Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

MORE "NORMAL" THAN NORMAL: SCALING DISTRIBUTIONS AND COMPLEX SYSTEMS

Walter Willinger

AT&T Labs-Research
180 Park Ave., Room B207
Florham Park, NJ 07932, U.S.A.

David Alderson, John C. Doyle, Lun Li

Department of Control and Dynamical Systems
California Institute of Technology
Pasadena, CA 91125, U.S.A.

ABSTRACT

One feature of many naturally occurring or engineered complex systems is tremendous variability in event sizes. To account for it, the behavior of these systems is often described using power law relationships or scaling distributions, which tend to be viewed as "exotic" because of their unusual properties (e.g., infinite moments). An alternate view is based on mathematical, statistical, and data-analytic arguments and suggests that scaling distributions should be viewed as "more normal than Normal". In support of this latter view that has been advocated by Mandelbrot for the last 40 years, we review in this paper some relevant results from probability theory and illustrate a powerful statistical approach for deciding whether the variability associated with observed event sizes is consistent with an underlying Gaussian-type (finite variance) or scaling-type (infinite variance) distribution. We contrast this approach with traditional model fitting techniques and discuss its implications for future modeling of complex systems.

1 INTRODUCTION

A common research theme in the study of complex systems is the pursuit of universal properties that transcend specific system details. It is in exactly what those properties are, and the theories to explain and exploit them, where sharp differences arise. One aspect of many complex systems that has received considerable attention is a tendency toward tremendous variability in event sizes, such that they can be reasonably represented by a so-called "power law" relationship. That is, the cumulative probability $P(X > l)$ of observing events greater than a given size l is given by $P(X > l) \approx l^{-\alpha}$ and manifests itself as a straight line of slope $-\alpha$ in a $\log(P)$ vs. $\log(l)$ plot (for large values of l, and for $\alpha > 0$). For example, consider the relative sizes of the largest disaster events during the 20^{th} Century (Figure 1). Simple inspection of the data shows a striking relationship between the size and frequency of large events, namely that they are reasonably approximated by a power law having $\alpha = 1$. Power law relationships have been observed within many naturally occurring and man made systems, including species within plant genera (Yule 1925); mutants in old bacterial populations (Luria and Delbrück 1943); a number of applications in the social sciences (Simon 1955), including economics (income distributions, city populations) and linguistics (word frequencies) (Mandelbrot 1997); forest fires (Malamud, Morein, and Turcotte 1998); Internet traffic (flow sizes, file sizes, Web documents) (Crovella and Bestavros 1997) and Internet topology (node degrees of physical or virtual connectivity graphs) (Faloutsos, Faloutsos, and Faloutsos 1999); and metabolic networks (Barabasi and Oltavi 2004). That such a diversity of systems exhibits similar scaling features has prompted many researchers to ask whether or not there are universal drivers of these phenomena.

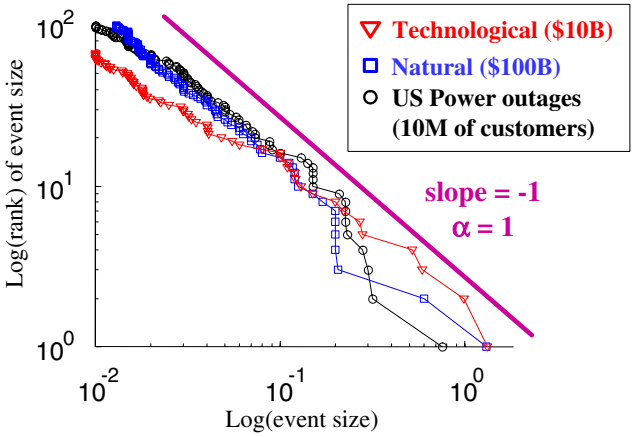

Figure 1: Log-log Plot of Event Size Versus Event Rank (100 Largest Disasters of the 20^{th} Century)

Given the discovery of such "emergent" properties of complex systems and the ability to describe them with power law-type relationships or scaling distributions, a fundamental issue underlying the attempts by complex systems researchers to understand and explain these highly variable event sizes has been the extent to which such high variability should be viewed as "exotic" (in the sense of

surprising) or fully expected. For example, the traditional statistical physics perspective views scaling distributions as a signature of an internal self-sustaining critical state, where the details associated with the initiation of events are a statistically inconsequential factor in determining the events' sizes. As a result, scaling distributions and their quantitative features are generally taken at face value and considered as simultaneously ubiquitous, arcane, and exotic (Bak 1996, Buchanan 2001, Barabasi 2002), which in turn argues for the construction of specialized models that produce or "explain" observed scaling behavior. Examples include various models based on self-organized criticality (SOC), edge-of-chaos (EOC), and more recently, scale-free networks (SFN), and they attempt to describe the observed phenomena (including high variability) as adaptive, self-organizing, far-from-equilibrium, or nonlinear. Developed and refined over the last several decades, these models are "exotic" in the sense that they rely on mechanisms that are generic or universal and independent of system-specific details. They assume that interactions are essentially random, but have some macroscopic statistic tuned to a special point, such as a bifurcation point (EOC), a critical density (SOC), or a power-law degree distribution (SFN). In addition, they are consistent in that they treat the appearance of power law (scaling) relationships as evidence of some critical phenomenon, which is in turn indicative of universal features that contribute to the large-scale properties of all complex systems. Applications of this approach to many important complex systems have been documented in the literature.

A completely different approach to dealing with high variability is based on *Highly Optimized Tolerance or Tradeoff (HOT)*, a recently introduced conceptual framework for capturing the highly organized, optimized, and "robust yet fragile" structure of complex systems such as the Internet (Carlson and Doyle 1999, Carlson and Doyle 2002). HOT is capable of accounting for many essential features of these systems with abstract models that are surprisingly simple yet contrast sharply with their "exotic" statistical physics-based counterparts (Carlson and Doyle 1999, Li et al. 2004). The idea of HOT is to put design or evolution in explicitly, yet use when possible the simple models of statistical physics to illustrate the essential tradeoffs that arise. To this end, the HOT-based approach argues that high variability in complex systems should come as no surprise but arises naturally as result of a rational design process, reflecting an inherent need for tradeoffs between resources and constraints. While sampling and data analytic methods can make inferring scaling distributions inevitable either correctly in the presence of high variability data or incorrectly via strong biases or inadequate statistics, such that assuming scaling distributions often requires a leap of faith, the qualitative assessment of the presence or absence of high variability in a given data set is usually less error-prone. By focusing on high variability and not on scaling distributions or power law relationships per se, the overriding concern of the HOT-based approach is to understand the main mechanisms that cause complex systems to exhibit high variability and not to produce yet another "exotic" explanation for an observed power law relationship.

This HOT perspective of high variability is fully consistent with a view advocated by Mandelbrot, who has provided for the last 40 years mathematical, statistical, and data-analytic arguments that demonstrate that highly variable event sizes are in a sense just as "normal"—or even more "normal"—than Gaussian-type event sizes. The main purpose of this paper is to pay tribute to Mandelbrot's ground-breaking work in this area as summarized in (Mandelbrot 1997) and review his key probabilistic and statistical arguments that form the basis for a rigorous treatment of high variability in observed event sizes. When reviewing in Section 2 the mathematical results and illustrating in Section 3 the statistical techniques, we follow closely (Willinger et al. 2004), where the focus is mainly on the observed high variability in Internet-related measurements. In Section 4 we discuss with a few examples how the HOT-based approach to high variability yields findings that are fully consistent with the real system (via supporting measurements), but contrast sharply with those that are based on the "exotic" statistical physics-based models.

2 MODELING HIGH VARIABILITY

We introduce in this section the class of subexponential distributions which provides a rigorous and convenient mathematical framework for dealing with high variability phenomena. To further distinguish between finite versus infinite variance distributions, we consider a subclass of the subexponential distributions, called heavy-tailed or scaling distributions, and elaborate on some key mathematical properties of the latter. For a more comprehensive treatment of these topics, we refer to a survey on subexponential distributions by Goldie and Klüppelberg (1998) and to early works by Mandelbrot on scaling distributions reproduced in Mandelbrot (1997).

2.1 Heavy-Tailed or Scaling Distributions

Focusing throughout this paper on non-negative random variables X, let $F(x) = P[X \leq x]$, $x \geq 0$, denote the *cumulative distribution function (CDF) of* X and $\bar{F}(x) = 1 - F(x)$ the *complementary CDF (CCDF)*. A typical feature of commonly-used distribution functions is that their (right) tails decrease exponentially fast, implying that all moments, including exponential moments, exist and are finite. In practice, this property ensures that X exhibits low variability and thus concentrates tightly around its mean. To describe in a mathematically convenient way high variability phenomena, we introduce next the class of subexponential distribution

functions. Following Goldie and Klüppelberg (1998), we call F (or X) *subexponential* if

$$\lim_{x\to\infty} \frac{P[X+Y>x]}{P[X>x]} = 2,$$

where Y is an independent copy of X. This definition can be shown to be equivalent to

$$\lim_{x\to\infty} \frac{P[X_1+\cdots+X_n>x]}{P[\max(X_1,\ldots,X_n)>x]} = 1 \text{ for some (all) } n \geq 2,$$

where $X_1, X_2, \ldots$ are IID non-negative random variables with distribution function F. This shows that in contrast to low variability distributions, the sum of n IID subexponential random variables is likely to be large if and only if their maximum is likely to be large, and accounts for the non-negligible probability that there will be extremely large values in a subexponential sample. This definition also implies that for subexponential distributions, we have

$$\bar{F}(x)/e^{-\epsilon x} \to \infty \text{ as } x \to \infty \text{ for all } \epsilon > 0;$$

that is, the (right) tail of a subexponential distribution decays more slowly than any exponential, implying that all exponential moments of a subexponential are infinite. Well-known examples of subexponential distributions include the Lognormal, Weibull, Pareto of the first or second kind, and stable laws, while the Gaussian, exponential, and Gamma are examples that are not in the class of subexponentials. It is sometimes convenient to consider the slightly more general class of *long-tailed distribution functions* (Goldie and Klüppelberg 1998), but for the purpose of this paper, this generalization is not needed.

To distinguish between subexponential distributions whose regular moments can also be infinite (e.g., between Lognormal and Pareto), we next consider the subclass of subexponentials consisting of the heavy-tailed or scaling distributions. To this end, a subexponential distribution function $F(x)$ or random variable X is called *heavy-tailed* or *scaling* if for some $0 < \alpha < 2$

$$P[X > x] \approx cx^{-\alpha} \quad \text{as} \quad x \to \infty \tag{1}$$

where $0 < c < \infty$. The parameter α is called the *tail index*. For $1 \leq \alpha < 2$, F has infinite variance but finite mean; for $0 < \alpha < 1$, F has not only infinite variance, but also infinite mean. In general, all moments of F of order $\beta \geq \alpha$ are infinite. Heavy-tailed distributions are called scaling distributions because the sole response to conditioning is a change in scale; that is, if X is heavy-tailed with index α and $x > w$, the conditional distribution of X given that $X > w$ satisfies

$$P[X > x|X > w] = \frac{P[X>x]}{P[X>w]} \approx c_1 x^{-\alpha},$$

which—at least for large values of x—is identical to the (unconditional) distribution $P[X > x]$, except for a change in scale. In contrast, the non-heavy-tailed *exponential distribution* gives

$$P(X > x|X > w) = e^{-\lambda(x-w)},$$

that is, the conditional distribution is also identical to the (unconditional) distribution, except for a change of location rather than scale. A more general definition involving regularly varying tails is possible (Goldie and Klüppelberg 1998), but such a generalization makes applying and inferring scaling behavior cumbersome.

Scaling distributions are also called *power law distributions*, and we will use below the notions of scaling, heavy-tailed, and power law distributions interchangeably and only insist that the right tail of the distribution satisfies property (1). One of the most publicized features of scaling distributions which follows trivially from (1) is that their CCDF, when plotted on a log-log scale, appears as a straight line, at least asymptotically. The CCDF plots for a number of well-known distributions are shown in Figure 2.

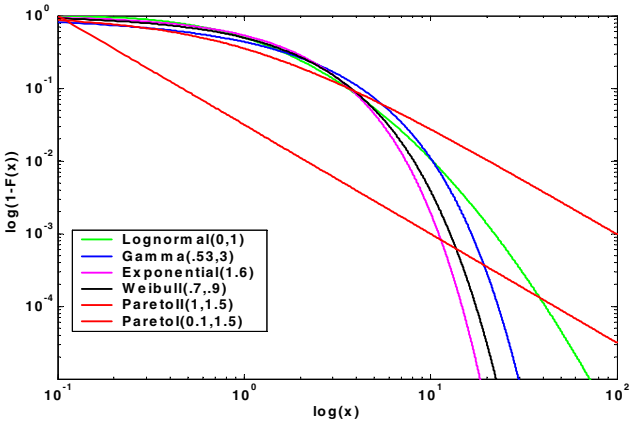

Figure 2: CCDF Plots of Some Known Distributions

An important but more obscure feature of scaling distributions that distinguishes them from their commonly-considered non-heavy-tailed counterparts concerns their *mean residual lifetime* defined as $E(X - x|X > x)$. While it is well-known that the mean residual lifetime of an exponential distribution with parameter λ is constant, i.e.,

$$E(X - x|X > x) = \frac{1}{\lambda},$$

the majority of non-heavy-tailed distributions have decreasing mean residual lifetime. In stark contrast, the mean

residual lifetimes of scaling distributions are linearly increasing, i.e.,

$$E(X - x|X > x) \approx cx.$$

Finally, some simple constructions that yield scaling distributions include the following.

- For U uniform in $[0, 1]$, set $X = 1/U$, then X is heavy-tailed with $\alpha = 1$.
- For E (standard) exponential, set $X = \exp(E)$, then X is heavy-tailed with $\alpha = 1$.
- The mixture of exponential distributions with parameter $1/\delta$ having a (centered) Gamma(a, b) distribution is a Pareto distribution with $\alpha = a$.
- The distribution of the time between consecutive visits of a symmetric random walk to zero is heavy-tailed with $\alpha = 1/2$.

2.2 Invariance Properties

Scaling distributions enjoy a number of invariance properties that (sometimes uniquely) characterize them. We follow here the presentations in Mandelbrot (1997), show that scaling distributions are essentially invariant under transformations such as aggregation, mixture, maximization, and marginalization, and discuss some practical implications of this invariance property.

2.2.1 Aggregation

The classical central limit theorem (CLT) is often cited as the reason for the ubiquity with which Gaussian (normal) distributions occur in nature. While more general versions of the CLT are available and can be found, for example, in Feller (1971), in its standard form (e.g., Whitt 2002), the classical CLT states:

Theorem 1 *Suppose that $(X_n : n \geq 1)$ is a sequence of IID random variables with distribution function F, where F has finite mean m and finite variance σ^2). Let $S_n = X_1 + \cdots X_n, n \geq 1$ denote the n^{th} partial sum. Then, as $n \to \infty$,*

$$n^{-1/2}(S_n - mn) \Rightarrow \sigma N(0, 1),$$

where $N(0, 1)$ is the standard Gaussian (normal) distribution having mean 0 and variance 1.

For a somewhat less well-known version of the CLT, we recall that a random variable U is said to have a *stable law (with index $0 < \alpha \leq 2$)* if for any $n \geq 2$, there is a real number d_n such that

$$U_1 + U_2 + \cdots + U_n = n^{1/\alpha}U + d_n,$$

where $U_1, U_2, \ldots, U_n$ are independent copies of U. Following Samorodnitsky and Taqqu (1994), the stable laws on the real line can be represented as a four-parameter family $S_\alpha(\sigma, \beta, \mu)$, with the *index* α, $0 < \alpha \leq 2$; the *scale parameter* $\sigma > 0$; the *skewness parameter* β, $-1 \leq \beta \leq 1$; and the *location (shift) parameter* μ, $-\infty < \mu < \infty$. When $1 < \alpha < 2$, the shift parameter is the mean, but for $\alpha \leq 1$, the mean is infinite. There is an abrupt change in tail behavior of stable laws at the boundary $\alpha = 2$. While for $\alpha < 2$, all stable laws are heavy-tailed in the sense that they satisfy condition (1), the case $\alpha = 2$ is special and represents a familiar, not heavy-tailed distribution—the Gaussian (normal) distribution; i.e., $S_2(\sigma, 0, \mu) = N(\mu, 2\sigma^2)$. The following is a simple version of the stable-law CLT.

Theorem 2 *Suppose that $(X_n : n \geq 1)$ is a sequence of non-negative, IID random variables with scaling distribution function F with $1 < \alpha < 2$ (implying finite mean m but infinite variance). Let $S_n = X_1 + \cdots X_n, n \geq 1$ denote the n^{th} partial sum. Then, as $n \to \infty$,*

$$n^{-1/\alpha}(S_n - mn) \Rightarrow S_\alpha(1, \beta, 0).$$

Again, more general versions of this non-classical CLT are available and can be found, for example, in Feller (1971) or Whitt (2002). For a detailed treatment of stable distributions, we refer to Samorodnitsky and Taqqu (1994). Together, these results show that the Gaussian and scaling distributions are both invariant under aggregation. More precisely, the classical and non-classical CLTs state that the stable distributions are the only fixed points of the renormalization group transformation (i.e., aggregation) and that Gaussian distributions are, in fact, a very special case (i.e., $\alpha = 2$).

2.2.2 Maximizing Choices

Consider the case of n independent random variables denoted $X_1, X_2, \ldots, X_n$ and assume that their distribution functions are scaling distributions with the same tail index parameter α, but possibly with different scale coefficients $c_i > 0$; that is,

$$P(X_i > x) \approx c_i x^{-\alpha} \text{ for } (1 \leq i \leq n).$$

For $1 \leq k \leq n$, define the random variables M_k to be the k-th successive maxima given by

$$M_k = \max(X_1, X_2, \ldots, X_k).$$

Using that $P(M_k \leq x) = \prod_{1 \leq i \leq k} P(X_i \leq x)$, it is easy to show that for large x,

$$P[M_k > x] \approx c_{M_k} x^{-\alpha},$$

where $c_{M_k} = \sum_{1 \le i \le k} c_i$. Thus, the k-th successive maxima of scaling distributions are also scaling, with the same tail index α, but different scale coefficients than the individual X_i's.

As for the converse (i.e., M_k is scaling only if the X_i's are scaling), for the invariance-up-to-scale to hold formally, the distributions of the X_i's need not follow the scaling distribution exactly. In fact, a result from extreme value theory (see for example Resnick 1987) identifies the invariant distributions as the Frechet distributions and characterizes the distributions of the X_i's that are in the domain of attraction of the Frechet distribution. The Frechet distribution is defined by $P[M > x] = 1 - \exp(-x^{-\alpha}),\ x > 0$ and is clearly scaling for large x. As a consequence, the individual X_i' s must be so close to being scaling distributions as to be scaling for all practical purposes. In this sense, scaling distributions are the only distributions that are invariant under the transformation of maximization. In particular, Gaussian distributions lack this invariance property.

2.2.3 Weighted Mixtures

As before, assume that $X_1, X_2, \ldots, X_n$ are n independent random variables with scaling distribution functions F_i, all with the same tail index parameter α, but possibly with different scale coefficients $c_i > 0$. Consider the *weighted mixture* W_n of the X_i's, and denote by p_i the probability that $W_n = X_i$. It is easy to show that

$$P[W_n > x] = \sum p_i P[X_i > x] \approx c_{W_n} x^{-\alpha},$$

where $c_{W_n} = \sum p_i c_i$ is the *weighted average* of the separate scale coefficients c_i. Thus, the distribution of the weighted mixture of scaling distributions is also scaling, with the same tail index α, but a different scale coefficient than the individual X_n's.

Mathematically, the converse (i.e., W_n is scaling only if the X_i's are scaling) holds only to a first approximation. In fact, in the limit as $n \to \infty$, the invariant "distributions" for W are of the form $P[W > x] = cx^{-\alpha}, x \ge 0$, which are improper distribution functions because they yield an infinite total probability. However, for all practical purposes, the X_i's are typically restricted by some relation of the form $0 < a \le x$ which results in perfectly well-defined (conditional) distribution functions of the scaling type. With these qualifications, scaling distributions are the only distributions that are invariant under the transformation of weighted mixture.

2.2.4 Marginalization

Recall that stable distributions are trivially scaling. For the sake of completeness, we note that the stable distributions, like the Gaussian, have natural extensions to the multivariate case. Indeed, the multivariate stable distributions can be characterized as being those for which every linear combination of the coordinates has a (scalar) stable distribution. We call this transformation marginalization and refer to Samorodnitsky and Taqqu (1994) for an in-depth treatment of stable distributions and their properties.

2.3 Scaling Distributions: More Normal than Normal

Aggregation, mixture, maximization, and marginalization are transformations that occur frequently in natural and engineered systems and are inherently part of many measured observations that are collected about them. For example, aggregate incomes are more widely collected and reported than each type of income separately; the flow or file/document sizes transmitted across the Internet and observed at a particular link within the network are naturally a mixture of distributions of the different file/document sizes residing on the various file/Web servers; for historical data such as natural or technological disasters (i.e., droughts, floods, earthquakes, hurricanes, blackouts, nuclear accidents), the fully recorded and commonly available observations reflect a maximizing choice and correspond to the exceptional (i.e., largest, or most catastrophic) events; and the marginalization transformation is relevant for dealing with a variety of multidimensional economic quantities. In turn, these invariance properties suggest that the presence of scaling distributions in data obtained from complex natural or engineered systems should be considered the norm rather than the exception and should not require "special" explanations.

However, there is an implicit tradeoff between Gaussians being the norm for low variability data and scaling distributions being the norm for high variability data. In the former case, the (traditional) CLT imposes only minimal conditions on the distribution of the individual constituent (i.e., finite variance), but as a result, invariance properties can only be obtained for aggregation and marginalization. In contrast, for high variability data, the relevant CLT requires the (right) tail of the distribution of the individual constituents to decay at a certain rate, and as a result of this more restrictive assumption, the individual constituents are not only invariant under aggregation and marginalization, but also under maximization and weighted mixtures. The pragmatic approach to dealing with high variability data advocated in this paper then consists of viewing Gaussians as the natural null hypothesis for low variability data, where variance estimates exist, are finite, and converge robustly to their theoretical value as the number of observations increases. Similarly, it views scaling distributions as the natural and parsimonious null hypothesis for high variability data, where variance estimates tend to be ill-behaved and converge either very slowly or fail to converge all together as the size of the data set increases. In addition, it fully

exploits the different invariance properties exhibited by low versus high variability data.

3 INFERRING HIGH VARIABILITY

Increasingly, conventional model fitting is experiencing the dilemma that when faced with large data sets or with data having non-traditional characteristics (e.g., high variability), standard goodness-of-fit tests to select among alternate models are in general inadequate and fail to choose the "best" model. In this section, we suggest an alternative approach to distinguishing between competing models that uses conventional model fitting not as an end in itself, but applies it iteratively to increasingly larger subsets of the data set at hand and checks for self-consistency among the resulting models.

3.1 Conventional Model Fitting: An End in Itself

In simplified terms, conventional model fitting proceeds in four steps. It starts by considering a given data set "as is", that is, all the available observations are taken at once and at face value. This is followed by selecting parameterized models or model classes that are deemed appropriate for the data at hand. In a third step, the full data set is used to estimate the necessary model parameters, and the last step consists of selecting the model that fits the data "best" according to some goodness-of-fit criterion.

Figure 3 illustrates steps 1–3 of this approach with a single data set and two different models. The data set is from Willinger and Paxson (1998) and consists of some 240,000 HTTP connection sizes (in bytes) collected at LBL's WAN (in- and outbound) for a 24-hour period in June of 1996. The two models are the 2-parameter Lognormal(μ, σ) distribution and the 2-parameter Pareto(β, α) distribution (e.g., see Johnson, Kotz, and Balakrishnan 1994). Fitting of the Lognormal was done by conventional moment-matching techniques, and fitting of the Pareto involved the "naive" tail index estimate (i.e., slope of fitted straight line through the tail of the CCDF, where the CCDF is plotted on a log-log scale).

While more sophisticated parameter estimation techniques could be used, the outcome of this standard model fitting exercise is highly predictable. Reasonable models will provide a reasonable fit, with more highly parameterized models typically yielding a better fit than more parsimonious ones. Moreover, because of the voluminous data set and some "unusual" features in the data (e.g., extreme values that are genuine and cannot be dismissed as outliers; possible dependencies), commonly-used goodness-of-fit measures to choose among comparable candidate models generally fail to identify the "best" model. For example, models that are excellent approximations tend to be rejected in large

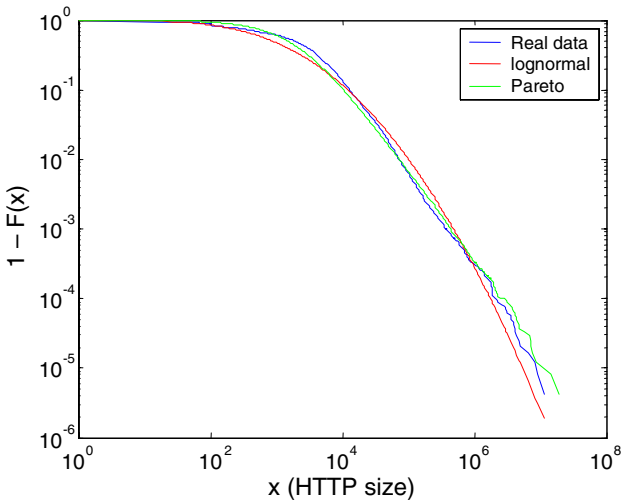

Figure 3: Model Fitting by Example: CCDF Plots of Fitted Lognormal, Fitted Pareto, and Original Data (HTTP Data Set)

samples, no matter how closely they seem to fit the data, resulting in similar discussions as, for example, in Downey (2001) about whether Lognormal or Pareto is a better model for a range of Internet traffic-related quantities. In view of G. P. E. Box's comment that "all models are wrong, but some models are useful", conventional model fitting applied to large data sets offers increasingly less guidance as to which models are indeed useful and has left, for example, Internet traffic modeling in a rather precarious state, where the same set of measurements are fitted with very different, but apparently equally "good" models, which in turn can give rise to completely opposite explanations and theories for one and the same observed phenomenon (see discussion in Section 4).

3.2 Beyond Conventional Model Fitting: Borrowing Strength

To find a way out of the above dilemma, we first note that taking a data set "as is" in step one of the described model fitting process is somewhat arbitrary. For example, in the case of the HTTP data set, we may just as well have ended up with only 1 hour, or half a day, or maybe even with two days worth of measurements, depending on the circumstances under which this measurement effort took place. Thus instead of viewing a given data set as "static", we propose taking a more "dynamic" view of the data at hand and apply Tukey's principle of *"borrowing strength from large data sets"* (Tukey 1986) in practice. To this end, let D denote the original data set of size N, start with an initial subset $D_0 \subset D$ of size N_0, and consider successively larger subsets $D_1, D_2, \ldots, D_n$ satisfying $D_1 \subset D_2 \subset \cdots \subset D_n$ and of length $N_0 < N_1 < \cdots < N_n$, with $N_n \approx N$.

The main motivation for taking this dynamic view of the data set D is that it allows for a careful exploration of the consistency of an assumed model (e.g., a Lognormal or Pareto distribution) as the number of observations increases. In particular, making the commonly-used assumption that one and the same (unknown) underlying process generated the data at hand in the first place, increasing the number of observations as we examine the sets D_0 through D_n should have only the following two main effects. First, the parameter estimates of the fitted model M_i should stabilize, and second, their accuracy expressed in terms of the widths of their corresponding 95% confidence intervals CI_i should improve in such a way that ultimately (i.e., as i tends to n), the confidence intervals CI_i should become roughly nested, with $CI_i \supseteq CI_{i+1}$.

To examine whether the fitted models M_i are indeed self-consistent, we combine Tukey's borrowing strength principle with Mandelbrot's *"sequential (moment) estimate plots"* (Mandelbrot 1997). The latter is simply a method that plots the "running (moment) estimates"; that is, the value of a model parameter estimate or a moment estimate of the data is plotted as a function of the number of observations used in the estimation of the parameter/moment. For example, Figure 4 shows the sequential standard deviation plot for the HTTP data set. Clearly, while for any fixed n, the sample standard deviation $S(n)$ always exists and is finite, $S(n)$ does not seem to converge as n increases, suggesting that it is conceivable to assume that the second moment does not exist, i.e, the data set is a sample from an underlying infinite variance distribution. To compare and become more familiar with interpreting sequential moment plots, Figure 4 also shows sequential standard deviation plots for (i) a random permutation of the HTTP data set, (ii) same-sized samples generated from the fitted Lognormal in Section 3.1, (iii) same-sized samples generated from the fitted Pareto in Section 3.1, and (iv) same-sized samples generated from fitted exponential distributions. While the sequential standard deviation plot for the randomized sample confirms the non-existence of a second moment for the data, plots (ii)–(iv) clearly depict the differences between a subexponential distribution with finite variance (e.g., Lognormal), a subexponential distribution with infinite variance (e.g., Pareto), and a non-subexponential distribution (e.g., Exponential).

The observed lack of convergence of the sequential standard deviation plot for the data in Figure 4 translates directly into inconsistencies of models that assume finite moments upfront, either implicitly or explicitly. To illustrate, for the same HTTP data set, Figure 5 shows the sequential estimates $\hat{\sigma}^2(i)$ of the variance parameter $\sigma^2(i)$ of fitted Lognormal models M_i, together with their 95% confidence intervals CI_i (here we used $n_i = 1000 * i$). Figure 5 is evidence that the parameter estimates $\hat{\sigma}^2(i)$ don't converge and that successive 95% confidence intervals are so small so as to have little chance to overlap. In short, while for any

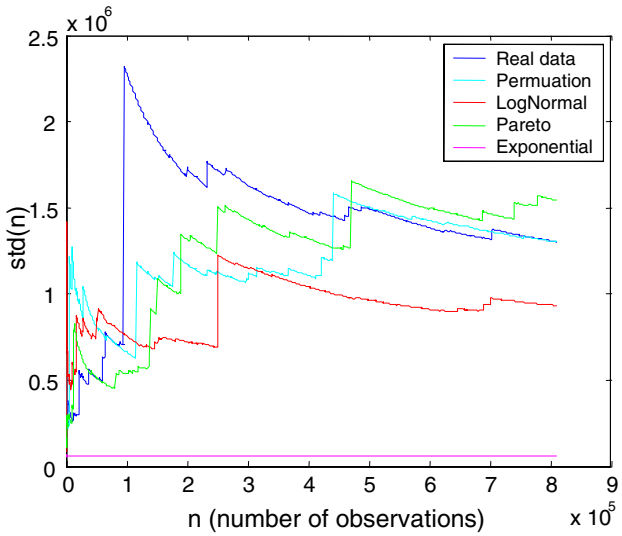

Figure 4: Sequential Standard Deviation Plot: Original Data (HTTP Data Set) and Fitted Distributions

fixed i, the fitted Lognormal model M_i appears to provide an adequate fit for the data set D_i, when viewed together, the disadvantage of using Lognormal distributions to fit our data sets becomes evident; the resulting models M_i are clearly inconsistent with one another, and while increasing the number of observations produces more accurate parameter estimates, an apparent lack of convergence of the latter renders the more precise estimates useless. To quote Mandelbrot (1997, p. 21), "when exactitude is elusive, it is better to be approximately right than certifiably wrong." For the data sets at hand, using the proposed framework shows that fitting Lognormals is a case of being "certifiably wrong."

We next apply our approach to show that fitting Pareto models to our data is indeed a case of being "approximately right." To this end, Figure 6 shows the sequential estimates $\hat{\alpha}(i)$ of the tail index parameter α of fitted Pareto models M_i, together with their 95% confidence intervals CI_i, where we used again $n_i = 1000 * i$. More precisely, we use here the well-known *Hill estimator* to estimate the tail index α of a Pareto distribution and exploit the fact that Hill's estimator is asymptotically normal to compute approximate 95% confidence intervals for $\hat{\alpha}(i)$. For details about Hill's estimator, conditions under which it is asymptotically normal, and an expression for the 95% confidence intervals, see for example Resnick (1997). The contrast between Figures 5 and 6 is telling. Not only is there evidence that the tail index estimates $\hat{\alpha}(i)$ converge as i increases to the full size of the data sets, but their corresponding CI_i's are such that the fitted Pareto models M_i are by and large fully consistent with one another. We take this as strong evidence that in the case of our data, Pareto models are not only "useful" but in fact "better" than Lognormal models. In this sense, model

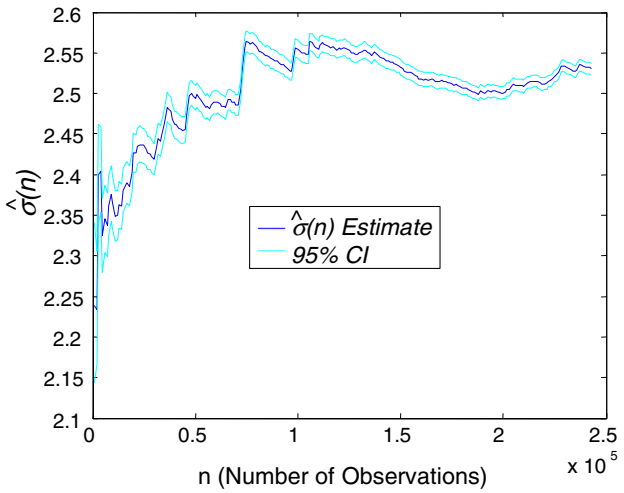

Figure 5: Sequential $\hat{\sigma}$ Estimate Plot, with Corresponding 95% Confidence Intervals for the HTTP Data Set

consistency is a powerful requirement and represents an effective criterion for selecting among otherwise omparable alternate models. It also benefits tremendously from the availability of voluminous data sets.

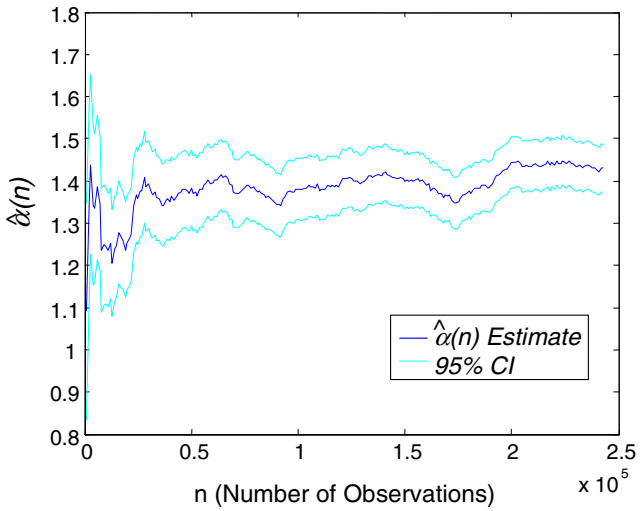

Figure 6: Sequential $\hat{\alpha}$ Estimate Plot, with Corresponding 95% Confidence Intervals for the HTTP Data Set

3.3 Beyond Borrowing Strength

While we have illustrated our approach with an example where a Pareto model is picked over a Lognormal model, the same method succeeds just as well in selecting a Lognormal over a Pareto when the underlying data is not consistent with a scaling distribution. For example, it is easy to see why and how our proposed framework would reject a Pareto model that was fitted to a large sample generated from a Lognormal distribution. In particular, our framework does not only apply to choosing among otherwise comparable distributions, but works just as well for selecting among alternate stochastic process models that are deemed reasonable for having generated the time series at hand in the first place. In fact, insisting on model consistency turns out to provide an especially powerful and elegant tool for determining whether the time series in question is consistent with long-range dependence or whether a short-range dependent process is a "better" model for the data (see for example, Beran 1994). Of course, there will always be situations where our approach will fail to identify the "best" model among competing candidates, but in this case, it almost certainly will be able to reveal whether the fitted candidate models are all uniformly "good" or "bad" with respect to the model consistency requirement.

While we advocate here that future modeling efforts should adhere more faithfully to Tukey's "borrowing strength" principle and thus to making model consistency a general requirement, the networking community already practices another data analysis principle that is also attributed to Tukey and is called *"broadening the basis"*. While related to "borrowing strength", "broadening the basis" refers more explicitly to attempts on generalizing a finding by drawing on a wider variety of data (Draper et al. 1993), collected under similar or even dissimilar conditions, at different points in space and time. Thus, in the networking context, broadening the basis is an approach that attempts to find law-like relationships that describe not a single set of measurements, but apply to many data sets collected from the same (or a similar) network or perhaps from very different networks, at different places within the network, over different period of time, and under varying networking conditions. IP flows are a perfect example where applying this principle has produced overwhelming evidence in favor of the scaling property of the size distribution of these basic constituents of aggregate network traffic.

4 DISCUSSION

4.1 Practical Considerations: Ambiguity in the Data

The mere observation that Gaussians and scaling distributions are both invariant under aggregation and marginalization suggests that the ubiquity with which the latter occur should be no more surprising than the wide-spread presence of the former. The fact that under the earlier-mentioned qualifications, scaling distributions are also invariant under maximization and mixture while Gaussians are not has a number of very practical implications for scientific modeling in general. For example, these stronger invariance properties make scaling distributions insensitive to a wide range of ambiguities that occur when measuring various quantities (see for example Bookstein 1990). Ambiguities commonly exist in levels of aggregation (e.g., grouping into

classes, choice of time segment), changing environments (e.g., entries or exits from a population, varying growth rates, different time segments), differences in accounting (e.g., treatment of multiple authorship) or measuring (e.g., off sets in clock times), etc. As a result of such robustness properties, the power of empirical studies can be vastly expanded by demonstrating, for example, that a given set of observations is not only consistent with a hypothesized model or distribution, but also that the finding is in fact insensitive to the ambiguities that are inherent in the process of obtaining the measurements in the first place (for more details, see Willinger et al. 2004). Moreover, the inherent robustness properties of the scaling distributions greatly facilitate scientific discovery, because they essentially ensure that detecting and identifying scaling laws in real data is not only feasible, but can be fully expected to succeed despite imperfect measurements as well as a wide range of ambiguities associated with the actual processes of measuring, accounting, and reporting the data. Properties of a system that require perfect measurements and tolerate no such ambiguities are highly unlikely to be useful, let alone be discovered.

4.2 Power Laws vs. High Variability

While the arguments found in the statistics literature concerning the use of scaling distributions for modeling high variability/infinite variance phenomena have hardly changed since Mandelbrot's attempts in the 1960s to bring scaling distributions into mainstream statistics, discovering and explaining strict power law relationships has become a minor industry in the complex science literature. Unfortunately, a closer look at the fascination within the complex science community with power law relationships reveals a very cavalier attitude towards inferring power law relationships or strict power law distributions from measurements. In fact, Figure 7 illustrates a commonly-used technique and widely-accepted approach for inferring power law behavior. Denoting by x the "size" of an object and by $f(x)$ its relative frequency of occurrence (with $\int f(x)dx = 1$), the plot on the left shows for a set of 1,000 observations the corresponding size-frequency relation on log-log scale and suggests that the data support a relation of the form $\log(f(x)) = log(c) - (1+\alpha)\log(x)$ for some $\alpha > 0$ (i.e., $f(x) = Cx^{-(1+\alpha)}$). Moreover, an estimate of the tail index α is readily available as the slope of a line that fits the data well. However, as illustrated this popular procedure lacks rigor and typically results in vastly different tail index estimates, depending on the degree of sophistication of the "eyeballing" technique used when fitting a straight line to the size-frequency data. In the case of Figure 7, the 1,000 observations are generated from a Pareto distribution with $\alpha = 1$, but the left plot suggests that α estimates between 0 and 1 would be appropriate.

To demonstrate why size-frequency plots should be altogether avoided when inferring power law relationships, we contrast their use in Figure 7 (left plot) with the use of the CCDF plots (right plot). Examining the CCDF plot corresponding to the same 1,000 observations as considered before, we observe that only tail index estimates around 1.0 are consistent with the data. In this sense, many of the discoveries of power law relationships reported in the complex science literature are simply the result of a non-rigorous and inadequate analysis of the underlying data and have to be revised/modified in view of the evidence provided by a statistically more robust analysis of the very same data. Unfortunately, the cavalier approach to dealing with power law relationships advocated by the complex science community has its appeal and has occasionally been adopted by other research communities. Other issues that—when ignored—readily invite claims of alleged power law relationships are discussed by Mandelbrot (1997) and concern, for example, a limited range of observed x-values, large tail index estimates, and ignoring naturally occurring small or large "outliers."

4.3 The Internet as a Case Study

To many researchers, the Internet offers an especially attractive case study of a large-scale complex system. The appealing allusion of a simple, robust, and homogeneous resource has led to a proliferation of specious claims about it in the physics and popular scientific literature. For example, the assumption that the Internet is sufficiently homogeneous and large-scale to be anemable to statistical physics-inspired analysis techniques has led to SOC-based models for explaining the self-similar nature of Internet traffic and has recently inspired the popular scale-free network models for explaining reported power law-type relationships associated with Internet connectivity. However, when trying to explain the presence of high variability in observed Internet traffic or Internet topologies, it is important to keep in mind that the vast majority of measurements from complex systems such as the Internet are almost never perfect, but are plagued by all kinds of ambiguities. Moreover, for a number of reasons, many quantities can hardly ever be measured directly and tend to be altered by manipulations and transformations of other measurements that allow for (hopefully) sufficiently accurate approximations of the quantities in question.

As far as Internet traffic is concerned, scaling distributions entered by way of a mathematical theorem originally due to Mandelbrot and Cox (see for example Leland et al. 1994, Crovella and Bestavros 1997 and references therein). The result states that the self-similar scaling behavior of the aggregate link traffic is caused by the high variability/infinite variance property of the individual constituents that make up the aggregate traffic. It also invites a direct validation of this explanation by identifying the nec-

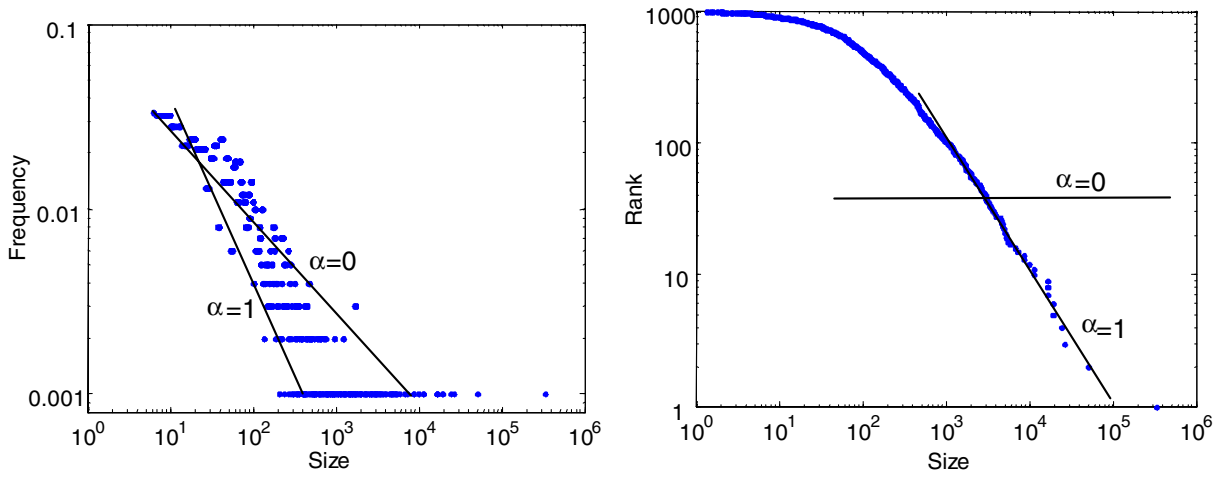

Figure 7: Size-frequency Plot on Log-log Scale (left) and Complementary Cumulative Distribution Plot on Log-log Scale (right) for 1000 Observations Sampled from a Pareto Distribution of the Second Kind with Parameters $\alpha = 1$ and $\beta = 100$

essary measurements (i.e., TCP connection sizes, IP flow sizes) and checking for consistency with scaling distributions or, equivalently, with heavy-tailed distributions, where the latter term is used in much of the self-similar traffic literature. A simple and popular descriptive model of this explanation is one in which most files ("mice") have few packets, while most packets are in large files ("elephants"). New results that expand on the theme of *Highly Optimized Tolerance (HOT)* (Carlson and Doyle 1999, Doyle and Carlson 2000, Carlson and Doyle 2002) consider Web layout design in the spirit of source coding for data compression and rate distortion theory. These results not only complete the present explanation of the observed self-similar Internet traffic, but they are a promising starting point for a more complete source/channel coding theory analogous to that from Shannon information theory for conventional communication problems, though necessarily differing greatly in detail (Zhu, Yu, and Doyle 2001). This new treatment builds on theories from robust control and duality in optimization and implies that scaling distributions at the level of traffic sources must be embraced because they are not an artifact of current network-specific features (e.g., applications, protocols), but are likely to be a permanent and essential characteristic of future network traffic (the bad news?). At the same time, the new approach also shows that not only can a new theory be developed to handle source-level scaling distributions, but if properly exploited, such behavior at the source level is in fact ideal for efficient and reliable transport over Internet-like networks (the good news!). Comparisons with statistical physics-inspired approaches have appeared elsewhere (e.g. Carlson and Doyle 1999, Doyle and Carlson 2000, Carlson and Doyle 2002, Willinger and Doyle 2003) and demonstrate that these SOC/EOC-based "explanations"—serving at best as simple null hypotheses—can be convincingly debunked and are easy to reject.

The use of scaling distributions (in the sense of strict power law relationships) to describe Internet topology entered by way of empirical studies, first reported by Faloutsos et al. (1999), who based their empirical findings on traceroute data that had been collected earlier by Pansiot and Grad (1998). The presence of a scaling relationship in node degree (i.e., number of connections) for routers within the Internet is critical to the scale-free story (Barabasi 2002), but it is well-known among networking researchers that the inferred node degrees are extremely ambiguous, and inferred node degree distributions that are not resilient to some of the most serious ambiguities are likely to be of little scientific value. When applied to Internet connectivity data (Willinger et al. 2004), our proposed approach provides evidence that in contrast to previously made claims, router-level node degrees are not consistent with scaling distributions. This finding is shifting the current research efforts from trying to explain scaling router-level degree distributions to understanding the high variability in measurements of IP-level connectivity. In particular, when trying to develop a modeling approach to router-level Internet connectivity that is truly *explanatory* in the sense of Willinger et al. (2002), the recently popular "scale-free" network models (Barabasi and Albert 1999) are not only inconsistent with more carefully interpreted router-level connectivity data, but are in almost every theoretical and practical aspect completely opposite from the real Internet. This realization is already yielding the beginnings of a first-principles approach to router-level topology modeling that reflects practical constraints and tradeoffs in networking technology and economics (Li et al. 2004).

REFERENCES

Bak, P. 1996. *How nature works: the science of self-organized criticality*. Copernicus.

Barabási, A.-L. 2002. *Linked: The New Science of Networks.* Perseus Publishing.

Barabási, A.-L. and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286: 509–512.

Barabási, A.-L. and Z.N. Oltavi. 2004. Network Biology: Understanding the cell's functional organization. *Nature Reviews Genetics* 5: 101–114.

Beran, J. 1994. *Statistics for Long-Memory Processes*. Chapman & Hall.

Bookstein, A. 1990. Informetric distributions, Part II: Resilience to ambiguity. *Journal of the American Society for Information Science* 41 (5): 376–386.

Buchanan, M. 2001. *Ubiquity: The Science of History…or Why the World Is Simpler Than We Think*. Crown.

Carlson, J. M. and J. C. Doyle. 1999. Highly Optimized Tolerance: a mechanism for power laws in designed systems. *Physics Review E* 60:1412–1428.

Carlson, J.M. and J.C. Doyle. 2002. Complexity and Robustness. *Proceedings of the National Academy of Science* 99(1): 2539–2545.

Crovella, M.E. and A. Bestavros. 1997. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking* 5(6): 835–846.

Downey, A.B. 2001. Evidence for long-tail distributions in the Internet. In *Proceedings of the 1st ACM SIGCOMM Internet Measurement Workshop*. 229–241. San Francisco, CA.

Doyle, J.C. and J. M. Carlson. 2000. Power laws, Highly Optimized Tolerance and generalized source coding. *Physics Review Letters* 84(24):5656–5659.

Draper, D., J.S. Hodges, C.L. Mallows, and D. Pregibon. 1993. Exchangability and data analysis. *Journal of the Royal Statistical Society A* 156: 9–37.

Faloutsos, M., P. Faloutsos, and C. Faloutsos. 1999. On Power-Law Relationships of the Internet Topology. *Proceedings of the ACM SIGCOMM*.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications, Volume 2*. New York: John Wiley & Sons.

Goldie, C.M. and C. Klüppelberg. 1998. Subexponential distributions. In *A Practical Guide to Heavy Tails*, R.J. Adler, R.E. Feldman, and M.S. Taqqu, eds. Boston: Birkhauser.

Johnson, N.L., S. Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions*. Vol. 1, 2nd Ed. New York: John Wiley & Sons.

Leland, W.E., M.S. Taqqu, W. Willinger, and D.V. Wilson. 1994. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking* 2 (1): 1–15.

Li, L., D. Alderson, J. Doyle, and W. Willinger. 2004. A First-Principles Approach to Understanding the Internet's Router-level Topology. *Proceedings of the ACM SIGCOMM*, Portland, Oregon.

Luria, S.E. and M. Delbrück. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28: 491–511.

Malamud, B.D., Morein, G., and D.L. Turcotte. 1998. Forest fires: an example of self-organized criticality. *Science* 281: 1840–1842.

Mandelbrot, B.B. 1997. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk*. New York: Springer-Verlag.

Pansiot, J.-J. and D. Grad. 1998. On routers and multicast trees in the Internet. *ACM Computer Communication Review* 28(1):41-50.

Resnick, S.I. 1987. *Extreme Values, Regular Variation, and Point Processes*. New York: Springer-Verlag.

Resnick, S.I. 1997. Heavy tail modeling in teletraffic data. *Annals of Statistics* 25: 1805–1869.

Samorodnitsky, G. and M.S. Taqqu. 1994. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. New York - London: Chapman and Hall.

Simon. H.A. 1955. On a class of skew distribution functions. *Biometrika* 42 (3/4): 425–440.

Tukey, J.W. 1986. Data analysis and behaviorial science or learning to bear the quantitative's man burden by shunning badmandments. In *The Collected Works of John W. Tukey*, ed. L.W. Jones, Vol. III. Wadsworth. Monterey, CA.

Whitt, W. 2002. *Stochastic-Process Limits*. New York: Springer-Verlag.

Willinger, W., D. Alderson, J.C. Doyle, and L. Li. 2004. A pragmatic approach to dealing with high variability measurements. *Proceedings of the ACM SIGCOMM Internet Measurement Conference 2004*. Taormina, Sicily, Italy.

Willinger, W. and J. C. Doyle. 2003. Robustness and the Internet: Design and Evolution. In *Robust design: A Repertoire of Biological, Ecological, and Engineering Case Studies*, ed. E. Jen, Oxford University Press (to appear).

Willinger, W., R. Govindan, S. Jamin, V. Paxson and S. Shenker. 2002. Scaling Phenomena in the Internet: Critically examining Criticality. *Proceedings of the National Academy of Science*, 99 (1): 2573–2580.

Willinger, W. and V. Paxson. 1998. Where mathematics meets the Internet. *Notices of the AMS* 45: 961–970.

Willinger, W., M.S. Taqqu, R. Sherman, and D.V. Wilson. 1997. Self-similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking* 5(1): 71–86.

Yule, G. 1925. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis *F.R.S. Philisophical Transactions of the Royal Society of London (Series B)* 213: 21–87.

Zhu, X., J. Yu, and J.C. Doyle. 2001. Heavy Tails, Generalized Coding, and Optimal Web Layout. *Proceedings of the IEEE Infocom*.

AUTHOR BIOGRAPHIES

WALTER WILLINGER is currently a member of the Information and Software Systems Research Center at AT&T Labs-Research, Florham Park, NJ. Before that, he was a Member of Technical Staff at Bellcore (1986-1996). He received the Diplom (Dipl. Math.) from the ETH Zurich, Switzerland, and the M.S. and Ph.D. degrees from the School of ORIE, Cornell University, Ithaca, NY. He has been a leader of the work on the self-similar ("fractal") nature of data network traffic and is co-recipient of the 1996 IEEE W.R.G. Baker Prize Award and the 1994 W.R. Bennett Prize Paper Award. He is a member of IEEE, ACM, SIAM, and INFORMS. His e-mail address is <walter@research.att.com>.

DAVID ALDERSON is a postdoctoral scholar in the Division of Engineering and Applied Sciences at Caltech. He received a B.S.E. in Civil Engineering and Operations Research from Princeton University and the M.S. and Ph.D. degrees from the Department of Management Science and Engineering at Stanford University. He is a member of INFORMS. His e-mail address is <alderd@caltech.edu>, and his web page is <www.cds.caltech.edu/~alderd>.

JOHN C. DOYLE is Professor of Control and Dynamical Systems, Bioengineering, and Electrical Engineering at Caltech. He has a BS and MS in EE, from MIT, 1977 and a PhD in mathematics, UC-Berkeley, 1984. Prize papers include the IEEE Baker (also ranked in the top 10 "most important" papers world-wide in pure and applied mathematics from 1981-1993), the IEEE AC Transactions Axelby (twice), and the AACC Schuck. Individual awards include the IEEE Centennial Outstanding Young Engineer, the IEEE Hickernell, the American Automatic Control Council (AACC) Eckman, and the Bernard Friedman. His e-mail address is <doyle@caltech.edu>, and his web page is <www.cds.caltech.edu/~doyle>.

LUN LI is a Ph.D. candidate in the Electrical Engineering department at Caltech. She received a B.E. from Tsinghua University and an M.S. in Mechanical Engineering from U.C. Berkeley. Her e-mail address is <lun@cds.caltech.edu>.