 ± 138
bird species sightings	591	3384.36	10 952.34	138 705	6679 ± 2463	2.1(2)	66 ± 41
blackouts ($\times 10^3$)	211	253.87	610.31	7500	230 ± 90	2.3(3)	59 ± 35
sales of books ($\times 10^3$)	633	1986.67	1396.60	19 077	2400 ± 430	3.7(3)	139 ± 115
population of cities ($\times 10^3$)	19 447	9.00	77.83	8 009	52.46 ± 11.88	2.37(8)	580 ± 177
email address books size	4581	12.45	21.49	333	57 ± 21	3.5(6)	196 ± 449
forest fire size (acres)	203 785	0.90	20.99	4121	6324 ± 3487	2.2(3)	521 ± 6801
solar flare intensity	12 773	689.41	6520.59	231 300	323 ± 89	1.79(2)	1711 ± 384
quake intensity ($\times 10^3$)	19 302	24.54	563.83	63 096	0.794 ± 80.198	1.64(4)	$11\,697 \pm 2159$
religious followers ($\times 10^6$)	103	27.36	136.64	1050	3.85 ± 1.60	1.8(1)	39 ± 26
freq. of surnames ($\times 10^3$)	2753	50.59	113.99	2502	111.92 ± 40.67	2.5(2)	239 ± 215
net worth (mil. USD)	400	2388.69	4 167.35	46 000	900 ± 364	2.3(1)	302 ± 77
citations to papers	415 229	16.17	44.02	8904	160 ± 35	3.16(6)	3455 ± 1859
papers authored	401 445	7.21	16.52	1416	133 ± 13	4.3(1)	988 ± 377
hits to web sites	119 724	9.83	392.52	129 641	2 ± 13	1.81(8)	$50\,981 \pm 16\,898$
links to web sites	241 428 853	9.15	106 871.65	1 199 466	3684 ± 151	2.336(9)	$28\,986 \pm 1560$

TABLE V Basic parameters of the data sets described in this section, along with their power-law fits.

v) The number of academic papers authored or coauthored by mathematicians listed in the American Mathematical Society's MathSciNet database. Data compiled by J. Grossman.

w) The number of "hits" received by web sites from customers of the America Online Internet service in a single day (Adamic and Huberman, 2000).

x) The number of links to web sites found in a 1997 web crawl of about 200 million web pages (Broder *et al.*, 2000).

Many of these data sets are only limited samples of much larger collections (such as the web sites, which are only a small fraction of the entire web). In some cases it is known that the sampling procedure used can be biased, e.g., for the protein interactions (Sprinzak *et al.*, 2003), citations and authorships (Bhattacharya and Getoor, 2007), and the Internet (Achlioptas *et al.*, 2005; Dimitropoulos *et al.*, 2007). The cited references provide details of these issues, but our analysis does not attempt to estimate or correct for the biases.

In Table V we show results for the best fits of the power-law form to each of these data sets, using the methods described in Section III, along with a variety of generic statistics such as mean, standard deviation, and maximum value.

As an indication of the importance of correct analysis methods for data such as these, we note that many of the values we find for the scaling parameters differ considerably from those deduced by previous authors from the same data, using *ad hoc* methods. For instance, the scaling parameter for the protein interaction network of Ito *et al.* (2000) has been reported to take a value of 2.44 (Yook *et al.*, 2004), which is quite different from, and incompatible with, the value we find of 3.1 ± 0.3. Similarly, the citation distribution data of Redner (1998) have been reported to have a scaling parameter of either 2.9 (Tsallis and de Albuquerque, 1999) or 2.5 (Krapivsky *et al.*, 2000), neither of which are compatible with our maximum likelihood figure of 3.16 ± 0.06.

In Tables VI and VII we show the results of our tests of the power-law hypothesis, which quantify whether the hypothesis is tenable for the data in question. Based on the results of these tests, we summarize in the final column of the table how convincing the power-law model is as a fit to each data set. In some cases, the p-values alone give us a fairly unambiguous measure of how believable the power law is, while in others, the likelihood ratio tests are also needed.

For most of the data sets considered the power-law model is in fact a plausible one, meaning that the p-value for the best fit is large. Other distributions may be a better fit, but the power law is not ruled out, especially if it is backed by additional physical insights that indicate it to be the correct distribution. In just one case—the distribution of the frequencies of occurrence of words in English text—the power law appears to be truly convincing in the sense that it is an excellent fit to the data and none of the alternatives carries any weight.

For seven of the data sets, on the other hand, the

data set	power law p	log-normal LR	log-normal p	exponential LR	exponential p	stretched exp. LR	stretched exp. p	power law + cut-off LR	power law + cut-off p	support for power law
birds	**0.55**	−0.850	0.40	1.87	**0.06**	−0.882	0.38	−1.24	0.12	moderate
blackouts	**0.62**	−0.412	0.68	1.21	0.23	−0.417	0.68	−0.382	0.38	moderate
book sales	**0.66**	−0.267	0.79	2.70	**0.01**	3.885	**0.00**	−0.140	0.60	moderate
cities	**0.76**	−0.090	0.93	3.65	**0.00**	0.204	0.84	−0.123	0.62	moderate
fires	0.05	−1.78	**0.08**	4.00	**0.00**	−1.82	**0.07**	−5.02	**0.00**	with cut-off
flares	**1.00**	−0.803	0.42	13.7	**0.00**	−0.546	0.59	−4.52	**0.00**	with cut-off
HTTP	0.00	1.77	**0.08**	11.8	**0.00**	2.65	**0.01**	0.000	1.00	none
quakes	0.00	−7.14	**0.00**	11.6	**0.00**	−7.09	**0.00**	−24.4	**0.00**	with cut-off
religions	**0.42**	−0.073	0.94	1.59	0.11	1.75	**0.08**	−0.167	0.56	moderate
surnames	**0.20**	−0.836	0.40	2.89	**0.00**	−0.844	0.40	−1.36	**0.10**	with cut-off
wars	**0.20**	−0.737	0.46	3.68	**0.00**	−0.767	0.44	−0.847	0.19	moderate
wealth	0.00	0.249	0.80	6.20	**0.00**	8.05	**0.00**	−0.142	0.59	none
web hits	0.00	−10.21	**0.00**	8.55	**0.00**	10.94	**0.00**	−74.66	**0.00**	with cut-off
web links	0.00	−2.24	**0.03**	25.3	**0.00**	−1.08	0.28	−21.2	**0.00**	with cut-off

TABLE VI Tests of power-law behavior in the data sets studied here that are comprised of continuous (non-discrete) data. (Results for the discrete data sets are given in Table VII.) For each data set we give a p-value for the fit to the power-law model and likelihood ratios for the alternatives. We also quote p-values for the significance of each of the likelihood ratio tests. Statistically significant p-values are denoted in **bold**. Positive values of the log likelihood ratios indicate that the power-law model is favored over the alternative. For non-nested alternatives, we give the normalized log likelihood ratio $n^{-1/2}\mathcal{R}/\sigma$ which appears in Eq. (32), while for the power law with exponential cut-off we give the actual log likelihood ratio. The final column of the table lists our judgment of the statistical support for the power-law hypothesis for each data set. "None" indicates data sets that are probably not power-law distributed; "moderate" indicates that the power law is a good fit but that there are other plausible alternatives as well; "good" indicates that the power law is a good fit and that none of the alternatives considered is plausible. (None of the data sets in this table earned a rating of "good," but one data set in Table VII, for the frequencies of words, is so designated.) In some cases we write "with cut-off," meaning that the power law with exponential cutoff is clearly favored over the pure power law. However, in each of the latter cases some of the alternative distributions are also good fits, such as the log-normal or the stretched exponential distribution.

data set	p	Poisson LR	Poisson p	log-normal LR	log-normal p	exponential LR	exponential p	stretched exp. LR	stretched exp. p	power law + cut-off LR	power law + cut-off p	support for power law
Internet	**0.29**	5.31	**0.00**	−0.807	0.42	6.49	**0.00**	0.493	0.62	−1.97	**0.05**	with cut-off
calls	**0.63**	17.9	**0.00**	−2.03	**0.04**	35.0	**0.00**	14.3	**0.00**	−30.2	**0.00**	with cut-off
citations	**0.20**	6.54	**0.00**	−0.141	0.89	5.91	**0.00**	1.72	**0.09**	−0.007	0.91	moderate
email	**0.16**	4.65	**0.00**	−1.10	0.27	0.639	0.52	−1.13	0.26	−1.89	**0.05**	with cut-off
metabolic	0.00	3.53	**0.00**	−1.05	0.29	5.59	**0.00**	3.66	**0.00**	0.000	1.00	none
papers	**0.90**	5.71	**0.00**	−0.091	0.93	3.08	**0.00**	0.709	0.48	−0.016	0.86	moderate
proteins	**0.31**	3.05	**0.00**	−0.456	0.65	2.21	**0.03**	0.055	0.96	−0.414	0.36	moderate
species	**0.10**	5.04	**0.00**	−1.63	0.10	2.39	**0.02**	−1.59	0.11	−3.80	**0.01**	with cut-off
terrorism	**0.68**	1.81	**0.07**	−0.278	0.78	2.457	**0.01**	0.772	0.44	−0.077	0.70	moderate
words	**0.49**	4.43	**0.00**	0.395	0.69	9.09	**0.00**	4.13	**0.00**	−0.899	0.18	good

TABLE VII Tests of power-law behavior in the data sets studied here that are comprised of discrete (integer) data. Statistically significant p-values are denoted in **bold**. Results for the continuous data sets are given in Table VI; see that table for a description of the individual column entries.

p-value is sufficiently small that the power-law model can be firmly ruled out. In particular, the distributions for the HTTP connections, earthquakes, web links, fires, wealth, web hits, and the metabolic network cannot plausibly be considered to follow a power law; the probability of getting a fit as poor as that observed purely by chance is very small in each case and one would have to be unreasonably optimistic to see power-law behavior in any of these data sets. (For two data sets—the HTTP connections and wealth distribution—the power law, while not a good fit, is nonetheless better than the alternatives, implying that these data sets are not well-characterized by any of the functional forms considered here.)

Of the remaining data sets, each is compatible with the power-law hypothesis according to the KS test, but we can gain more insight by looking at the likelihood ratio tests, which tell us whether other distributions may be a good fit as well.

We find that for all the data sets save three, we can rule out the exponential distribution as a possible fit—the likelihood ratio test firmly favors the power law over the exponential. The three exceptions are the blackouts, religions, and email address books, for which the power law is favored over the exponential but the accompanying

p-value is large enough that the results cannot be trusted. For the discrete data sets we can also rule out the Poisson distribution in every case.

For the log-normal and stretched exponential distributions the likelihood ratio tests are in most cases inconclusive: the p-values indicate that the observed likelihood ratios have a high probability of being purely the result of chance. This mean that for most of the data sets the power law may be a plausible model but the log-normal and stretched exponential are also plausible. In cases such as these, it will be important to look at physical motivating factors to make a sensible judgment about the true distribution—we must consider whether there is a mechanistic reason to believe one distribution or another to be correct.

In other cases the likelihood ratio tests do give conclusive answers. In some the tests rule firmly in favor of the power law. For instance, the stretched exponential is ruled out for the book sales, telephone calls, and citation counts. In other cases the tests rule in favor of the alternative—the stretched exponential is strongly favored over the power law for the forest fires and earthquakes, for example. Note however that the log-normal is not ruled out for any of our data sets, save the HTTP connections. In every case it is a plausible alternative and in a few it is strongly favored. In fact, we find that it is in general extremely difficult to tell the difference between a log-normal and true power-law behavior. Indeed over realistic ranges of x the two distributions are very closely equal, so it appears unlikely that any test would be able to tell them apart unless we have an extremely large data set. Thus one must again rely on physical intuition to draw any final conclusions. Otherwise, the best that one can say is that the data do not rule out the power law, but that other distributions are as good or better a fit to the data.

Finally a word concerning the cut-off power law is in order. Since this model is a superset of the power-law model, it can, as discussed in Section IV.D, never truly be ruled out, as reflected in the fact that the likelihood ratio is always either zero or negative; the power law with cut-off can never be a worse fit than the pure power law. In many cases, however, the corresponding p-value shows that the likelihood ratio is not significant and there is no statistical reason to prefer the cut-off form. For almost a dozen data sets, however—the forest fires, solar flares, earthquakes, web hits, web links, telephone calls, Internet, email address books, and mammal species—the cut-off form is clearly favored. For surnames, the cut-off form is favored over the pure form but only weakly, as the p-value is very close to our threshold. In each of these cases some of the other models, such as log-normal or stretched exponential, are also plausible, so again some physical insight must be added into the mix to reach a conclusion about the true underlying distribution.

VI. OTHER TECHNIQUES

Before giving our concluding thoughts, we would be remiss should we fail to mention some of the other techniques for the analysis of power-law distributions, particularly those developed within the statistics and finance communities. We give only a very brief summary of this material here; readers interested in pursuing the topic further are encouraged to consult the books by Adler *et al.* (1998) and Resnick (2006) for a more thorough explanation.[6]

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

$$p(x) \propto L(x)\, x^{-\alpha}, \tag{34}$$

where $L(x)$ is some slowly varying function, i.e., in the limit of large x, $L(c\,x)/L(x) \to 1$ for any $c > 0$. An important issue in this case—as it is in the calculations presented in this paper—is deciding the point $x_{\min}$ at which the $x^{-\alpha}$ dominates over the non-asymptotic behavior of the function $L(x)$, a task that can be tricky if the data span only a limited dynamic range or if $|L(x) - L(\infty)|$ decays only a little faster than $x^{-\alpha}$. A common approach involves plotting an estimate $\hat{\alpha}$ of the scaling parameter as a function of $x_{\min}$ and choosing for $\hat{x}_{\min}$ the value beyond which $\hat{\alpha}$ appears stable. If we use the popular Hill estimator for α (Hill, 1975), which is equivalent to our maximum likelihood estimator for continuous data, Eq. (16), such a plot is called a Hill plot. Other estimators, however, can often yield more useful results—see, for example, Kratz and Resnick (1996) and Stoev *et al.* (2006). An alternative approach, quite common in the economics 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 (Farmer *et al.*, 2004).

The methods we describe in Section III offer several advantages over these visual or heuristic techniques. In particular, the goodness-of-fit-based approach gives accurate estimates of $x_{\min}$ with synthetic data (Fig. 5b) and appears to give reasonable estimates in real-world situations too (Fig. 8). Moreover, its simple implementation and low computational cost readily lends it to further analyses such as the calculation of p-values in Section IV.A.[7] And because our method removes the non-power-law portion of the data entirely from the estimation of the scaling parameter, we end up fitting simpler functional forms to the data, which allows us more easily

[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 de Hann and Ferreira (2006).

[7] We further note that the goodness-of-fit-based approach for estimating $x_{\min}$ can easily be adapted to estimating a lower cut-off for other distributions.

to test the statistical agreement between the data and the best-fit model.

VII. CONCLUSIONS

The study of power-law distributed quantities spans an impressive variety of disciplines, including physics, computer and information sciences, the earth sciences, molecular and cellular biology, evolutionary biology, ecology, economics, political science, sociology, and statistics. Unfortunately, well founded methods for analyzing power-law data have not yet taken root in all, or even most, of these areas and in many cases hypothesized distributions are not tested rigorously against the data, which leaves open the possibility that apparent power-law behavior is, in some cases at least, merely an figment of the researcher's imagination.

The common practice of identifying and quantifying power-law distributions by the approximately straight-line behavior of a histogram on a doubly logarithmic plot is known to give biased results and should not be trusted. In this paper we have described a collection of simple but reliable alternative techniques that can be used to search for power-law behavior in real-world data. These techniques include methods for fitting a power law or related form to data, methods for gauging the range over which power-law behavior holds, methods for assessing whether the power law is a good fit to the data, methods for comparing the quality of fit to that of competing distributions, and a number of related tools such as methods for generating random numbers from power-law distributions for use in Monte Carlo or bootstrap calculations.

Applying these methods to data sets from a broad range of different fields, we find an interesting picture. In many of the cases studied in this paper the power-law hypothesis turns out to be, statistically speaking, a reasonable description of the data. That is, the data are compatible with the hypothesis that they are drawn from a power-law distribution, at least over a part of their range, although in many cases they are compatible with other distributions as well, such as log-normal or stretched exponential distributions. In the remaining cases the power-law hypothesis is found to be incompatible with the observed data, although in a number of instances the power law becomes plausible again if one allows for the possibility of an exponential cut-off that truncates the tail of the distribution. Thus it appears that the conclusions drawn from *ad hoc* methods of analysis are sometimes correct and sometimes not, a result that highlights the inadequacy of these methods. The methods described here, by contrast, give us solid evidence to back up our claims when we do find power-law behavior and in cases where two or more competing hypotheses appear plausible they allow us to make quantitative statements about relative merit.

In some cases, the interesting scientific conclusions do not rest upon a quantity having a perfect power-law distribution. It may be enough merely that the quantity have a heavy-tailed distribution. For instance, in studies of the Internet the distributions of many quantities, such as file sizes, HTTP connections, node degrees, and so forth, have heavy tails and appear visually to follow a power law, but upon more careful analysis it proves impossible to make a strong case for the power-law hypothesis; typically the power-law distribution is not ruled out but competing distributions may offer a better fit. Whether this constitutes a problem for the researcher depends largely on his or her scientific goals. For a computer engineer, simply quantifying the heavy tail may help address important questions concerning, for instance, future infrastructure needs or the risk of overload from large but rare events. Thus in some cases pure power-law behavior may not be fundamentally more interesting than any other heavy-tailed distribution. (In such cases, non-parametric estimates of the distribution may be useful, though making such estimates for heavy-tailed data presents special difficulties (Markovitch and Krieger, 2000).) If, on the other hand, our goal is to infer plausible mechanisms that might underlie the formation and evolution of Internet structure or traffic patterns, then it may matter greatly whether the observed quantity follows a true power law or some other form.

In closing, we echo comments made by Ijiri and Simon (1977) more than thirty years ago and similar thoughts expressed more recently by Mitzenmacher (2006). They argue that the characterization of empirical distributions is only a part of the challenge that faces us in explaining the ubiquity of power laws in the sciences. In addition we also need methods to validate the models that have been proposed to explain those power laws. They also urge that, wherever possible, we consider to what practical purposes these robust and interesting behaviors can be put. We hope that the methods given here will prove useful in all of these endeavors, and that these long-delayed hopes will at last be fulfilled.

Acknowledgments

The authors thank Sandra Chapman, Allen Downey, Doyne Farmer, Chris Genovese, Joel Greenhouse, Luwen Huang, Kristina Klinkner, Joshua Ladau, Michael Mitzenmacher, Cristopher Moore, Sidney Resnick, Stilian Stoev, Valérie Ventura, Larry Wasserman, Nick Watkins, Michael Wheatland, and Maxwell Young for helpful conversations and comments and Lada Adamic, Alison Boyer, Andrei Broder, Allen Downey, Petter Holme, Mikael Huss, Josh Karlin, Sidney Redner, Janet Wiener, and Walter Willinger for generously sharing data. 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).

Computer code implementing many of the analysis

methods described in this paper can be found online at `http://www.santafe.edu/~aaronc/powerlaws/`.

APPENDIX A: Problems with linear regression and power laws

The most common approach for testing empirical data against a hypothesized power-law distribution is to observe that the power law $p(x) \sim x^{-\alpha}$ implies the linear form

$$\log p(x) = \alpha \log x + c. \tag{A1}$$

The probability density $p(x)$ can be estimated by constructing a histogram of the data or one can construct the cumulative distribution function by a simple rank ordering of the data and the resulting distribution fitted to the linear form by least-squares linear regression. The slope of the fit is interpreted as the estimate $\hat{\alpha}$ of the scaling parameter. Many standard packages exist that can perform this kind of line-fitting and also provide standard errors for the estimated slope and calculate the fraction r^2 of variance accounted for by the fitted line, which is taken as an indicator of the quality of the fit.

Although this procedure appears frequently in the literature there are several problems with it. As we saw in Section III, the estimates of the slope are subject to systematic and potentially large errors (see Table IV and Fig. 2) and there are a number of other serious problems as well. First, the errors are hard to estimate because they are not well-described by the usual regression formulas, which are based on assumptions that do not apply in this case. Second, a fit to a power-law distribution can account for a large fraction of the variance even when the fitted data do not follow a power law, and hence high values of r^2 cannot be taken as evidence in favor of the power-law form. And third, the fits extracted by regression methods usually do not satisfy basic requirements on probability distributions, such as normalization, and hence cannot be correct. Let us look at each of these objections in a little more detail.

1. Calculation of standard errors

The ordinary formula for the calculation of the standard error on the slope of a regression line is correct when the assumptions of linear regression hold, which include independent, Gaussian noise in the dependent variable at each value of the independent variable. When fitting a histogram of the PDF, the noise, though independent, is Gaussian (actually Poisson) in the frequency estimates $p(x)$ themselves, so the noise in the *logarithms* of those frequency estimates cannot also be Gaussian. (For $\ln p(x)$ to have Gaussian fluctuations, $p(x)$ would have to have log-normal fluctuations, which would violate the central limit theorem.) Thus the formula for the error is inapplicable in this case.

For fits to the CDF the noise in the individual values $P(x)$ is also Gaussian (since it is the sum of independent Gaussian variables), but the noise in the logarithm of $P(x)$ again is not. Furthermore, the assumption of independence now fails, because $P(x) = P(x+1) + p(x)$ and hence adjacent values of the CDF are strongly correlated. Fits to the CDF are, as we showed in Section III, empirically more accurate as a method for determining the scaling parameter α, but this is not because the assumptions of the fit are any more valid. The improvement arises because the statistical fluctuations in the CDF are typically much smaller than those in the PDF. The error on the scaling parameter is thus smaller but this does not mean that the estimate of the error is any better. (In fact, it is typically a gross underestimate because of the failure to account for correlations in the noise.)

2. Validation

If our data are truly drawn from a power-law distribution and n is large, then the probability of getting a low r^2 in a straight-line fit is small, so a low value of r^2 can be used to reject the power-law hypothesis. Unfortunately, as we saw in Section IV.A, distributions that are nothing like a power-law can appear as such for small samples and some, like the log-normal, can approximate a power law closely over many orders of magnitude, resulting in high values of r^2. And even when the fitted distribution approximates a power law quite poorly, it can still account for a significant fraction of the variance, although less than the true power law. Thus, though a low r^2 is informative, we in practice rarely see a low r^2, regardless of the actual form of the distribution, so that the value of r^2 tells us little. In the terminology of statistical theory, the value of r^2 has very little *power* as a hypothesis test because the probability of successfully detecting a violation of the power-law assumption is low.

3. Regression lines are not valid distributions

The CDF must take the value 1 at $x_{\min}$ if the probability distribution above $x_{\min}$ is properly normalized. Ordinary linear regression, however, does not incorporate such constraints and hence, in general, the regression line does not respect them. Similar considerations apply for the PDF, which must integrate to 1 over the range from $x_{\min}$ to ∞. Standard methods exist to incorporate constraints like these into the regression analysis (Weisberg, 1985), but they are not used to any significant extent in the literature on power laws.

APPENDIX B: Maximum likelihood estimators for the power law

In this section we give derivations of the maximum likelihood estimators for the scaling parameter of a power

law.

1. Continuous data

In the continuous case the maximum likelihood estimator for the scaling parameter, first derived (to our knowledge) by Muniruzzaman (1957), is equivalent to the well-known Hill estimator (Hill, 1975). Consider the continuous power-law distribution,

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}, \tag{B1}$$

where α is the scaling parameter and $x_{\min}$ is the minimum value at which power-law behavior holds. Given a data set containing n observations $x_i \geq x_{\min}$, we would like to know the value of α for the power-law model that is most likely to have generated our data. The probability that the data were drawn from the model is proportional to

$$p(x \,|\, \alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha}. \tag{B2}$$

This probability is called the *likelihood* of the data given the model. The data are mostly likely to have been generated by the model with scaling parameter α that maximizes this function. Commonly we actually work with the logarithm of the likelihood, which has its maximum in the same place:

$$\begin{aligned} \mathcal{L} = \ln p(x \,|\, \alpha) &= \ln \prod_{i=1}^{n} \frac{\alpha - 1}{x_{\min}} \left(\frac{x_i}{x_{\min}} \right)^{-\alpha} \\ &= \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln x_{\min} - \alpha \ln \frac{x_i}{x_{\min}} \right] \\ &= n \ln(\alpha - 1) - n \ln x_{\min} - \alpha \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}. \end{aligned} \tag{B3}$$

Setting $\partial \mathcal{L} / \partial \alpha = 0$ and solving for α, we obtain the *maximum likelihood estimate* $\hat{\alpha}$ for the scaling parameter:

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right]^{-1}. \tag{B4}$$

There are a number of theorems in mathematical statistics that motivate and support the use of the MLE. We describe them briefly below and show that they apply. Note that these theorems are independent of any assumptions regarding the prior distribution of values for the scaling parameter, the equating of observed values with expectations, and so forth.

Theorem 1 *Under mild regularity conditions, if the data are independent, identically-distributed draws from a distribution with parameter α, then as the sample size $n \to \infty$, $\hat{\alpha} \to \alpha$ almost surely.*

Proof: See, for instance, Pitman (1979). (Note that his proof is stronger than the usual consistency result for the MLE, which gives only convergence in probability, $\Pr(|\hat{\alpha} - \alpha| > \epsilon) \to 0$ for all $\epsilon > 0$.)

Proposition 1 (Muniruzzaman (1957)) *The maximum likelihood estimator $\hat{\alpha}$ of the continuous power law converges almost surely on the true α.*

Proof: It is easily verified that $\ln(x/x_{\min})$ has an exponential distribution with rate $\alpha - 1$. By the strong law of large numbers, therefore, $\frac{1}{n} \sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}}$ converges almost surely on the expectation value of $\ln(x/x_{\min})$, which is $(\alpha - 1)^{-1}$.

Theorem 2 *If the MLE is consistent, and there exists an interval $(\alpha - \epsilon, \alpha + \epsilon)$ around the true parameter value α where, for any α_1, α_2 in that interval,*

$$\frac{\partial^3 \mathcal{L}(\alpha_1) / \partial \alpha^3}{\partial^2 \mathcal{L}(\alpha_2) / \partial \alpha^2} \tag{B5}$$

is bounded for all x, then asymptotically $\hat{\alpha}$ has a Gaussian distribution centered on α, whose variance is $1/nI(\alpha)$, where

$$I(\alpha) = -\mathbf{E} \left[\frac{\partial^2 \log p(X|\alpha)}{\partial \alpha^2} \right] \tag{B6}$$

which is called the Fisher information *at α. Moreover, $\partial^2 \mathcal{L}(\hat{\alpha}) / \partial \alpha^2 \to I(\alpha)$.*

Proof: For the quoted version of this result, see Barndorff-Nielsen and Cox (1995, ch. 3). The first version of a proof of the asymptotic Gaussian distribution of the MLE, and its relation to the Fisher information, may be found in Fisher (1922).

Proposition 2 (Muniruzzaman (1957)) *The MLE of the continuous power law is asymptotically Gaussian, with variance $(\alpha - 1)^2 / n$.*

Proof: By applying the preceding theorem. Simple calculation shows that $\partial^2 \log \mathcal{L}(\alpha) / \partial \alpha^2 = -n(\alpha - 1)^{-2}$ and $\partial^3 \log \mathcal{L}(\alpha) / \partial \alpha^3 = 2n(\alpha - 1)^{-3}$, so that the ratio in question is $2(\alpha_2 - 1)^2 / (\alpha_1 - 1)^3$. Since $\alpha > 1$, this ratio is bounded on any sufficiently small interval around any α, and the hypotheses of the theorem are satisfied.

A further standard result, the *Cramér-Rao inequality*, asserts that for *any* unbiased estimator of α, the variance is at least $1/nI(\alpha)$. (See Cramér (1945, §32.3), or, for an elementary proof, Pitman (1979).) The MLE is thus said to be *asymptotically efficient*, since it attains this lower bound.

Proposition 2 yields approximate standard error and Gaussian confidence intervals for $\hat{\alpha}$, becoming exact as n gets large. Corrections depend on how $x_{\min}$ is estimated, and the resulting coupling between that estimate and $\hat{\alpha}$. As they are $\mathrm{O}(1/n)$, however, while the leading

terms are $O(1/\sqrt{n})$, we have neglected them in the main text. The corrections can be deduced from the "sampling distribution" of $\hat{\alpha}$, i.e., the distribution of its deviations from α due to finite-sample fluctuations. (See Cramér (1945) or Wasserman (2003) for introductions to sampling distributions.) In general, these are hard to obtain analytically, but may be found by bootstrapping (Efron and Tibshirani, 1993; Wasserman, 2003). An important exception is when $x_{\min}$ is either known *a priori* or an effective $x_{\min}$ is simply chosen by fiat (as in the Hill estimator). Starting from the distribution of $\ln x$, it is then easy to show that $(\hat{\alpha}-1)/n$ has an inverse gamma distribution with shape parameter n and scale parameter $\alpha-1$. This implies (Johnson *et al.*, 1994) that $\hat{\alpha}$ has a mean of

$$\alpha\frac{n}{n-1}-\frac{1}{n-1}\ ,$$

and a standard deviation of

$$(\alpha-1)\frac{n}{(n-1)\sqrt{n-2}}\ ,$$

differing, as promised, from the large-n values by $O(1/n)$.

2. Discrete data

We define the power-law distribution over an integer variable by

$$p(x)=\frac{x^{-\alpha}}{\zeta(\alpha,x_{\min})}, \tag{B7}$$

where $\zeta(\alpha,x_{\min})$ is the generalized Riemann zeta function. For the case $x_{\min}=1$, Seal (1952) derived the maximum likelihood estimator. One can also derive an estimator for the more general case as follows.

Following an argument similar to the one we gave for the continuous power law, we can write down the log-likelihood function

$$\begin{aligned}\mathcal{L}&=\ln\prod_{i=1}^{n}\frac{x_i^{-\alpha}}{\zeta(\alpha,x_{\min})}\\&=-n\ln\zeta(\alpha,x_{\min})-\alpha\sum_{i=1}^{n}\ln x_i.\end{aligned} \tag{B8}$$

Setting $\partial\mathcal{L}/\partial\alpha=0$ we then find

$$\frac{-n}{\zeta(\alpha,x_{\min})}\frac{\partial}{\partial\alpha}\zeta(\alpha,x_{\min})-\sum_{i=1}^{n}\ln x_i=0. \tag{B9}$$

Thus, the MLE $\hat{\alpha}$ for the scaling parameter is the solution of

$$\frac{\zeta'(\hat{\alpha},x_{\min})}{\zeta(\hat{\alpha},x_{\min})}=-\frac{1}{n}\sum_{i=1}^{n}\ln x_i. \tag{B10}$$

This equation can be solved numerically in a straightforward manner. Alternatively, one can directly maximize the log-likelihood function itself, Eq. (B8).

The consistency and asymptotic efficiency of the MLE for the discrete power law can be proved by applying Theorems 1 and 2. As the calculations involved are long and messy, however, we omit them here. Brave readers can consult Arnold (1983) for the details.

Equation (B10) is somewhat cumbersome. If $x_{\min}$ is moderately large then a reasonable figure for α can be estimated using the much more convenient approximate formula derived in the next section.

3. Approximate estimator for the scaling parameter of the discrete power law

Given a differentiable function $f(x)$, with indefinite integral $F(x)$, such that $F'(x)=f(x)$,

$$\begin{aligned}\int_{x-\frac{1}{2}}^{x+\frac{1}{2}}f(t)\,\mathrm{d}t&=F\bigl(x+\tfrac{1}{2}\bigr)-F\bigl(x-\tfrac{1}{2}\bigr)\\&=\bigl[F(x)+\tfrac{1}{2}F'(x)+\tfrac{1}{8}F''(x)+\tfrac{1}{48}F'''(x)\bigr]\\&\quad-\bigl[F(x)-\tfrac{1}{2}F'(x)+\tfrac{1}{8}F''(x)-\tfrac{1}{48}F'''(x)\bigr]+\ldots\\&=f(x)+\tfrac{1}{24}f''(x)+\ldots\end{aligned} \tag{B11}$$

Summing over integer x, we then get

$$\int_{x_{\min}-\frac{1}{2}}^{\infty}f(t)\,\mathrm{d}t=\sum_{x=x_{\min}}^{\infty}f(x)+\frac{1}{24}\sum_{x=x_{\min}}^{\infty}f''(x)+\ldots \tag{B12}$$

For instance, if $f(x)=x^{-\alpha}$ for some constant α, then we have

$$\begin{aligned}\int_{x_{\min}-\frac{1}{2}}^{\infty}t^{-\alpha}\,\mathrm{d}t&=\frac{\bigl(x_{\min}-\frac{1}{2}\bigr)^{-\alpha+1}}{\alpha-1}\\&=\sum_{x=x_{\min}}^{\infty}x^{-\alpha}+\frac{1}{24\alpha(\alpha+1)}\sum_{x=x_{\min}}^{\infty}x^{-\alpha-2}+\ldots\\&=\zeta(\alpha,x_{\min})\bigl[1+O\bigl(x_{\min}^{-2}\bigr)\bigr],\end{aligned} \tag{B13}$$

where we have made use of the fact that $x^{-2}\le x_{\min}^{-2}$ for all terms in the second sum. Thus

$$\zeta(\alpha,x_{\min})=\frac{\bigl(x_{\min}-\frac{1}{2}\bigr)^{-\alpha+1}}{\alpha-1}\bigl[1+O\bigl(x_{\min}^{-2}\bigr)\bigr]. \tag{B14}$$

Similarly, putting $f(x)=x^{-\alpha}\ln x$ we get

$$\begin{aligned}\zeta'(\alpha,x_{\min})=-\frac{\bigl(x_{\min}-\frac{1}{2}\bigr)^{-\alpha+1}}{\alpha-1}\biggl[\frac{1}{\alpha-1}+\ln\bigl(x_{\min}-\tfrac{1}{2}\bigr)\biggr]\\\times\bigl[1+O\bigl(x_{\min}^{-2}\bigr)\bigr]+\zeta(\alpha,x_{\min})O\bigl(x_{\min}^{-2}\bigr).\end{aligned} \tag{B15}$$

We can use these expressions to derive an approximation to the maximum likelihood estimator for the scaling

parameter α of the discrete power law, Eq. (B10). The ratio of zeta functions in Eq. (B10) becomes

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = \left[\frac{1}{\hat{\alpha}-1} - \ln\bigl(x_{\min} - \tfrac{1}{2}\bigr)\right]\bigl[1 + \mathrm{O}\bigl(x_{\min}^{-2}\bigr)\bigr] + \mathrm{O}\bigl(x_{\min}^{-2}\bigr), \tag{B16}$$

and, neglecting quantities of order $x_{\min}^{-2}$ by comparison with quantities of order 1, we have

$$\hat{\alpha} \simeq 1 + n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min} - \frac{1}{2}}\right]^{-1}, \tag{B17}$$

which is in fact identical to the MLE for the continuous case except for the $-\frac{1}{2}$ in the denominator.

Numerical comparisons of Eq. (B17) to the exact discrete MLE, Eq. (B10), show that Eq. (B17) is a good approximation when $x_{\min} \gtrsim 6$—see Fig. 3.

References

Abello, J., A. Buchsbaum, and J. Westbrook, 1998, in *Proceedings of the 6th European Symposium on Algorithms* (Springer, Berlin).

Achlioptas, D., A. Clauset, D. Kempe, and C. Moore, 2005, in *Proceedings of the 37th ACM Symposium on Theory of Computing* (Association of Computing Machinery, New York).

Adamic, L. A., and B. A. Huberman, 2000, Quarterly Journal of Electronic Commerce **1**, 512.

Adler, R. J., R. E. Feldman, and M. S. Taqqu (eds.), 1998, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications* (Birkhäuser, Boston).

Aiello, W., F. Chung, and L. Lu, 2000, in *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York), pp. 171–180.

Arnold, B. C., 1983, *Pareto Distributions* (International Cooperative Publishing House, Fairland, Maryland).

Barndorff-Nielsen, O. E., and D. R. Cox, 1995, *Inference and Asymptotics* (Chapman and Hall, London).

Bhattacharya, I., and L. Getoor, 2007, ACM Transactions on Knowledge Discovery from Data **1**(1), 5.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, 2000, Computer Networks **33**, 309.

Clauset, A., M. Young, and K. S. Gleditsch, 2007, Journal of Conflict Resolution **51**, 58.

Cramér, H., 1945, *Mathematical Methods of Statistics* (Almqvist and Wiksells, Uppsala).

D'Agostino, R. B., and M. A. Stephens (eds.), 1986, *Goodness-of-fit Techniques* (Marcel Dekker, New York).

Dimitropoulos, X., D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, k. claffy, and G. Riley, 2007, ACM SIGCOMM Computer Communication Review **37**(1), 29.

Efron, B., and R. J. Tibshirani, 1993, *An Introduction to the Bootstrap* (Chapman and Hall, New York).

Farmer, J. D., L. Gillemot, F. Lillo, S. Mike, and A. Sen, 2004, Quantitative Finance **4**, 383.

Fisher, R. A., 1922, Philosophical Transactions of the Royal Society A **222**, 309.

Goldstein, M. L., S. A. Morris, and G. G. Yen, 2004, European Physics Journal B **41**, 255.

Hackett, A. P., 1967, *70 Years of Best Sellers, 1895–1965* (R. R. Bowker Company, New York).

Hall, P., 1982, Journal of the Royal Statistical Society B **44**, 37.

Handcock, M. S., and J. H. Jones, 2004, Theoretical Population Biology **65**, 413.

de Hann, L., and A. Ferreira, 2006, *Extreme Value Theory: An Introduction* (Springer-Verlag, New York).

Hill, B. M., 1975, Annals of Statistics **3**, 1163.

Holme, P., J. Karlin, and S. Forrest, 2007, Proceedings of the Royal Society A **463**, 1231.

Huss, M., and P. Holme, 2006, *Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks*, Preprint q-bio/0603038.

Ijiri, Y., and H. A. Simon, 1977, *Skew Distributions and the Sizes of Business Firms* (North-Holland, Amsterdam), with Charles P. Bonini and Theodore A. van Wormer.

Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, 2000, Proceedings of the National Academy of Sciences (USA) **97**, 1143.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994, *Continuous Univariate Distributions* (John Wiley, New York).

Krapivsky, P. L., S. Redner, and F. Leyvraz, 2000, Physical Review Letters **85**, 4629.

Kratz, M., and S. I. Resnick, 1996, Stochastic Models **12**, 699.

Markovitch, N. M., and U. R. Krieger, 2000, Performance Evaluation **42**, 205.

Mason, D., 1982, Annals of Probability **10**, 754.

Mayo, D. G., 1996, *Error and the Growth of Experimental Knowledge* (University of Chicago Press, Chicago).

Mayo, D. G., and D. R. Cox, 2006, in *Optimality: The Second Erich L. Lehmann Symposium*, edited by J. Rojo (Institute of Mathematical Statistics, Bethesda, Maryland), pp. 77–97.

Mitzenmacher, M., 2004, Internet Mathematics **1**, 226.

Mitzenmacher, M., 2006, Internet Mathematics **2**, 525.

Muniruzzaman, A. N. M., 1957, Bulletin of the Calcutta Statistical Association **7**, 115.

Newman, M. E. J., 2005, Contemporary Physics **46**, 323.

Newman, M. E. J., S. Forrest, and J. Balthrop, 2002, Physical Review E **66**, 035101.

Pitman, E. J. G., 1979, *Some Basic Theory for Statistical Inference* (Chapman and Hall, London).

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, England), 2nd edition.

Redner, S., 1998, European Physical Journal B **4**, 131.

Resnick, S. I., 2006, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling* (Springer-Verlag, New York).

Roberts, D. C., and D. L. Turcotte, 1998, Fractals **6**, 351.

Schwarz, G., 1978, Annals of Statistics **6**, 461.

Seal, H. L., 1952, Journal of the Institute of Actuaries **78**, 115.

Shalizi, C. R., 2007, Bayesian learning, evolutionary dynamics, and information theory, In preparation.

Small, M., and J. D. Singer, 1982, *Resort to Arms: International and Civil Wars, 1816-1980* (Sage Publications, Beverley Hills).

Smith, F. A., S. K. Lyons, S. K. M. Ernest, K. E. Jones, D. M. Kaufman, T. Dayan, P. A. Marquet, J. H. Brown, and J. P.

Haskell, 2003, Ecology **84**, 3403.
Sprinzak, E., S. Sattath, and H. Margalit, 2003, Journal of Molecular Biology **327**, 919.
Stoev, S. A., G. Michailidis, and M. S. Taqqu, 2006, *Estimating heavy-tail exponents through max self-similarity*, Preprint math/0609163.
Stouffer, D. B., R. D. Malmgren, and L. A. N. Amaral, 2005, *Comment on Barabási, Nature* **435***, 207 (2005)*, Preprint physics/0510216.
Tsallis, C., and M. P. de Albuquerque, 1999, European Physical Journal B **13**, 777.
Vuong, Q. H., 1989, Econometrica **57**, 307.
Wasserman, L., 2003, *All of Statistics: A Concise Course in Statistical Inference* (Springer-Verlag, Berlin).
Wasserman, L., 2006, Bayesian Analysis **1**, 451.
Weisberg, S., 1985, *Applied Linear Regression* (Wiley, New York), 2nd edition.
Wheatland, M. S., 2004, Astrophysical Journal **609**, 1134.
Wilks, S. S., 1938, Annals of Mathematical Statistics **9**, 60.
Willinger, W., and V. Paxson, 1998, Notices of the American Mathematical Society **45**, 961.
Yook, S.-H., Z. N. Oltvai, and A.-L. Barabási, 2004, Proteomics **4**, 928.