Statistical Science
2005, Vol. 20, No. 1, 68–88
DOI 10.1214/088342304000000215

Strong, Weak and False Inverse Power Laws

Richard Perline

Abstract. Pareto, Zipf and numerous subsequent investigators of inverse power distributions have often represented their findings as though their data conformed to a power law form for all ranges of the variable of interest. I refer to this ideal case as a *strong* inverse power law (SIPL). However, many of the examples used by Pareto and Zipf, as well as others who have followed them, have been truncated data sets, and if one looks more carefully in the lower range of values that was originally excluded, the power law behavior usually breaks down at some point. This breakdown seems to fall into two broad cases, called here (1) *weak* and (2) *false* inverse power laws (WIPL and FIPL, resp.). Case 1 refers to the situation where the sample data fit a distribution that has an approximate inverse power form only in some upper range of values. Case 2 refers to the situation where a highly truncated sample from certain exponential-type (and in particular, "lognormal-like") distributions can convincingly *mimic* a power law. The main objectives of this paper are (a) to show how the discovery of Pareto–Zipf-type laws is closely associated with truncated data sets; (b) to elaborate on the categories of strong, weak and false inverse power laws; and (c) to analyze FIPLs in some detail. I conclude that *many*, but *not all*, Pareto–Zipf examples are likely to be FIPL finite mixture distributions and that there are few genuine instances of SIPLs.

Key words and phrases: Pareto–Zipf laws, Pareto distribution, lognormal distribution, extreme value theory, mixture distributions, Gumbel distributions.

1. INTRODUCTION

Empirical distributions that conform approximately to an inverse power form are now widely recognized as common occurrences in numerous scientific disciplines. Probably the first examples of this type came from Pareto's (1897) investigations of personal income distributions. It is also fitting to acknowledge the significant contribution of the philologist George Kingsley Zipf (1949), whose compilation of many examples published in his book, *Human Behavior and the Principle of Least Effort*, was called by Kendall (1961) "one of the most fascinating quantitative studies ever done."

Richard Perline is a Senior Scientist at NuTech Solutions, 34-50 80th Street, Jackson Heights, New York 11372, USA (e-mail: rich.perline@nutechsolutions.com).

Yet even from the beginning, certain considerations suggested that the story behind Pareto–Zipf laws is more complicated. In particular, questions with regard to how well an inverse power form fits the data have frequently been raised, and I believe much misunderstanding and controversy stem from the poorly appreciated role that truncation and some associated difficulties have played. To explain this, I have organized this paper around the ideas of *strong*, *weak* and *false* inverse power laws, abbreviated SIPL, WIPL and FIPL. Briefly, the term SIPL refers to the case where an inverse power law fits the full, untruncated range of the distribution of interest; the term WIPL refers to the case where only some upper portion of the distribution follows an approximate inverse power law; and the term FIPL refers to the poorly understood case where the largest observations (extremes) of samples drawn from certain exponential-type—and especially,

"lognormal-like"—distributions can closely mimic an inverse power law in a way that will be made clear.

In Section 2, I discuss the historical significance of truncation and, more generally, some problems with the casual treatment of data in relation to longstanding confusion regarding the exact form of Pareto–Zipf laws. In particular, I observe that when researchers look more carefully in the lower range of values that were originally excluded from their data sets, they almost always abandon the idea of an SIPL.

In Section 3, I study three FIPLs (the lognormal, finite mixtures of lognormals and the Poison–lognormal) in detail using simulations, empirical examples and analytical results. This material seems to be not well known in the literature, so I have devoted the greatest part of this paper to focus on it. Indeed, except for Parr and Suzuki (1973) and Perline (1982), the key simulations I present that show power law mimicry in truncated samples drawn from the lognormal and finite mixtures of the lognormal (as distinct from looking at the parent distribution itself) have not, to the best of my knowledge, appeared elsewhere. Power law mimicry in the truncated Poisson–lognormal distribution (Section 3.3) has not been noted anywhere before, and the proof that the Poisson–lognormal has the same asymptotic extreme value behavior as its underlying mixing lognormal distribution was first given in Perline (1998) and independently, but less directly, in Asumussen, Klüppelberg and Sigman (1999). Very surprisingly, as basic as it is, except for my own work (Perline, 1982), I have never seen the classic asymptotic approximation $E(X_{i:n}) \approx (a_n + b_n\gamma) - b_n H_{i-1}$, $i \ll n$ (as discussed in Section 3), mentioned in connection with power law mimicry. In fact, as noted below, there have been strong statements that seem to deny what the eyes see in Figures 6, 8, 9 and 10 and what this asymptotic approximation begins to help account for. [On the other hand, someone familiar with the broad view of extreme value theory in terms of the generalized Pareto distribution (Embrechts, Klüppelberg and Mikosch, 1997) would probably not be too surprised by anything in Section 3.]

My concluding section (Section 4) also, very briefly, touches upon why I do *not* think that *all* Pareto–Zipf examples can be explained in terms of power law mimicry. This article emphasizes how oversimplification and overstatement in the literature of inverse power laws have produced a confused picture of things, and I certainly do not want to add another layer to the problem. The references in Section 4 to Montroll and Shlesinger (1982, 1983), Reed (2001) and Perline (1996) point to a connected body of work that can be viewed as naturally related to, and an extension of, the discussion of finite mixtures of lognormal distributions in Section 3.2. Details of these connections will be presented elsewhere, but my closing questions/challenges in Section 4 indicate my overall view.

2. COMPLICATIONS AND CONFUSION—SEVERAL EXAMPLES

The distribution Pareto (1895) *first* proposed as a good representation of the distribution of observed personal incomes within a country in modern notation has a probability density function (p.d.f.) $f(y) = (\alpha A^\alpha / y^{\alpha+1})$ for $y \geq A > 0$, $\alpha > 0$, with cumulative distribution function (c.d.f.) $F(y) = 1 - (A/y)^\alpha$. Pareto presented his results a little differently and stated them in terms of the sample tail probability—actually, he used the sample ranks—writing his law in the form (I am changing the notation slightly) $N_y = B/y^\alpha$, where N_y is the number of people with income greater than or equal to y (i.e., the rank from the top of an individual with income y) and where his constant B was not explicitly related to a lower bound on y. This form has an obvious problem at the *low* end of the income distribution where incomes can approach zero (or, indeed, be negative!). This has led to comments such as by Stamp (1914, page 203), who remarked that "...strict applications of Pareto's law fail, otherwise there would be an enormous population far below the subsistence level." As a matter of fact, Pareto (1897) proposed a more general distribution with a tail probability (or survival function or complementary c.d.f.), which is written today (Johnson, Kotz and Balakrishnan, 1994) as $1 - F(y) = Ce^{-\beta y}/(y + C)^\alpha$ $(C, \alpha, \beta, y > 0)$, that avoids this difficulty. He also asserted (Pareto, 1897, pages 305–306) that the β term is usually negligible; it seems to have been needed in only one case in the empirical income statistics he analyzed.

Using the language of the title of this article, I say that Pareto modified his concept of the theoretical form of income distributions from a *strong* inverse power law—the simpler, first distribution given above—to a *false* inverse power law, which corresponds to his more general distribution. I use the term *false* because in today's nomenclature for statistical distributions, Pareto's general form is "exponential type," not a "power law type." Roughly speaking (for now), by this I mean that the dominant characteristic of the tail behavior is driven by the $e^{-\beta y}$ term for large values of y. However, with β close to 0, which he asserted was

usually the case for his data, in samples where $y \gg C$ and βy remains small, the upper tail behavior approximates an inverse power closely for a significant range of y. Ultimately, of course, as $y \to \infty$, the exponential character of the distribution reveals itself. If $\beta = 0$ is taken identically, the distribution takes the form of what I call a *weak* inverse power law, where the upper tail behavior becomes, as $y \to \infty$, asymptotically equivalent to an inverse power law.

Note that, at least implicitly, the three categories, SIPL, WIPL and FIPL, were introduced by Pareto himself. It is now universally acknowledged that income distributions can be described as approximate power laws only in the upper range of incomes and that they deviate substantially from this form in the lower ranges (Arnold, 1983). Mandelbrot (1960) referred to a "weak Pareto law" precisely to describe this situation.

Examining one of Pareto's (1897) original data sets is instructive. Table 1 gives the income statistics for England in 1893–1894 that he used for one of his analyses. For graphical investigations, it is natural to linearize the representation of power laws using log–log plots, as most researchers have been doing since Pareto. If the relationship $N_y = B/y^\alpha$ holds approximately, then $\log y \approx \log B/\alpha - \log N_y/\alpha$, so a plot of $\log y$ against $\log N_y$ looks linear with slope $-1/\alpha$. The log–log plot of the data in Table 1 is shown in Figure 1 and indicates that an approximate inverse power law holds. However, looking more closely at the data, which are based on tax statistics for shop owners and

TABLE 1
One of Pareto's (1897) *original data sets*: *the distribution of incomes in Great Britain in 1893–1894 based on tax statistics for shop owners and professionals whose income exceeded* £150

Income £	**Frequencies**
150	400,648
200	234,185
300	121,996
400	74,041
500	54,419
600	42,072
700	34,269
800	29,311
900	25,033
1,000	22,896
2,000	9,880
3,000	6,069
4,000	4,161
5,000	3,081
10,000	1,104

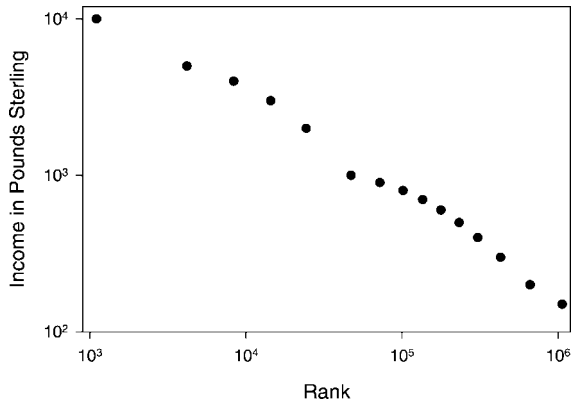

FIG. 1. *A log–log plot of Pareto's original income data shown in Table* 1. *The income statistics are based on approximately* 1,000,000 *individuals in the commercial and professional class whose annual income exceeded* £150 *in the year 1893–1894. The population of Great Britain was around* 30,000,000 *at that time and* £150 *was a relatively high income, so this is a substantially truncated distribution.*

the professional classes, a lower cutoff is observed on income of £150 based on a total count of approximately 1,000,000 individuals, and it is easy to prove that this was a relatively high income threshold. For example, the average annual income for agricultural workers in England and Wales was about £40 in 1893 (Bowley, 1899). Furthermore, the population of England at that time exceeded 30,000,000. How would the distribution look if accurate income data for the lower classes below the £150 threshold were available? The universal income tax in many countries, as well as survey data, have greatly enlarged our view of the matter and confirm that the strong form of Pareto's law does not hold for income distributions collected by modern methods aimed at properly counting all income classes. [Figure 7(a) shows an example that is discussed in the next section.]

This historical example of Pareto's power law is similar in two essential ways to many others that have been discovered: (1) initial research on a truncated data set reveals what is thought to be an SIPL and (2) subsequent investigation deeper into the lower ranges shows that the power law breaks down. As a second example of this sort, consider the city-size law first reported by Auerbach (1913), who showed a good power law fit to the rank ordered populations of the 94 largest German cities from a 1910 census. His article suggests that he viewed his results as supporting an SIPL, but his data fell far short of proving one. By his own estimate, there were something like 100,000 small villages in Germany during this period, yet he explicitly reported

population data only for the top 94 cities. He did observe that the power law still held down to rank 481 for a town with a population of only 10,000 (certainly intending to demonstrate the wide range of validity of the power law), but what about the roughly 99,500 remaining communities? Again, as in Pareto's case, modern studies examining population data for the tens of thousands of small villages existing in a large country have conclusively shown that the power law breaks down beyond the upper tail (Parr and Suzuki, 1973). Note also the switch in the frame of reference here: a town of 10,000 individuals usually is considered small, yet from the wider perspective, it falls in the upper tail.

Most, and perhaps all, of the other well-known power laws also seem to be associated with highly truncated data. Consider, for example, Korčák's (1938) island-area law that asserts that the areas of individual islands in an island chain follow an inverse power law. However, what is an island and what is a rocky outcrop or a little reef? Again, we see a truncation effect that leads to the exclusion of myriad smaller values. Lake systems also follow an inverse power law (Korčák, 1938; Mandelbrot, 1982, page 272), but what is a lake and what is a pond or even a puddle? Similarly, the Gutenberg–Richter (Bak, 1996) power law of earthquake magnitudes involves cutoffs that are as arbitrary as the distinction between a lake and a pond. Inevitably, studies of earthquakes begin from the top down, eliminating some unknown, but vast, proportion of lower energy seismic events.

Figure 2 is another informative example of how researchers can be led astray by truncation. The log–log rank-size graph on the left of Figure 2 plots ingot capacities for U.S. and Canadian steel manufacturing plants in 1954. Simon and Bonini (1958) presented only the first 10 values (enclosed in the rectangle) using published data from a list that gave statistics for the 10 largest U.S. and Canadian steel plants. These same 10 values were then also presented by Kendall (1961) as the first of his several examples of power law distributions in his presidential address before the Royal Statistical Society, and they have been used as an illustrative data set by others, as well. I obtained the remaining 75 values from *all* the listings in the 28th edition

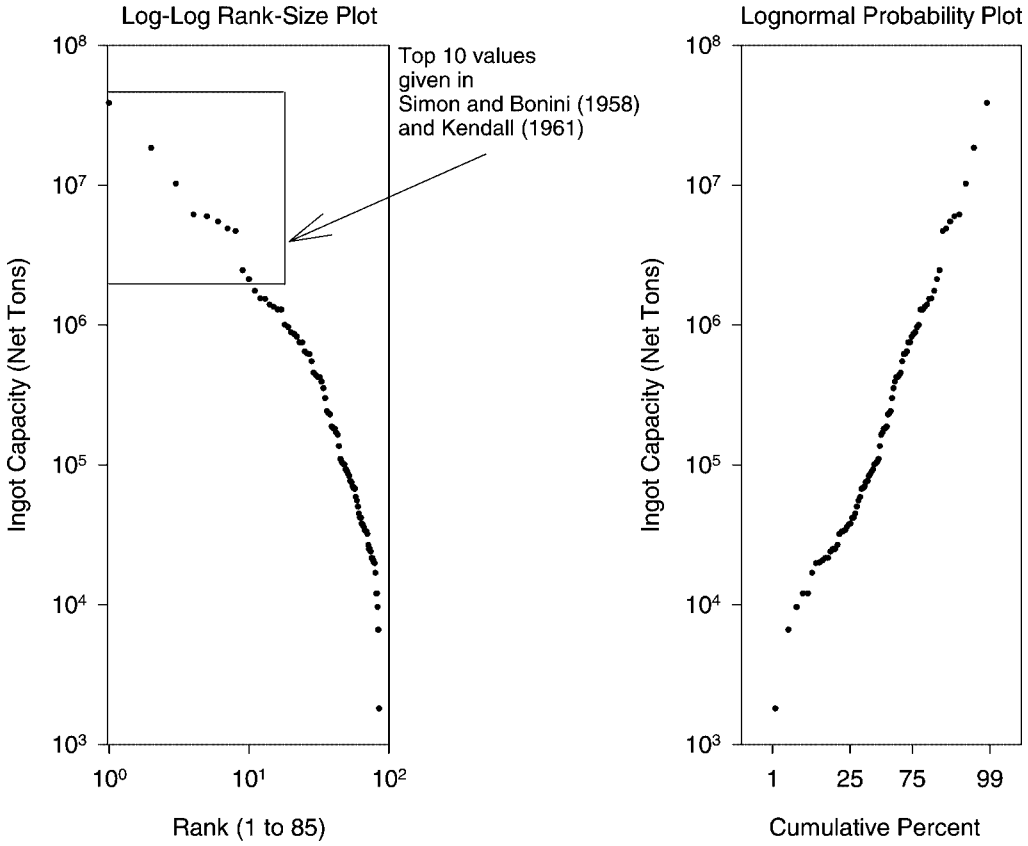

FIG. 2. *Comparison of log–log rank-size plot* (left) *versus lognormal probability plot* (right) *for 1954 ingot capacity data. Only the first* 10 *observations were originally presented by Simon and Bonini* (1958) *and Kendall* (1961). *The departure from the Pareto fit becomes obvious when the rest of the available data are shown. The fit to the lognormal appears reasonably good except for the smallest observation.*

of the *Directory of Iron and Steel Works of the United States and Canada* (American Iron and Steel Institute, 1957, pages 437–439). Looking at the full data set, the inverse power law breaks down, although the lognormal probability plot on the right-hand side of Figure 2 shows an approximate lognormal fit over the entire distribution, except for the smallest data point.

This example shows that Simon was unaware of the significance of truncation in much the same way as Pareto. Simon's (1955) seminal theoretical work on power law distributions (Ijiri and Simon, 1977) carried the parallel even further and illustrated how the three categories of models (SIPL, WIPL, FIPL), implicitly introduced by Pareto (1897), continued to be used casually in a way that invites confusion. First, Simon (1955) stated that the tails of these types of distributions "…can generally be approximated by a function of the form $f(i) = (a/i)^{\kappa} b^{i}$, where a, b, and κ are constants; and where b is so close to unity that in first approximation the final factor has a significant effect on $f(i)$ only for large values of i." [In his notation, $f(i)$, $i = 1, 2, \ldots$, is the sample histogram function that represents counts of the entities of interest, such as words, publications, town populations and personal income.] The convergence factor b is understood to be less than 1, and Simon gave one estimate of its value from a particular data set as 0.999667. The constant b serves the same role, of course, as Pareto's β noted above, and renders $f(i)$ an FIPL. Taken as stated, Simon could be interpreted as meaning that the empirical data fit a power law in approximately some upper, but not-too-upper, part of the distribution.

After suggesting this FIPL approximation, Simon then specified a theoretical model based on a stochastic process that has as its stationary solution the Yule distribution, $f(i) = \rho \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(i+\rho+1)}$, $i = 1, 2, \ldots$, for some constant $\rho > 0$, where $\Gamma(i)$ is the standard gamma function. (Note that this is a one-parameter model and it is truncated at $i \geq 1$.) As Simon remarked, $\Gamma(i)/\Gamma(i+\rho+1)$ is asymptotic to $i^{-(\rho+1)}$ as $i \to \infty$, so that the Yule distribution has an asymptotic power law form and is, therefore, a WIPL. However, he never addressed the problem that $f(i) = (a/i)^{\kappa} b^{i}$ and $f(i) = \rho \frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(i+\rho+1)}$, $i = 1, 2, \ldots$, ultimately have very different tail behavior.

To make things more confusing, the Yule distribution is, as a practical matter, approximately an SIPL. Ijiri and Simon (1977, page 72, Figure 3.1) showed this themselves in one of their own log–log plots, but they did not comment on its consequences. I have generated a similar plot in Figure 3. In this figure, the tail probability $1 - F(i)$ is plotted against i on a log–log scale. [See page 67 of Ijiri and Simon, 1977, for the calculation of $1 - F(i) = \rho B(i, \rho)$.] As their graph and mine make clear, for $0.3 \leq \rho \leq 3$, the distribution is very close to an SIPL. This results in a cloudy state of affairs where, when confronted with empirical data that do not fit an SIPL, Simon and others who have used this model have been forced to ignore or downplay the bottom portion of the distribution. Often, there is some appeal to how the model assumptions are likely

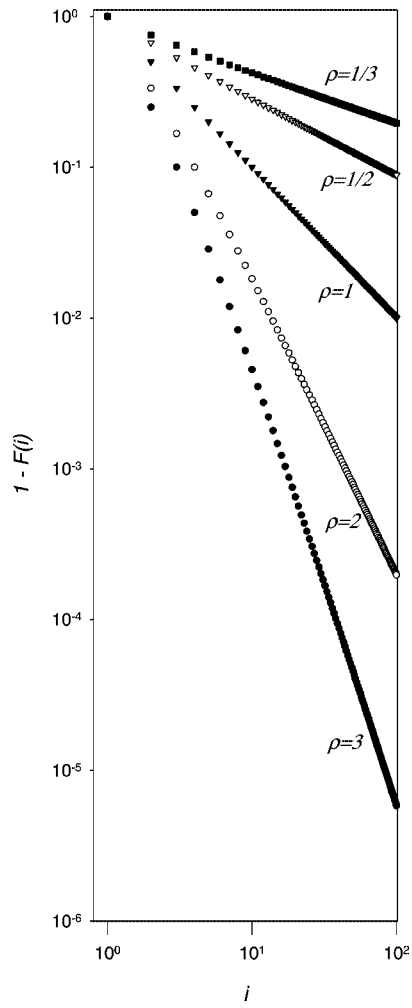

FIG. 3. *A plot similar to that of Ijiri and Simon* (1977, *page* 73) *that shows the complement of the Yule c.d.f.,* $1 - F(i)$, *plotted against* i *in log–log coordinates for different values of the parameter* ρ. *Although it is true that the Yule distribution is only asymptotically an inverse power law, the strongly linear character of the curves makes it clear that, practically speaking, it approximates an SIPL very closely in the range of values of* ρ *likely to be encountered with empirical data.*

TABLE 2
Distribution of lengths of one-way trips extending beyond city limits for truck data used by Zipf (1949, *page* 401). *The raw data are taken from Paddock and Rodgers* (1939, *page* 50, *Table* 9)

Length of one-way trip (miles)	Number of trips	Standardized frequencies	Mid value of trip length interval	Mid value of ranks in interval
0–4.99	184,952	184,952.00	2.5	452,068.5
5–9.99	138,916	138,916.00	7.5	290,134.5
10–19.99	113,521	56,760.50	15.0	163,916.0
20–29.99	41,855	20,927.50	25.0	86,228.0
30–39.99	20,030	10,015.00	35.0	55,285.5
40–49.99	10,262	5,131.00	45.0	40,139.5
50–99.99	23,908	2,390.80	75.0	23,054.5
100–249.99	9,911	330.37	175.0	6,145.0
250–499.99	1,034	20.68	375.0	672.5
500–999.99	110	1.10	750.0	100.5
1000+	45	—	—	—

to fail below an arbitrary threshold value, as in Simon's (1955) comment that for the city-size law, the Yule distribution "…could only be expected to hold down to some minimum size—say, 5000 or 10,000." This floating, arbitrary threshold is determined ad hoc after examining the data and frequently leads to excluding 90% or more of the observations.

In short, Simon derived a theoretical WIPL, which actually approximates a SIPL, to represent empirical data that he believed conformed to a vaguely specified FIPL. It is of course much easier to see these inconsistencies with hindsight, but even current researchers, such as Krugman (1996), who referred to the Simon model as still the best explanation of city-size distributions, seem to be unaware of the serious shortcomings of this one-parameter distribution.

Similar problems can be found in the work of many other researchers in this area. Indeed, Zipf's (1949) book is filled with some odd data sets that look rather different on closer scrutiny. One of these is his (1949, page 401, Figure 9-19) log–log histogram plots on the length of one-way trips made by passenger cars and trucks outside city limits in 11 states in the year 1936. I focus only on the truck data, which are given in Table 2 as taken from a report by Paddock and Rodgers (1939, page 50, Table 9) cited by Zipf as his source. There are at least two difficulties with Zipf's figure: (1) he used some ad hoc constants to shift the horizontal scale with the apparent effect of making his graphs look straighter in log–log form and (2) because the original Paddock and Rodgers data are given using unequal bin intervals, one needs to standardize interval lengths by the counts within the intervals, although it is not clear to me that Zipf did this accurately. My corrected version of the histogram, as given in Table 2, is shown in the log–log plot of Figure 4(a) and certainly does not suggest much of a power law. Figure 4(b), which shows the data plotted in log–log rank-size form using midinterval values and midranks of the intervals as coordinates, tells a similar story. Finally, when the data are graphed as a lognormal probability plot [Figure 4(c)], they conform well to a lognormal model in the part of the distribution above 5 miles (about the top two-thirds of the data), although we do not have a detailed breakdown of the values in the lowest interval (0–5 miles).

A very recent example shows how these kinds of difficulties continue to persist. Research that involves networks of all kinds is leading to exciting discoveries that reveal common mathematical structure across areas as diverse as social, economic and ecological webs, transportation and telephone systems, metabolic nets and the internet. Barabási (2002) provided an interesting popular account by a key investigator in this new cross-disciplinary science. He, as well as many of his colleagues doing similar work, reported an explosion of new power laws, such as with the numbers of in- and out-links of nodes, for the distributions associated with their empirical networks. In some cases, this has become the key focus, and like their predecessors in the area of power laws, they have sometimes been unclear about the form of their models beyond the asymptotics of upper tail behavior and have tended to be casual about specifying what part of their empirical data actually fit a power law. I restrict my comments here to a single data set analyzed by Barabási and his colleagues and reported by them as a power law in the widely

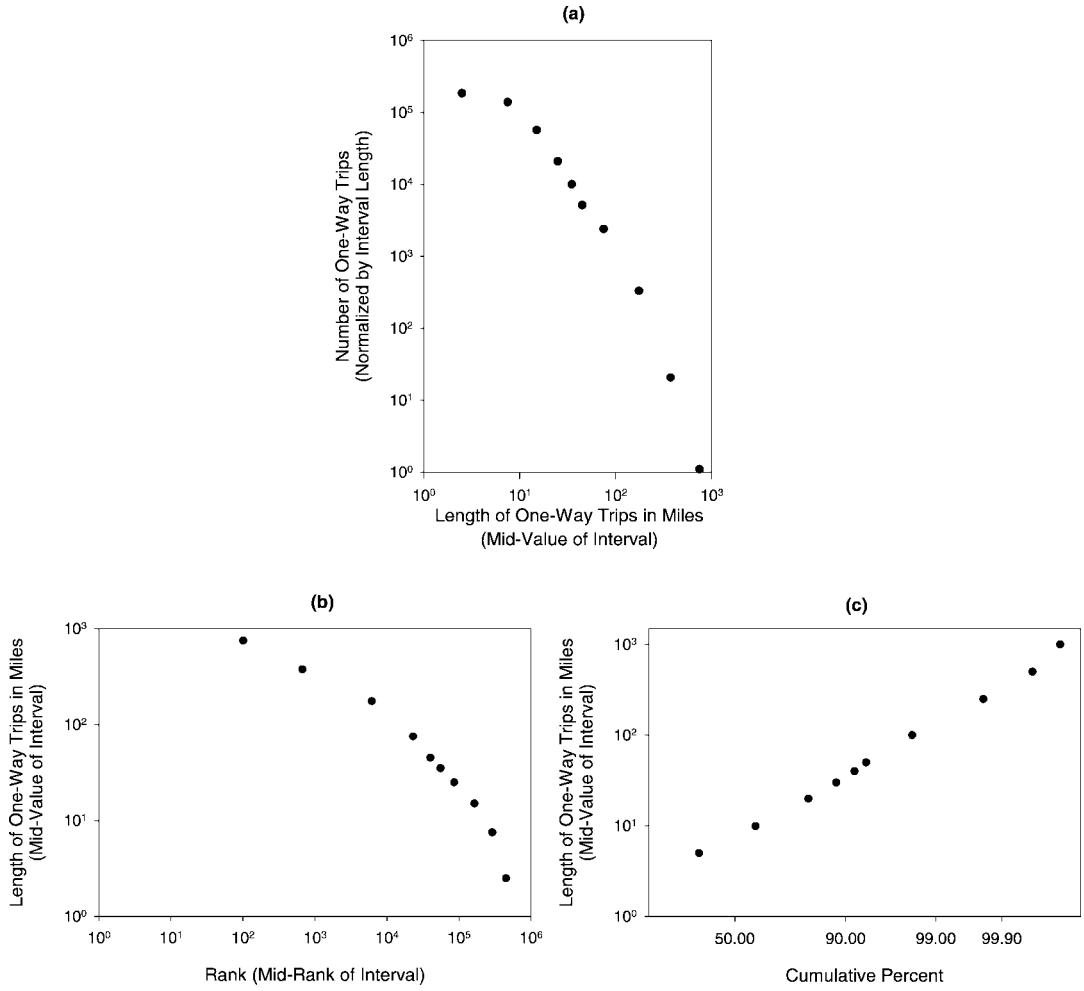

FIG. 4. (a) *Replot of a graph given by Zipf* (1947, *page* 401, *Figure* 9-19) *for truck mileage without using his additive constant for frequencies* (= 10,000). (b) *The data plotted in log–log rank-size form*. (c) *Lognormal probability plot, which shows a good lognormal fit for the part of the data that can be graphed. There appears to be no justification for interpreting the data as a power law.*

read journals *Science* (Barabási and Albert, 1999), *Nature* (Albert, Jeong and Barabási, 1999) and *Scientific American* (Barabási and Bonabeau, 2003).

Barabási and Albert (1999) discussed their finding of a power law distribution for the number of outgoing links from URL documents on the World Wide Web based on a complete map of the nd.edu (Notre Dame University) domain. Figure 1B of Barabási and Albert shows a log–log histogram of outgoing links that is quite linear over the whole distribution, but replotting the data (available in raw form at www.nd.edu/~networks/database/index.html) indicates that the good linear fit over the full range of data is not robust across different binning choices. (I thank Professor Jeong for explaining how geometric binning was used in the original plot, but I was unable to obtain the exact binning intervals used.) When I redo the plot using geometric binning with the intervals 1, 2–3, 4–7, 8–15, ... as given in Table 3, my results as shown in Figure 5(a) do not support an SIPL. The rank-size plot of Figure 5(b) further emphasizes this. In fact, it is as easily argued from the lognormal probability plot in Figure 5(c) that the top 10% of the distribution has a roughly lognormal tail. (I note, by the way, that the data set contains unusual clumping at particular values, with the most obvious case occurring for the number of out-links = 155.) *Furthermore, this is a highly truncated data set because it ignores the very large category of documents that have* 0 *links*! Indeed, as Table 3 shows, more than 58% of the 325,729 URL documents in the data set have no links to other documents. [The log–log histogram plot of the same data in Figure 1A of Albert, Jeong and Barabási (1999) includes the 0-link category by artificially adding 1 to the counts.]

TABLE 3
Distribution of out-link counts for documents on the nd.edu domain as described by Barabási and Albert (1999). *The raw data are available as the World-Wide-Web data set at http://www.nd.edu/~networks/database. The data set comprises* 325,729 *documents with a total of* 1,469,680 *out-links URL's*

Number of out-links	Frequencies	Standardized frequencies	Mid value of number of out-links interval
0	188,795	—	—
1	21,830	21,830.0000	1
2–3	29,647	14,823.5000	2.5
4–7	47,919	11,979.7500	5.5
8–15	19,202	2,400.2500	11.5
16–31	8,709	544.3125	23.5
32–63	6,730	210.3125	47.5
64–127	1,163	18.1719	95.5
128–255	1,568	12.2500	191.5
256–511	89	0.3477	383.5
512–1023	63	0.1230	767.5
1024–2047	11	0.0107	1535.5
2048–4095	3	0.0015	3071.5

Barabási and his colleagues first proposed a theoretical model of the link distribution for networks based on "preferential attachment," which both Watts (2003) and Mitzenmacher (2001) noticed is very similar to Simon's model. Consequently, it may well share similar weaknesses as those discussed above. In subsequent work, however, Barabási and his colleagues (Bianconi and Barabási, 2001) proposed models where, in some circumstances, the distribution of network links no longer even has a power law tail. In just a few short years the subject of power laws in networks has undergone a swift evolutionary recapitulation of earlier work in which initial reports of SIPLs have been followed by more detailed studies that back away from a strong power law model and sometimes even abandon the idea of power law tails. Watts's (2003, page 112) comment that "Some evidence that scale-free (i.e., power law) networks may not be as widespread as they first seemed appeared about a year after Barabási and Albert's original paper" refers to work by Amaral, Scala, Barthelemy and Stanley (2000) that may be the first critical response to claims of strong power laws in this field. Nevertheless, in an introductory article intended for wide popular dissemination, Barabási and Bonabeau (2003) made no mention of these issues. Indeed, similar introductory articles about Pareto–Zipf-type examples have been published through the years, each leaving a new set of readers without any real understanding of the "weak" character of these laws and the significance of truncation.

3. A DETAILED LOOK AT FIPLs

3.1 Power Law Mimicry with the Lognormal Distribution

Closely tied to the difficulties associated with truncated data and the question of the breakdown of SIPLs is the thorny topic of "power law mimicry," that is, the extent to which samples from distributions without a power law tail can nevertheless look like a power law under the right circumstances. Specifically, the question of whether and how samples from the lognormal distribution can mimic power law behavior has been a matter of considerable confusion over the years. This is surprising because it is so easy to show that this can happen—*if* there is substantial truncation.

To varying degrees, a few individuals have touched upon this. Taking a more general view of things, the economist Macauley (1922, page 368) commented long ago in connection with Pareto's income law that "The approximate linearity of the tail of a frequency distribution charted on a doubly logarithmic scale signifies relatively little, because it is such a common characteristic of frequency distributions of many and various types." In addition, Aitchison and Brown (1957, page 101), remarking on some of Zipf's empirical examples in their monograph on the lognormal distribution, stated that "…it is likely that many of these distributions can be regarded as lognormal, or *truncated* lognormal…" (italics added). Parr and Suzuki (1973) made a similar comment:

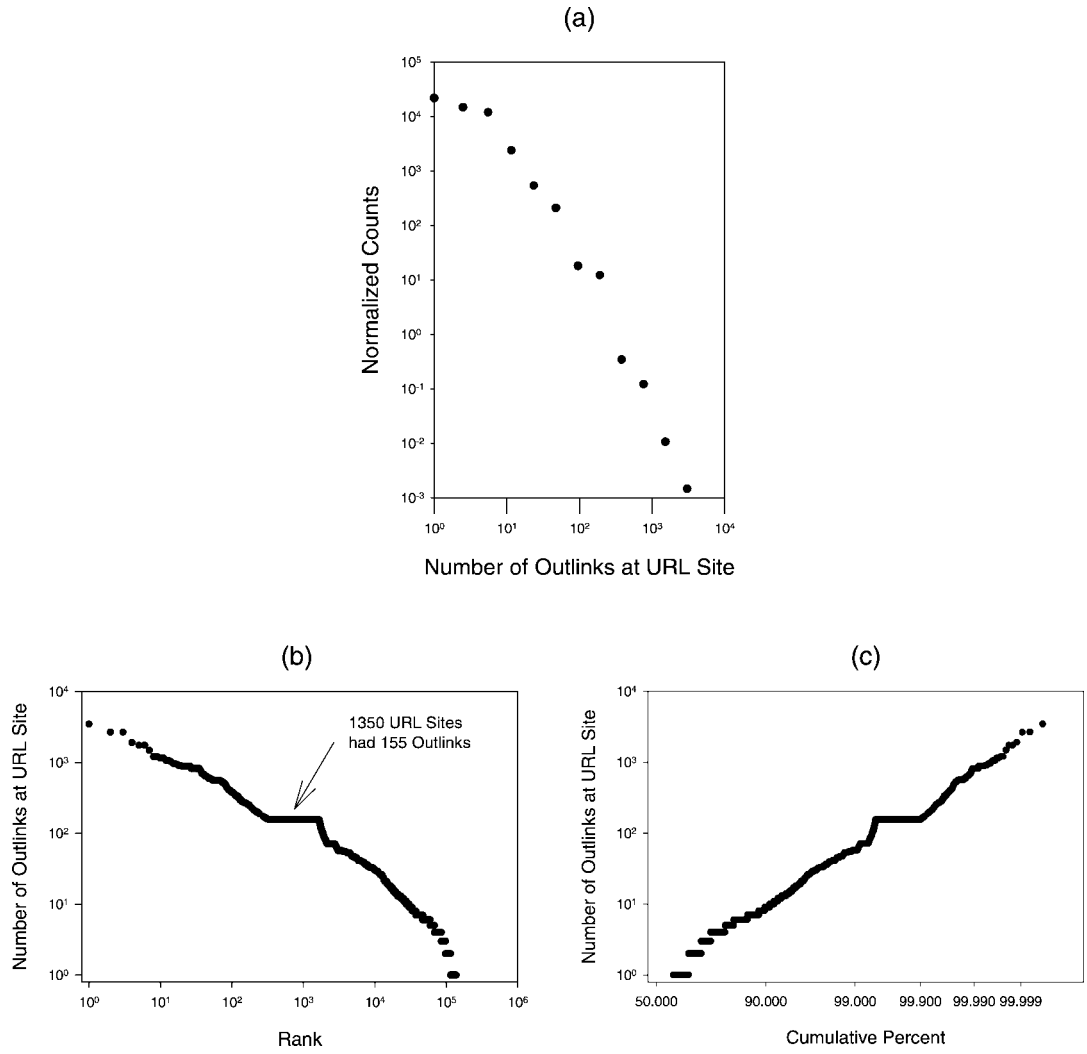

FIG. 5. (a) *Log–log histogram of the outgoing links from URL documents on the nd.edu domain using the data from Barabási and Albert* (1999). *The plot uses geometric binning with the bin intervals and coordinates given in Table* 3. (b) *and* (c) *The same data in log–log rank-size form and as a lognormal probability plot, respectively. URL sites with* 0 *out-links are excluded*; *thus only the top* 42% *of the data is shown. It is easily arguable from plot* (c) *that the distribution has a roughly lognormal tail, not a power law tail. Log–log histogram plots such as* (a) *can look more or less linear, depending on bin interval choices. For example, compare Figure* 1B *of Barabási and Albert* (1999) *with* (a) *above.*

"... truncation of the lognormal distribution at an appropriately high level enables the truncated portion to be regarded as not significantly different from the rank-size distribution (i.e., inverse power distribution)." They also presented numerical and graphical results that convincingly illustrated their assertion. Montroll and Shlesinger (1982, 1983) noted that the lognormal p.d.f. $f(y) = 1/(y\sqrt{2\pi\sigma^2})\exp(-(\log y - \mu)^2/(2\sigma^2))$ $(y, \sigma^2 > 0, -\infty < \mu < \infty)$, which is so different from the Pareto p.d.f. that it is puzzling how the two distributions could ever be confused with each other, looks approximately proportional to $1/y$ for large σ over "an intermediate range" (Montroll and Shlesinger, 1982) or a "certain range" of y (Montroll and Shlesinger, 1983). In more recent work, Gong, Liu, Misra and Towsley (2001), Mitzenmacher (2001) and Downey (2003) have all commented on the difficulties associated with comparing lognormal and power law tails.

Arguing against the significance of lognormal power law mimicry, Mandelbrot (1997, page 206) is quite dismissive of Aitchison and Brown's comment above. In his article titled "A Case Against the Lognormal Distribution," Mandelbrot showed a graph (1997, page 254, Figure E9-1) of a log–log plot of a sample of 9000 observations drawn from a lognormal distribution that

is supposed to reveal how the lognormal can *slightly* resemble a power law—but Mandelbrot's main point seems to be to emphasize how poor the resemblance is. Because he used an *untruncated* sample for his demonstration, however, he missed the point of how the lognormal and certain other related distributions can very convincingly pass for an inverse power law *when sufficient truncation is present*. This failure to appreciate the importance of truncation in the lognormal versus power law issue is longstanding. For example, Ijiri and Simon (1977, page 4) asserted that there is some difficulty in distinguishing a lognormal from a Pareto sample, but except for very small sample sizes, this is simply not the case unless the lognormal data are highly truncated.

The approach I take here, which emphasizes the study of sample extreme order statistics through simple graphical and numerical experiments, on the one hand, and analytic results, on the other, gives an easily understood picture of what is happening. Moreover, the asymptotic theory I use leads to insights about the Pareto-mimicking potential of other distributions besides the lognormal, the point that Macauley (1922) touched upon but which has not previously been given any analytic explanation.

The simulation results graphically represented in Figure 6 are my starting point. Let $Y_1, Y_2, \ldots, Y_n$ be n independent observations drawn from the basic Pareto distribution with c.d.f. $F(y) = 1 - (A/y)^\alpha$. These n observations rank ordered as $Y_{1:n} \geq Y_{2:n} \geq \cdots \geq Y_{n:n}$ are the order statistics of the sample (David, 1970). For my purposes, I say that a sample satisfies an *approximate inverse power law* if its order statistics in some approximate sense satisfy the relationship $Y_{j:n} \approx c_n/j^\beta$, $j = 1, \ldots, n$ ($\beta > 0$, $c_n > 0$). That is, if $Y_{j:n} \approx c_n/j^\beta$, then a plot of $\log Y_{j:n}$ on the vertical axis against $\log j$ on the horizontal axis should be approximately linear with slope $-\beta$. This linearization simplifies matters; consequently, I chose to work with the log-transformed data.

Why does the ordered sample of 1000 observations drawn from a Pareto distribution produce a linear log–log plot like that of curve 1 in Figure 6? The answer is given in terms of expectations of order statistics. If Y has the Pareto distribution $F(y) = 1 - (A/y)^\alpha$, then $X = \log Y$ has the c.d.f. $G(x) = 1 - A^\alpha e^{-x\alpha}$ and p.d.f. $G'(x) = g(x) = A^\alpha \alpha e^{-x\alpha}$ ($x \geq \log A$), that is, X is exponentially distributed. Therefore, any analysis using log-transformed Pareto order statistics, $\log Y_{j:n} = X_{j:n}$, actually can be thought of directly in terms of the exponential order statistics $X_{j:n}$. A key result (David, 1970) that pertains to the moments of exponential order statistics gives

$$E(X_{j:n}) = \frac{1}{\alpha}\sum_{r=j}^{n}\frac{1}{r} + \log A = \frac{1}{\alpha}(H_n - H_{j-1}) + \log A, \tag{1}$$

where H_j is the sum of the first j terms of the harmonic series and $H_0 = 0$. Detailed analysis of H_j (Graham, Knuth and Patashnik, 1994, page 278) shows that for any positive integer j,

$$H_j = \log j + \gamma + \frac{1}{2j} - \frac{1}{12j^2} + \frac{\varepsilon_j}{120j^4}, \tag{2}$$

where $\gamma = 0.577\ldots$ (Euler's constant) and $0 < \varepsilon_j < 1$ (all logarithms are to base e, unless otherwise indicated). From (1) and (2) and the fact that $\log(n+1) > \log n + (n+1)^{-1}$, the following bounds on $E(X_{j:n})$ are obtained (the lower bound is valid for $1 \leq j \leq n$ and the upper is valid for $2 \leq j \leq n$):

$$\frac{1}{\alpha}[\log n - \log j] + \log A < E(X_{j:n}) < \frac{1}{\alpha}[\log(n+1) - \log(j-1)] + \log A. \tag{3}$$

So $E(X_{j:n})$ can be approximated for most practical plotting applications by the value of its lower bound, $\log A + \frac{1}{\alpha}\log n - \frac{1}{\alpha}\log j$, which has the desired form $c'_n - \beta \log j$, where $\beta = 1/\alpha$ and $c'_n = \log A n^{1/\alpha}$.

Given the order statistics $Y_{j:n}$ for a sample of positive values, a plot of $\log Y_{j:n} = X_{j:n}$ against $\log j$ is called a *log–log rank-size* plot of the sample observations. With $A = \alpha = 1$, a log–log rank-size plot of the Pareto order statistics, which, as just seen, actually involves exponential order statistics, tends to produce a nearly straight line with intercept approximately $\log n$ and slope approximately -1. The lowest curve (labeled 1) in the graph on the left of Figure 6 shows such a plot for $n = 1000$ and $j = 1$ to 1000 (i.e., an untruncated sample). The curves above this one show n varied from 10,000 to 1,000,000 with j still between 1 and 1000. Thus, the second curve up (labeled 2) shows the top 1000 order statistics from a Pareto sample of $n = 10{,}000$ observations (i.e., a 90% truncated sample). As the sample size increases in the curves moving up, the degree of truncation also increases. Only the intercepts are changing, as the slopes all remain close to -1 since $E(\log Y_{j:n}) \approx \log n - \log j$.

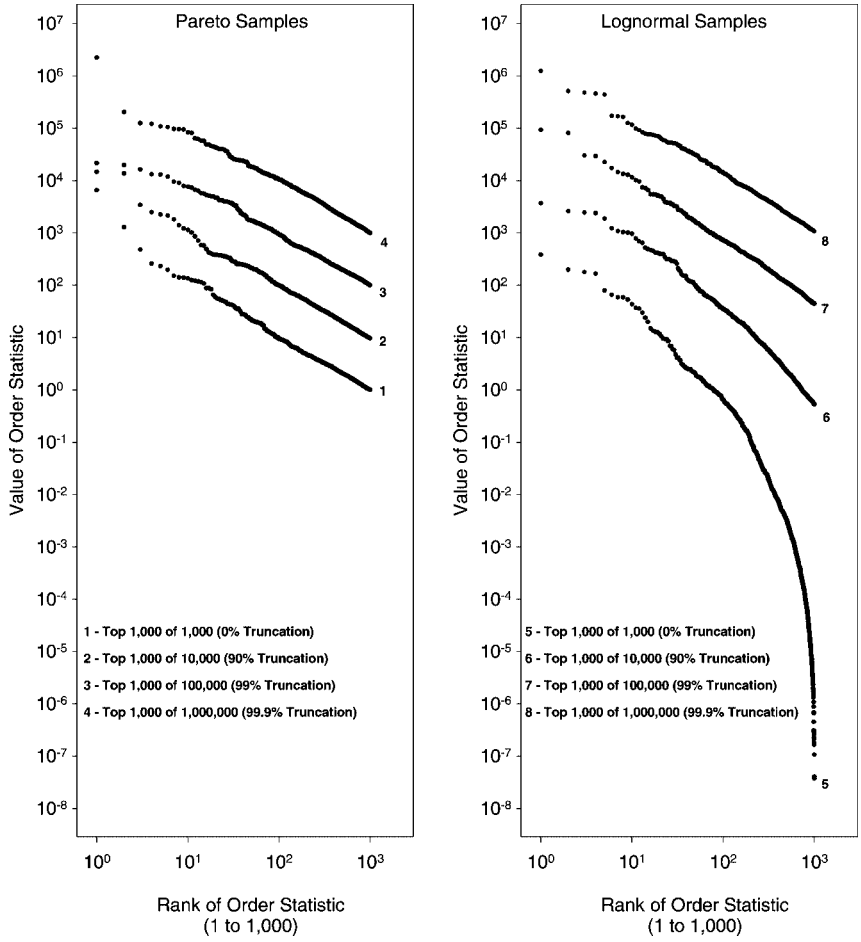

FIG. 6. *Log–log rank-size plots that compare the top order statistics from simulated Pareto and lognormal samples varying from* 0 *to* 99.9% *truncation and sample sizes varying from* 1000 *to* 1,000,000. *Note that lognormal samples with a high degree of truncation* (*plots labeled* 7 *and* 8 *on the right*) *produce quite linear log–log plots that essentially are indistinguishable from those of the Pareto samples on the left.*

It is the counterintuitive behavior demonstrated on the log–log plots on the right-hand side of Figure 6 that is most interesting. I emphasize *counterintuitive* because even experienced statisticians question what is shown, and the story that unfolds in the graphs has been poorly understood by many investigators. Indeed, it appears that the only similar published results, *based on truncated sample data*, are those of Parr and Suzuki (1973) and Perline (1982). In the plots in Figure 6, Y is lognormal [denoted $Y \sim \mathrm{L}(\mu, \sigma^2)$] with parameters $\mu = -5.79$, $\sigma^2 = 16.9$; that is, $X = \log Y$ is normally distributed $\mathrm{N}(-5.79, 16.9)$. The lowest curve (labeled 5) shows the values of an untruncated sample of 1000 observations. As the amount of truncation increases from the 1000 largest of 10,000 observations (90% truncation) in the second curve labeled 6 to the largest 1000 of 1,000,000 observations (99.9% truncation) in the highest curve labeled 8, the upper tail of the lognormal sample produces an increasingly straight log–log rank-size plot. The largest 1000 of 1,000,000 observations of the lognormal sample are for most practical purposes indistinguishable from the Pareto plots on the left and therefore closely mimic an inverse power law. This set of plots is much more revealing than Mandelbrot's (1997, page 254) plot, which does not examine or mention the effects of truncation. Furthermore, there is no confusing the *untruncated* Pareto and lognormal samples (labeled 1 and 5 in the bottom curves on the left and right of Figure 6), although the discussion in Ijiri and Simon (1977, page 4) suggested that distinguishing between a lognormal and a power law would be a difficult task. Again, they make no mention of truncation as the real culprit responsible for the difficulty.

What explains the Pareto-mimicking behavior of the upper tail of lognormal samples? Standard asymptotic estimates from the classic theory of extreme order statistics provide helpful insights, and because the theory

is applicable to a large class of distributions, other distributions that can exhibit this behavior can be found as well. The behavior of the order statistics of the exponential (i.e., log-transformed Pareto) distribution turns out to be, in a certain sense that needs to be carefully defined, typical of a surprisingly large class of distributions called *exponential type* or *Gumbel type*, provided only the top order statistics or largest extremes are considered. The Gumbel-type distributions of interest here are, essentially, all distributions with c.d.f. $F(y)$ and p.d.f. $f(y) = dF(y)/dy$ that satisfy the von Mises condition $\lim_{y\to\infty} \frac{d}{dy}\frac{1-F(y)}{f(y)} = 0$, which includes the exponential, normal, lognormal, gamma and Weibull distributions; however, my results also are extended to a certain class of integer-valued distributions that are closely related to Gumbel types, as subsequently discussed. All of these distributions, including the integer-valued ones that are considered, are said to belong to the Gumbel domain of attraction because the asymptotic behavior of their upper extreme order statistics can be characterized in the same general way. I briefly state the standard results required for the asymptotic estimates that are used here. These can be found in Embrechts, Klüppelberg and Mikosch (1997), unless otherwise indicated.

Suppose $Y_{j:n}$ is the jth order statistic of a sample drawn from a parent distribution such that $\log Y = X$ is Gumbel type. In the case of Pareto Y, X is exponential, which is Gumbel type. In the case of lognormal Y, X is normally distributed and so also Gumbel type. The expectations of normal order statistics do not have the simple form of the exponential, but for the largest order statistics, the asymptotic theory provides estimates of $E(X_{j:n})$ for the situation of fixed $j \ll n$ that help to explain the log–log linear results in the upper curves on the right of Figure 6.

First fix j. Then for X Gumbel type with c.d.f. $F(x)$, there exist two sequences of standardizing constants a_n and b_n, depending on $F(x)$, such that each of the standardized variables $(X_{i:n} - a_n)/b_n$, $1 \le i \le j$, as $n \to \infty$, converges in distribution to the limiting distribution with c.d.f. $G_i(x)$,

$$G_i(x) = \exp(-e^{-x}) \sum_{k=0}^{i-1} \frac{1}{\Gamma(k+1)} e^{-kx}, \tag{4}$$

and limiting p.d.f.

$$g_i(x) = \frac{d}{dx} G_i(x) = \frac{1}{\Gamma(i)} \exp(-ix - e^{-x}). \tag{5}$$

With $p > 0$ for well-behaved distributions like the abovementioned Gumbel types, Polfeldt (1970) justified the limiting moment convergence

$$\lim_{n\to\infty} E\left[\left|\frac{(X_{i:n} - a_n)}{b_n}\right|^p\right] = \int_{-\infty}^{\infty} |x|^p g_i(x)\, dx. \tag{6}$$

For $p = 1$, I get

$$\begin{aligned} &\lim_{n\to\infty} E\left[\frac{(X_{i:n} - a_n)}{b_n}\right] \\ &= \frac{1}{\Gamma(i)} \int_{-\infty}^{\infty} x \exp(-ix - e^{-x})\, dx \\ &= \gamma - \sum_{k=1}^{i-1} \frac{1}{k} = \gamma - H_{i-1} \quad (= \gamma \text{ for } i = 1). \end{aligned} \tag{7}$$

In general these standardizing constants can be computed from the relationships

$$F(a_n) = 1 - \frac{1}{n} \quad \text{and} \quad b_n = \frac{1}{nf(a_n)}. \tag{8}$$

Therefore, for Gumbel-type distributions with large n and $i = 1, \ldots, j \ll n$, $E(X_{i:n}) \approx (a_n + b_n\gamma) - b_n H_{i-1}$. Since H_{i-1} is close to $\log i$, a plot of $X_{i:n}$ against $\log i$ looks linear with intercept $(a_n + b_n\gamma)$ and slope $-b_n$.

Before looking more closely at the case of the normal distribution, for an instructive comparison the simpler case of the exponential is examined. Recall from (1) that if X is exponentially distributed with p.d.f. $\alpha A^\alpha e^{-\alpha x}$, then $E(X_{i:n}) = \log A + H_n/\alpha - H_{i-1}/\alpha$ *exactly* for *all* i, $i = 1, 2, \ldots, n$. Now to compare this exact result with the asymptotic approximation from extreme value theory, compute $a_n = \log A + \frac{1}{\alpha}\log n$ and $b_n = \frac{1}{\alpha}$ from the equations in (8). Then the asymptotic approximation for the ith top extreme for $i \ll n$ is $E(X_{i:n}) \approx (\log A + \frac{1}{\alpha}\log n + \frac{\gamma}{\alpha}) - \frac{1}{\alpha}H_{i-1}$. Because $\lim_{n\to\infty} H_n - \log n = \gamma$, the difference between the exact and asymptotic estimates goes to 0 for any fixed i as $n \to \infty$. In this case $b_n = 1/\alpha$ is a constant, and therefore plots of $X_{i:n}$ against $\log i$ tend to be approximately linear with slope $-1/\alpha$ *independent of the sample size* n, as we saw on the left-hand side of Figure 6.

Now consider the log-transformed lognormal variable, that is, the normal case represented graphically on the right-hand side of Figure 6. For $\log Y = X \sim \mathrm{N}(\mu, \sigma^2)$, the standardizing constants for the top j order statistics $X_{i:n}$, $1 \le i \le j$, can be computed as (see Embrechts, Klüppelberg and Mikosch, 1997, page 145)

$$\begin{aligned} a_n &= \mu + \sigma(2\log n)^{1/2} \\ &\quad - \sigma\frac{\log\log n + \log 4\pi}{2(2\log n)^{1/2}} \quad \text{and} \\ b_n &= \frac{\sigma}{(2\log n)^{1/2}}. \end{aligned} \tag{9}$$

Consequently, using the asymptotic approximation in this normal case, for $i = 1, \ldots, j \ll n$, $X_{i:n}$ plotted against $\log i$ is expected to give an approximately linear trend with intercept approximately $\mu + \sigma(2\log n)^{1/2}$ and slope approximately $-\sigma/(2\log n)^{1/2}$. Observe that the slope in this case is not a constant independent of n, as in the exponential case, but is of order $O(1/\sqrt{\log n})$ and so *slowly* goes to 0. Therefore, the lognormal upper extremes should be described as only mimicking a true power law because the slopes, not just the intercepts, of the log–log rank-size plots of extremes vary with n. (See the Appendix for further discussion of these estimates.)

It is also worth remarking that on a log scale the restriction $j \ll n$ can be less severe than might be supposed. If $j = 10^3$ and $n = 10^6$ as in curve 8 of Figure 6, then j is small relative to n, but on the log scale of the x axis, the ranks 1–10^3 span the same linear length as the ranks 10^3–10^6. That is, at least $1/2$ of the span of the curve on the x axis looks quite straight.

3.2 Power Law Mimicry with Finite Mixtures of Lognormal Distributions

Various researchers have commented on the tendency toward oversimplification in the quest for models of Pareto–Zipf-type distributions. The fact is that when domain experts bore in on the empirical data of their specialty, whether incomes, community populations, earthquakes or the numbers of publications of scientists, they can usually point to many sources of heterogeneity that support the idea of discrete subpopulations likely to differ in important characteristics. This motivates the logic of using *mixture distribution* models.

In the case of income distributions, Lebergott (1959), among many others, discussed the messy realities of segments of the population that consist of part-time and seasonal workers, the semiretired, different age and sex categories, members of the armed forces and so forth. The data for the U.S. depression-era income distribution graphed in Figures 7(a) and (b) were collected with survey information from some 729 component subpopulations with strong regional, occupational and rural–urban differences (National Resources Committee, 1938). The log–log rank-size plot of Figure 7(a) indicates that the depression-era income data do not fit a power law over the whole distribution. In addition, the income data do not fit a lognormal model, as is clear from the lognormal probability plot of Figure 7(b). It should not be surprising that simple models like the lognormal or Pareto or Yule–Simon do not fit well over the full range of values for distributions that comprise so many distinct subgroups.

The form of the income distribution plotted in Figure 7(a) and (b) is so common that it has led to the general opinion expressed by Klein (1962) that there "is a tendency towards the view that the Pareto distribution gives a better explanation of the upper tail and that the lognormal distribution gives a better explanation at lower income values." Figures 7(c) and (d) show a similar-looking distribution for data on the lengths of articles in the 11th edition of the *Encyclopedia Britannica*. Zipf (1949, page 177, Figure 5-3) exhibited a log–log histogram plot of a sample of articles from this encyclopedia, but again, his selection of bin intervals seems to have enhanced the linearity of the plot. Because of the continuous character of the lengths data, a log–log rank-size plot is preferred. He did not publish his raw data, but I have collected a sample of article lengths from the same edition of this encyclopedia by beginning with the first article starting on or after page 500 and continuing for a total of 20 articles for each of the 28 volumes. (My sample differs from Zipf's only in that he continued for 50 pages after page 500 for each volume, while for convenience, I took a fixed 20 articles from each volume.) The lengths in column inches for all 560 articles that I measured are shown as the log–log rank-size plot of Figure 7(c). This departs significantly from linearity, in contrast to Zipf's log–log histogram plot, which seems to have only one deviant point. It is also obvious from Figure 7(d) that the data do not fit a lognormal model either.

A third, neither Pareto nor lognormal distribution is shown in Figures 7(e) and (f). This is another data set that Zipf (1949, page 381, Figure 9-7) presented as a binned log–log histogram, but which I am graphing here as a log–log rank-size plot. The exact source of his data is some census data prepared by Edwards (1943, Table 2, pages 49–58), as cited in Zipf (1947). These census data give the counts of individuals working in 450 occupational categories. Zipf (1947, 1949) plotted the data for the 450 "Specific classes" and also for aggregated "Generic classes" that come from the same table. The plots in Figures 7(e) and (f) force us to reject the notion that the data could be as simply described as either Pareto or lognormal.

For income data there is evidence that when more homogeneous segments are examined, a lognormal fits well within each segment. For example, this is true for the distribution of weekly wages for full-time male manual workers in England beginning with the first wage survey of 1886 down to modern times (Thatcher,

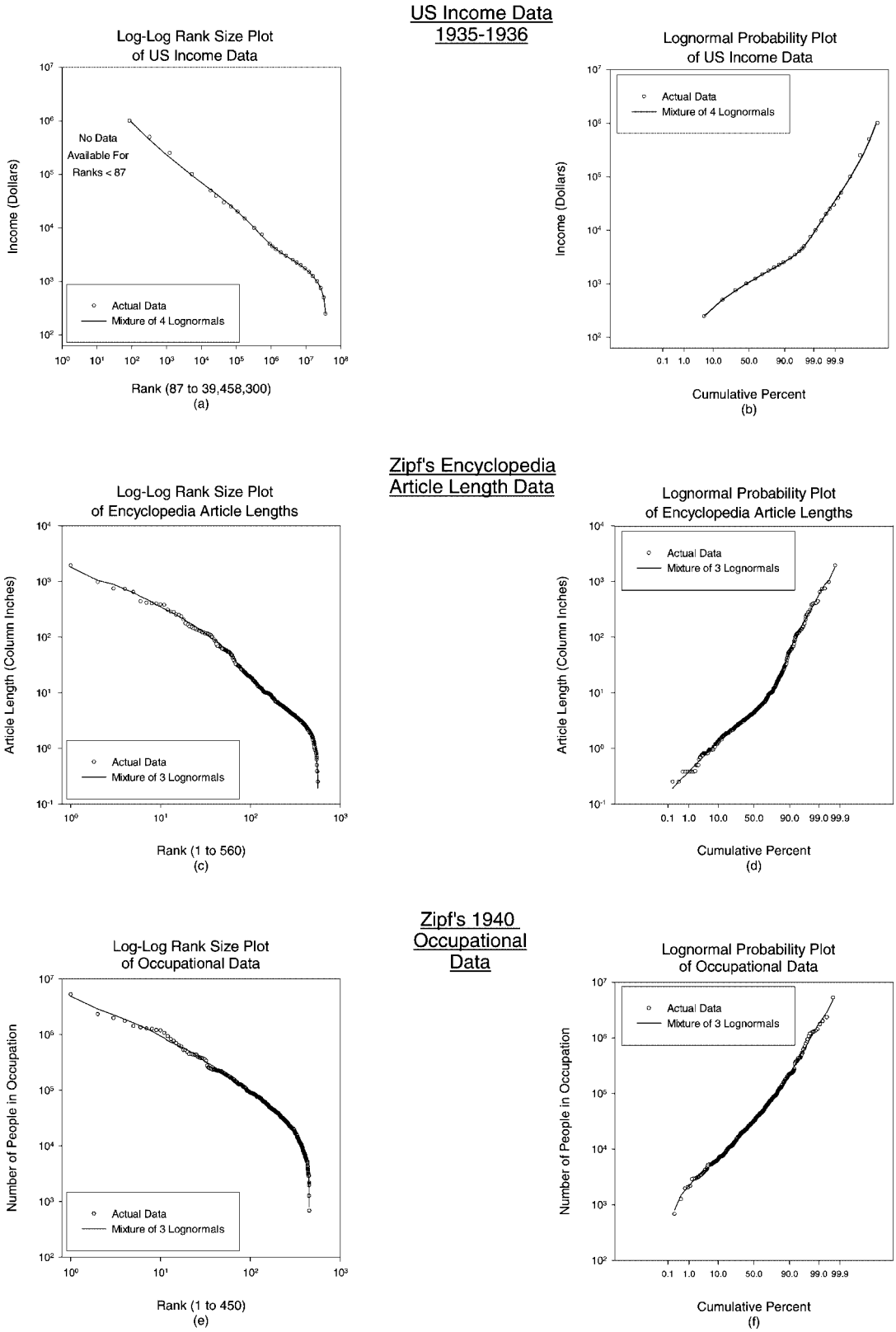

FIG. 7. *Fitting finite mixtures of lognormal distributions. None of these three examples fits a Pareto distribution, as is shown by the departures from linearity in the log–log plots of* (a), (c) *and* (e). *Similarly, the deviations from linearity in the plots of* (b), (d) *and* (f) *show that no examples fit a lognormal distribution. However, mixtures of lognormals, represented by the curves drawn through the data points, fit well.*

1976). The corresponding distribution for women, although it shows a large difference in earnings from the men, is nevertheless also approximately lognormal when separated out. The French economist Gibrat (1931) was the first to remark on the lognormal character of the distribution of English wages, but there is practically no overlap between the incomes of the segment of the population he studied and the wealthier English population in Pareto's data in my Table 1.

These examples point to the reasonableness of using finite mixtures of lognormals to model income and other Pareto–Zipf-type distributions. In this scheme, a model is assumed in which the random variable Y_{mix} has c.d.f. $F_{\text{mix}}(y)$ defined by $F_{\text{mix}}(y) = \sum_{i=1}^{n} p_i F_i(y)$, where the $p_i > 0$ are mixing proportions, $\sum_{i=1}^{n} p_i = 1$, and the component c.d.f.'s $F_i(y)$ correspond to those of n distinct lognormals $\mathrm{L}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. I used the expectation–maximization (McLachlan and Krishnan, 1997) algorithm to fit this model to each of the three data sets of Figure 7. The resulting fitted distributions are drawn through the plotted points in all the graphs and can be seen to fit very well. Quite recently, Downey (2003) also made the point that mixtures of lognormals can do a good job of fitting data (in his case, computer file size distributions) that others have put forth as having power law tails.

It is an easy matter to do simulations that show that sufficiently truncated samples from mixtures of lognormals can mimic a Pareto distribution. To do this, I repeated the earlier truncation experiment in Figure 6, but now using samples from a mixture of three lognormal distributions. For the experiments whose results are plotted in Figure 8, the mixture parameters were $(p_1, p_2, p_3) = (0.33, 0.33, 0.34)$, $(\mu_1, \mu_2, \mu_3) = (-6, -1, -7)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (5^2, 2^2, 1^2)$. A random sample of 1000 observations drawn from this mixture is graphed in the lognormal probability plot of Figure 8(a). The decidedly nonlinear plot would not be confused with a single lognormal sample, but the effects of increasing truncation displayed in the log–log rank-size plot of Figure 8(b) have basically the same increasingly linear appearance as the truncated samples from the single lognormal that was shown in Figure 6(b). (However, I also add the observation that my simulation results indicate a definite tendency for samples of mixtures of lognormals to exhibit Pareto-mimicking behavior over a *greater* range of the upper tail than a sample from a single lognormal.) It is not difficult to show that mixtures of normally distributed random variables are Gumbel-type distributions, so the asymptotic theory discussed in the previous section can also help motivate the Pareto-mimicking behavior displayed in the top curves labeled 2–4 in Figure 8(b).

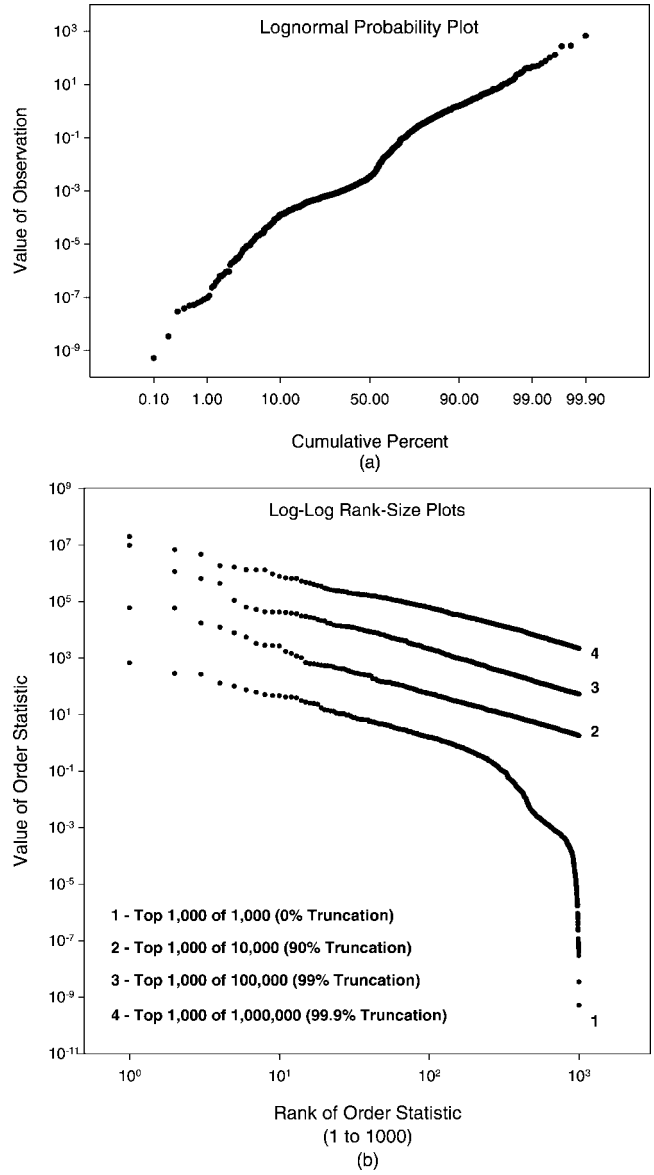

FIG. 8. (a) *Lognormal probability plot of a sample of* 1000 *drawn from a mixture of three lognormal distributions as described in the text.* (b) *The effects of increasing truncation in this mixture distribution. Pareto-mimicking behavior is obvious in curves* 2–4. *The data plotted in the untruncated curve labeled* 1 *are the same as those plotted in the lognormal probability plot of* (a).

3.3 Power Law Mimicry with the Poisson–Lognormal Distribution

Pareto-mimicking behavior is also easy to observe in another lognormal-like distribution, the Poisson–lognormal, and its *zero-truncated* form will be seen to be particularly relevant. The general idea of using mixed Poisson models to fit some of the integer-valued

distributions associated with power laws has been suggested on several occasions (Sichel, 1975; Bookstein, 1997). First consider the general definition of a mixed Poisson distribution: Start with a Poisson random variable with parameter $\lambda > 0$ and then let λ itself be a random variable whose distribution has the p.d.f. $f(\lambda)$. This generates a mixed Poisson distribution whose p.d.f. is given by

$$\begin{aligned} P_j &= P(Y = j) \\ &= \frac{1}{\Gamma(j+1)} \int_0^\infty e^{-\lambda} \lambda^j f(\lambda)\, d\lambda \qquad (10) \\ &\qquad\qquad (j = 0, 1, 2, \ldots). \end{aligned}$$

The p.d.f. $f(\lambda)$ in (10) is referred to as the *mixing density function*. The classic, tractable case assumes $f(\lambda) = \beta^\gamma \lambda^{\gamma-1} e^{-\beta\lambda} / \Gamma(\gamma)$ $(\gamma, \beta, \lambda > 0)$, the p.d.f. of a gamma distribution, so that the values of P_j have an exact, closed form solution. The resulting distribution, usually called a negative binomial (rather than a "Poisson–gamma") distribution, probably still remains the most commonly applied mixed Poisson distribution function because of its mathematical tractability.

Nevertheless, many other mixing density functions have been investigated. The lognormal p.d.f. has been considered potentially attractive (Bulmer, 1974), but is regarded as difficult to work with because the values P_j cannot, in this case, be computed in closed form. However, heuristic considerations (Bulmer, 1974; Grandell, 1997, page 48) previously suggested that the *tail* of the Poisson–lognormal takes on a simple form. This has now been rigorously shown to be true (Perline, 1998; see also Proposition 3.1 of Asmussen, Klüppelberg and Sigman, 1999) and leads to the conclusion that the Poisson–lognormal is *tail equivalent* to its mixing lognormal and therefore has very lognormal-like tail behavior.

Specifically, let $F(y)$ be the c.d.f. and let $f(y) = F'(y)$ be the p.d.f. of a lognormal distribution $\mathrm{L}(\mu, \sigma^2)$. Write $\mathrm{PL}(\mu, \sigma^2)$ to denote the Poisson–lognormal mixture generated with $f(y)$ as the mixing p.d.f. in (10) and let $F_{\mathrm{PL}}(y)$ be the c.d.f. of $\mathrm{PL}(\mu, \sigma^2)$, that is, $F_{\mathrm{PL}}(y) = \sum_{j=0}^{[y]} P_j$, where $[y]$ is the greatest integer function. Using a powerful asymptotic integral approximation due to Berg (1958), I was able to show that the two c.d.f.'s, $F(y)$ and $F_{\mathrm{PL}}(y)$, are tail equivalent, which is to say that $\lim_{y\to\infty}(1 - F(y))/(1 - F_{\mathrm{PL}}(y)) = 1$. As a consequence, the extreme order statistics of both $\mathrm{L}(\mu, \sigma^2)$ and $\mathrm{PL}(\mu, \sigma^2)$ have the same asymptotic behavior.

Power law mimicking in the Poisson–lognormal is accentuated by the fact that it is often the zero-truncated form of the distribution that is appropriate, because in many empirical situations only the events that occur at least once can be observed. I illustrate how this bears on the problem with two classic integer-valued examples of power laws: the word counts from James Joyce's novel *Ulysses* analyzed by Zipf (1949) and the publication counts of chemists analyzed by Lotka (1926).

I fit the zero-truncated Poisson–lognormal distribution to the *Ulysses* word frequency counts [compiled by Hanley (1937), which was Zipf's source, as well] using the maximum likelihood estimation procedure outlined in Bulmer (1974). When the zero class is unobservable, the untruncated form P_j of the mixed Poisson mass function is modified by the renormalization $P_j^* = P_j/(1 - P_0)$, $j \geq 1$. If P_0 turns out to be large, the situation is, in effect, like observing the upper tail of a truncated lognormal distribution, so that log–log rank-size plots of the data have an appearance similar to the highly truncated lognormal samples of Figure 6.

The results of a simulation graphed in the log–log rank-size plot on the right-hand side of Figure 9 make this concrete. The maximum likelihood estimates of the zero-truncated Poisson–lognormal model yielded $\widehat{\mu} = -5.62$ and $\widehat{\sigma}^2 = 9.75$ for the underlying lognormal. These estimates imply $\widehat{P}_0 \approx 0.931$. Because there are 29,899 distinct vocabulary words (*word types*) in the novel, this gives an estimate of $\widehat{V} = 29{,}899/(1 - \widehat{P}_0) \approx 433{,}319$ distinct words in the author's active vocabulary. (The notion of a finite vocabulary opens up a debate that is secondary to my main point of showing how well this distribution can mimic a power law.) To simulate the observed sample using the fitted model, a value λ_1 was sampled from a lognormal distribution with the fitted parameter values given above. This value was then used as a Poisson parameter to generate a random Poisson variable Y_1. If $Y_1 > 0$, a counter was incremented. This process was then repeated with another sample value λ_2, and so forth, until the counter reached 29,899. To obtain this number, an actual total of 441,418 (not the theoretical 433,319 from above) parameter values $\lambda_1, \ldots, \lambda_{441418}$ had to be sampled and then used to generate 441,418 Poisson random variables, about 6.5% of which were nonzero. The log–log rank-size plot of the simulated sample on the right-hand side of Figure 9 is therefore constructed from the 29,899 nonzero-order

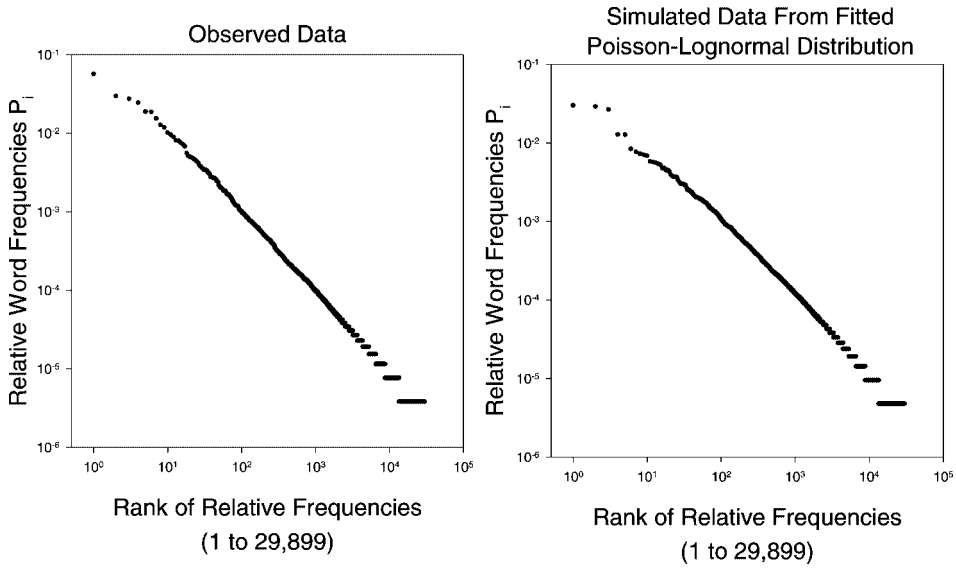

FIG. 9. Left: *Recreation of Zipf's log–log rank-size plot of the word frequencies in James Joyce's novel Ulysses*. Right: *Results of simulating a sample of word frequencies drawn from a zero-truncated Poisson–lognormal distribution fit to the observed data. The Poisson–lognormal distribution is analytically quite different from a power law form, but with sufficient truncation, it can masquerade as a power law, as shown in the plot.*

statistics $Y_{1:n} \geq Y_{2:n} \geq \cdots \geq Y_{29899:n} = 1$, where $n = 441{,}418$.

I also fit the zero-truncated Poisson–lognormal model to Lotka's (1926) famous data on the scientific productivity of chemists. My Figure 10 is like his Figure 2, which plots the log–log histogram compiled from the papers listed in *Chemical Abstracts 1907–1916* authored by 6891 individuals with last names beginning with the letters A and B. As Lotka did, I have plotted only the data for counts between 1 and 30. Because only someone with at least one listed paper appears in the sample, a zero-truncated model is required. The estimated parameters for these data were $\widehat{\mu} = -4.35$ and $\widehat{\sigma}^2 = 6.97$. Figure 10 plots Lotka's data together with the fitted $\widehat{P}_j^*$ values from the truncated Poisson–lognormal model, showing that the estimated probabilities exhibit a substantially linear log–log plot in the range shown.

[As Stewart (1994) pointed out, the Poisson–lognormal model is rejected for Lotka's data by a formal chi-square goodness-of-fit test. On the other hand, Price (1963) noted many years ago that the power law model is also a poor fit over the entire range of data. A more realistic model of both the *Ulysses* data and Lotka's counts would be based on mixtures of Poisson–lognormals, that is, a model that uses mixing distributions for λ that comprise several component lognormal distributions. However, in my conclusion (Section 4), I mention a class of hierarchically structured mixtures

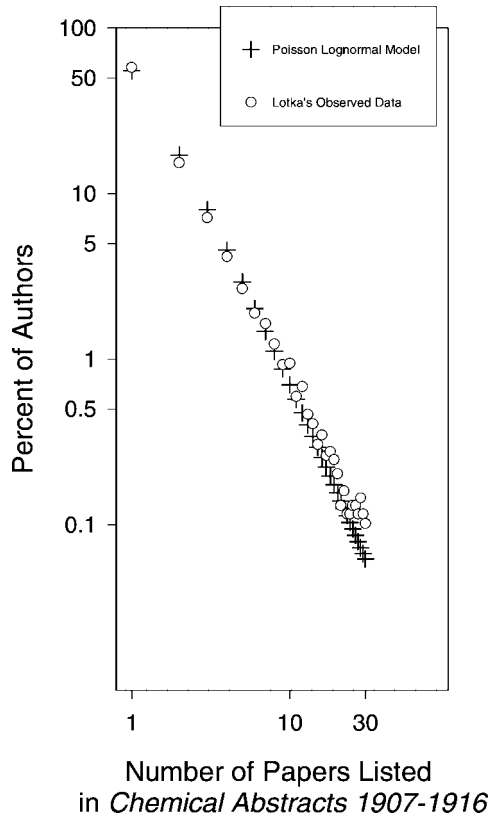

FIG. 10. *A recreation of Lotka's* (1926) *famous log–log plot of the publication counts of chemists. Also shown are the predicted counts from the fitted zero-truncated Poisson–lognormal distribution. The zero-truncated Poisson–lognormal mimics a power law over the range of values that Lokta used in his plot.*

of lognormal distributions that is much more likely to help explain Pareto–Zipf-type data when the slope parameter is highly stable across samples. For example, empirical word frequency distributions have a remarkably constant -1 slope for log–log rank-size plots.]

The *Ulysses* and Lotka data fit to the Poisson–lognormal distribution here exemplify a type of data that occurs quite commonly, but whose highly truncated character has usually not been appreciated. This type of data set is often found in "network science" research, as in the Barabási and Albert (1999) example or the URL links graphed in Figure 5. A salient common feature of these examples is that the sampling scheme is biased because it specifically excludes, or inaccurately assesses, the very large category of observations with $j = 0$. For the Lotka data, this means that, within a given time interval, authors who have higher probabilities of publishing are more likely to be observed, and the lower probability individuals tend not to show up with even a single paper. A similar situation holds for word frequencies. Indeed, a strong indication of a high degree of truncation is the occurrence of a large number of "singletons," that is, the observations that occur only once in the sample. For Lotka's data set above, fully 3991 of the 6891 authors had only one paper to their credit in the *Chemical Abstracts 1907–1916*. Also, from Hanley's (1937) index of the *Ulysses* word frequencies, 16,432 of the 29,899 distinct words in the novel occurred only once. With so many authors and words occurring only one time, it is obvious that there must be *potentially* many more that did not occur at all, but somehow need to be taken into account to get a true picture of the situation.

4. CONCLUSION AND RELATED RESULTS

It is surprising that the significance of truncation and Pareto-mimicking behavior are so little discussed in the literature of power law distributions. The basic facts are easily established with simulations, and useful insights can be obtained with standard asymptotic estimates from extreme value theory. Perhaps part of the explanation that truncation has been so often overlooked is due to the strong natural asymmetry that calls attention to the high end, while obscuring the low end, of many kinds of distributions. What is the height of the *shortest* building or the area of the *smallest* island in the world? What is the *dimmest* star in the sky? How do we estimate the number of chemists who did *not* publish any papers (the zero class) listed in *Chemical Abstracts 1907–1916*? It is in the nature of things that the low end, or very commonly, all but the upper tail, of many kinds of data is hidden because of definitional fuzziness and the difficulties associated with measurement below some threshold. At the same time, it is frequently the high end that is most important or most likely to capture our attention.

As I have attempted to show in Section 3.2, a more promising modeling approach takes into account issues of heterogeneity that have generally been naively ignored. Shoehorning the data into one- or two-parameter models, such as the Pareto or Yule or the lognormal, while simultaneously excluding some inconvenient portion of the distribution, has too long been the norm. Many of the examples of inverse power laws proposed through the years are probably FIPLs best represented by finite mixtures of distributions.

Nevertheless, I want to emphasize my strong belief in the existence of Pareto–Zipf-type distributions that cannot be explained as FIPLs. For instance, Zipf's word frequency law exhibits a rank-size power law exponent consistently close to -1 regardless of the sample used. Such striking stability of a parameter value is not accounted for by Pareto-mimicking FIPLs, which have "exponents" ("slopes" on a log–log scale) that depend on the sample size, as we saw in Section 3.1. [The upper tails of income distributions—the "weak Pareto law"—were found by Pareto to approximately satisfy $N_y = B/y^{\alpha}$ with α close to $3/2$, which implies a log-log rank-size slope of $-2/3$ as in Figure 1. However, the stability of this parameter value in other data sets (Arnold, 1983, Table C) is not nearly as impressive as that for Zipf's word frequency law.] It is important to point out, therefore, that there is a natural path that leads from FIPLs to WIPLs and SIPLs that seems particularly compelling to me as an explanation for how these distributions with "true" power law tails could occur. Instead of working with unstructured, finite mixtures of lognormal distributions (other exponential-type distributions can also be used) as in Section 3.2, there is a class of *hierarchically structured*, infinite mixture distributions that provides the mechanism for generating WIPLs and SIPLs from FIPL building blocks. One well-known example is a model of Pareto's income law given by Montroll and Shlesinger (1982, 1983) that is defined as an infinite mixture of increasingly richer income levels represented by the random variables $Y_1, Y_2, \ldots$, where each component distribution Y_j has the form $k^{j-1}Y$ for some constant $k > 1$ and Y has the lognormal distribution $\mathrm{L}(\mu, \sigma^2)$. Montroll and Shlesinger then defined the mixing proportions for the components as

a geometric distribution, so that Y_j has the mixing proportion qp^{j-1}, $p+q=1$. The parameter $k>1$ is an amplification constant that reflects the hierarchical structure of society in which individuals at higher income levels organize enterprises so that their incomes are "amplified through the efforts of others." Montroll and Shlesinger showed that the overall distribution generated from this infinite mixture is a WIPL with an asymptotic power law tail that has an exponent equal to $-\log p/\log k$.

Quite recently, Reed (2001) proposed a conceptually related model in the context of *geometric Brownian motion*. Reed defined a stochastic process based on an infinite mixture of lognormal distributions with parameter values that increase linearly over time using an exponential mixing distribution for the time variable. From this he then derived a new class of distributions called the *double Pareto*, which can be viewed as an odd form of WIPL with one power law in the upper part of the distribution and another in the lower part. Reed and Hughes (2002) showed how several forms of exponential growth can generate such a stochastic process.

Indeed, the idea of hierarchically structured mixture distributions—closely related to what Mandelbrot (1997, page 226) termed "compensation between two exponentials"—is an old and recurring theme associated with power law behavior. This crops up in the analysis of the monkey-at-the-typewriter generation of pseudo-words (Perline, 1996) and has been proposed in a variety of models of power laws spanning both physical and social sciences.

This leads to a research question of keen interest to me: Which Pareto–Zipf distributions are more plausibly modeled as FIPL unstructured, finite mixtures and which are better explained in terms of hierarchical, infinite mixture models? Finally, let me also pose my favorite Pareto–Zipf challenge: Are there *any* examples in Zipf (1949) that can reasonably be called SIPLs, but are not associated in any way with truncated data? I believe there is at least one, but I will save my argument for another occasion.

APPENDIX: DISCUSSION OF ASYMPTOTIC PARAMETER ESTIMATES

The theoretical slope estimates based on the parameter value b_n defined in (9) may be considerably discrepant from the slope computed from randomly generated sample data until n is sufficiently large. For example, for a truncated sample from the normal distribution (on the log scale) with $\sigma^2=16.9$ and $n=10^6$ as graphed in curve 8 of Figure 6, the theoretical approximation of the slope for the top 10^3 observations plotted against log rank is $-b_n=-0.782$. However, the slope computed from the actual simulated data using the least squares regression of $X_{j:n}=\log Y_{j:n}$ onto $\log j$ for the 10^3 values is -1.074—a large 27% discrepancy. I mention, therefore, that the constants a_n and b_n in (8) and (9) are not unique with respect to the limiting distributions and moments of (4) and (6). Those limit results remain identical if alternative standardizing constants A_n and B_n are substituted, provided that the two limits $\lim_{n\to\infty}(a_n-A_n)/b_n=0$ and $\lim_{n\to\infty} b_n/B_n=1$ hold. This is just a statement of the well-known "convergence to types" theorem (Embrechts, Klüppelberg and Mikosch, 1997) and, consequently, different choices of these constants will prove more useful for the task of slope estimation.

As an illustratation, Hall (1979) showed that for the standard normal distribution, the speed of convergence of the largest order statistic, $X_{1:n}$, to its limiting distribution is fastest by choosing a_n as the solution to $nf(a_n)=a_n$ and then setting $b_n=1/a_n$, where $f(x)=\exp(-x^2/2)/\sqrt{2\pi}$ is the standard normal p.d.f. [These constants, therefore, can be shown to satisfy the convergence to types theorem relative to the a_n and b_n of (9).] For $n=10^6$ and adjusting for $\sigma^2=16.9$, Hall's solution leads to a slope estimate of $b_n=-0.863$, which is a 20% discrepancy—already an improvement from above. Still further refinements based on other considerations can yield even better slope estimates.

A reviewer of an earlier version of this paper raised questions about the accuracy of the random values used in the simulations. The normal random deviates were produced by SAS©V8.2 programs employing a random generator based on the widely used procedure of generating uniform random deviates using a multiplicative congruential generator modulus $2^{31}-1$ (Fishman and Moore, 1982) followed by a Box–Muller (1958) transformation. This procedure is exceedingly unlikely to be producing poor quality simulations on samples the size of those used here, but there are several ways to check this. One approach is to use some classical bounds on the moments of order statistics in terms of distribution quantiles (David, 1970). Let $\Phi^{-1}(x)$ be the inverse of the standard normal c.d.f. It can be shown that the bounds $\Phi^{-1}((n-j)/n)\le E(X_{j:n})<\Phi^{-1}((n-j+0.5)/n)$ hold for all $1\le j\le n$. First, with $n=10^6$, I took the midpoints of these bounds (adjusted for $\sigma^2=16.9$ and $\mu=-5.79$) as estimates of the $E(X_{j:n})$ and then regressed them onto $\log j$, for $1\le j\le 10^3$. The slope

and intercept values obtained with this regression were within 0.1% of that obtained from the regression using the simulated sample. Also, based on 50 different simulated samples of 10^6 observations, the average values for the top 10^3 sample order statistics differed by only tiny amounts from the midpoint estimates from the classical bounds.

ACKNOWLEDGMENTS

I want to thank two previous Executive Editors of *Statistical Science*, Prof. George Casella and Prof. Leon Gleser, and their reviewers for numerous specific suggestions that greatly improved this article. Also, thanks to Chris Monroe for his helpful recommendations and to Abe Bookstein for introducing me to this topic (many years ago!).

REFERENCES

AITCHISON, J. and BROWN, J. A. C. (1957). *The Lognormal Distribution*. Cambridge Univ. Press.

ALBERT, R., JEONG, H. and BARABÁSI, A.-L. (1999). Diameter of the World-Wide Web. *Nature* **401** 130.

AMARAL, L. A. N., SCALA, A., BARTHELEMY, M. and STANLEY, H. E. (2000). Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.* **97** 11,149–11,152.

AMERICAN IRON AND STEEL INSTITUTE (1957). *Directory of Iron and Steel Works of the United States and Canada*, 28th ed. American Iron and Steel Institute, New York.

ARNOLD, B. C. (1983). *Pareto Distributions*. International Cooperative Publishing House, Burtonsville, MD.

ASMUSSEN, S., KLÜPPELBERG, C. and SIGMAN, K. (1999). Sampling at subexponential times, with queueing applications. *Stochastic Process. Appl.* **79** 265–286.

AUERBACH, F. (1913). Das Gesetz der Bevölkerungskonzentration. *Petermanns Geographische Mitteilungen* **59** 74–76.

BAK, P. (1996). *How Nature Works*. Copernicus, New York.

BARABÁSI, A.-L. (2002). *Linked*: *The New Science of Networks*. Perseus, Cambridge, MA.

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512.

BARABÁSI, A.-L. and BONABEAU, E. (2003). Scale-free networks. *Scientific American* **288** 60–69.

BERG, L. (1958). Asymptotische Darstellungen für Integrale und Reihen mit Anwendungen. *Math. Nachr.* **17** 101–135.

BIANCONI, G. and BARABÁSI, A.-L. (2001). Competition and multiscaling in evolving networks. *Europhys. Lett.* **54** 436–442.

BOOKSTEIN, A. (1997). Informetric distributions. III. Ambiguity and randomness. *J. American Society for Information Science* **48** 2–10.

BOWLEY, A. L. (1899). The statistics of wages in the United Kingdom during the last hundred years. Part IV. Agricultural wages. *J. Roy. Statist. Soc.* **62** 555–570.

BOX, G. E. P. and MULLER, M. E. (1958). A note on the generation of random normal deviates. *Ann. Math. Statist.* **29** 610–611.

BULMER, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* **30** 101–110.

DAVID, H. A. (1970). *Order Statistics*. Wiley, New York.

DOWNEY, A. B. (2003). Lognormal and Pareto distributions in the internet. Available at http://allendowney.com/research/longtail.

EDWARDS, A. M. (1943). *Sixteenth Census of the United States, 1940. Population. Comparative Occupation Statistics for the United States, 1870 to 1940*. U.S. Government Printing Office, Washington.

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events*. Springer, Berlin.

FISHMAN, G. S. and MOORE, L. R. (1982). A statistical evaluation of multiplicative congruential random number generators with modulus $2^{31} - 1$. *J. Amer. Statist. Assoc.* **77** 129–136.

GIBRAT, R. (1931). *Les Inégalités Économiques*. Libraire de Recueil Sirey, Paris.

GONG, W., LIU, Y., MISRA, V. and TOWSLEY, D. (2001). On the tails of Web file size distributions. In *Proc. 39th Annual Allerton Conference on Communication, Control and Computing*. Univ. Illinois Press, Champaign. Available at http://www1.cs.columbia.edu/~misra/pubs/allerton.pdf.

GRAHAM, R. L., KNUTH, D. E. and PATASHNIK, O. (1994). *Concrete Mathematics*: *A Foundation for Computer Science*, 2nd ed. Addison–Wesley, Reading, MA.

GRANDELL, J. (1997). *Mixed Poisson Processes*. Chapman and Hall, New York.

HALL, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16** 433–439.

HANLEY, M. L. (1937). *Word Index to James Joyce's Ulysses*. Univ. Wisconsin Press, Madison.

IJIRI, Y. and SIMON, H. A. (1977). *Skew Distributions and the Sizes of Business Firms*. North-Holland, Amsterdam.

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1994). *Distributions in Statistics*: *Continuous Univariate Distributions* **1**, 2nd ed. Wiley, New York.

KENDALL, M. G. (1961). Natural law in the social sciences. *J. Roy. Statist. Soc. Ser. A* **124** 1–16.

KLEIN, L. R. (1962). *An Introduction to Econometrics*. Prentice–Hall, Englewood Cliffs, NJ.

KORČÁK, J. (1938). Deux types fondamentaux de distribution statistique. *Bull. Inst. Internat. Statist.* **30**(3) 295–298.

KRUGMAN, P. (1996). *The Self-Organizing Economy*. Blackwell, Cambridge, MA.

LEBERGOTT, S. (1959). The shape of the income distribution. *American Economic Review* **49** 328–347.

LOTKA, A. J. (1926). The frequency distribution of scientific productivity. *J. Washington Academy of Sciences* **16** 317–323.

MACAULEY, F. (1922). Pareto's law and the general problem of mathematically describing the frequency distribution of income. In *Income of the United States. Its Amount and Distribution 1909–1919* **2** Chap. 23. National Bureau of Economic Research, New York.

MANDELBROT, B. (1960). The Pareto–Lévy law and the distribution of income. *Internat. Econom. Rev.* **1** 79–106.

MANDELBROT, B. (1982). *The Fractal Geometry of Nature*. W. H. Freeman, San Francisco.

MANDELBROT, B. (1997). *Fractals and Scaling in Finance*: *Discontinuity, Concentration, Risk*. Springer, New York.

MCLACHLAN, G. J. and KRISHNAN, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.

MITZENMACHER, M. (2001). A brief history of generative models for power law and lognormal distributions. In *Proc. 39th Annual Allerton Conference on Communication, Control and Computing* 182–191. Univ. Illinois Press, Champaign.

MONTROLL, E. and SHLESINGER, M. F. (1982). On $1/f$ noise and other distributions with long tails. *Proc. Natl. Acad. Sci. U.S.A.* **79** 3380–3383.

MONTROLL, E. and SHLESINGER, M. F. (1983). Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: A tale of tails. *J. Statist. Phys.* **32** 209–230.

NATIONAL RESOURCES COMMITTEE (1938). *Consumer Incomes in the United States*: *Their Distribution in 1935–36*. U.S. Government Printing Office, Washington, DC.

PADDOCK, R. H. and RODGERS, R. P. (1939). Preliminary results of road-use studies. *Public Roads* **20** 45–63.

PARETO, V. (1895). La legge della demanda. *Giornale degli Economisti* 45–63.

PARETO, V. (1897). *Cours d'Économie Politique* **2**. F. Rouge, Lausanne.

PARR, J. B. and SUZUKI, K. (1973). Settlement populations and the lognormal distribution. *Urban Studies* **10** 335–352.

PERLINE, R. (1982). An extreme value model of weakly harmonic (Pareto–Zipf type) laws. Ph.D. dissertation, Univ. Chicago.

PERLINE, R. (1996). Zipf's law, the central limit theorem, and the random division of the unit interval. *Phys. Rev. E* **54** 220–223.

PERLINE, R. (1998). Mixed Poisson distributions tail equivalent to their mixing distributions. *Statist. Probab. Lett.* **38** 229–233.

POLFELDT, T. (1970). Asymptotic results in non-regular estimation. *Skand. Aktuarietidskr.* **1970 suppl.** 1–78.

PRICE, D. J. DE S. (1963). *Little Science, Big Science*. Columbia Univ. Press, New York.

REED, W. J. (2001). The Pareto, Zipf and other power laws. *Econom. Lett.* **74** 15–19.

REED, W. J. and HUGHES, B. D. (2002). From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature. *Phys. Rev. E* **66** 067103.

SICHEL, H. S. (1975). On a distribution law for word frequencies. *J. Amer. Statist. Assoc.* **70** 542–547.

SIMON, H. (1955). On a class of skew distribution functions. *Biometrika* **52** 425–440. Also in Ijiri and Simon (1977).

SIMON, H. A. and BONINI, C. P. (1958). The size distribution of business firms. *American Economic Review* **48** 607–617. Also in Ijiri and Simon (1977).

STAMP, J. (1914). A new illustration of Pareto's law. *J. Roy. Statist. Soc.* **77** 200–204.

STEWART, J. (1994). The Poisson–lognormal model for bibliometric/scientometric distributions. *Information Processing and Management* **30** 239–251.

THATCHER, A. R. (1976). The new earnings survey and the distribution of earnings. In *The Personal Income Distribution* (A. B. Atkinson, ed.) 227–268. Westview Press, Boulder, CO.

WATTS, D. J. (2003). *Six Degrees*: *The Science of a Connected Age*. Norton, New York.

ZIPF, G. (1947). The frequency and diversity of business establishments and personal occupations: A study of social stereotypes and cultural roles. *J. Psychology* **24** 139–148.

ZIPF, G. K (1949). *Human Behavior and the Principle of Least Effort*. Addison–Wesley, Cambridge, MA.