May 28, 2002

From gene families and genera to incomes and internet file sizes: why power-laws are so common in nature

William J. Reed
Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3P4

Barry D. Hughes
Department of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia

We present a simple explanation for the occurrence of power-law tails in statistical distributions, by showing that if stochastic processes with exponential growth in expectation are killed (or observed) randomly, the distribution of the killed or observed state exhibits power-law behavior in one or both tails. This simple mechanism can explain power-law tails in the distributions of the sizes of incomes, cities, internet files, biological taxa and in gene family and protein family frequencies.

PACS numbers: 02.50.Cw, 02.50.Ey

Distributions with power-law behavior in one or both tails are ubiquitous in physics, biology, geography, economics, insurance, lexicography, internet ecology, *etc.* Associated names include *Zipf's law* (word frequencies, city sizes), *Pareto's law* (incomes) and the *rank-size property* (cities, firms *etc.*) There have been many attempts to explain why such distributions are so common, including notions of self-organized criticality [1] and highly optimized tolerance [2]. We shall show that power-law tail behavior can arise from a very simple mechanism that can explain its occurrence in many instances.

The basic idea is that if a process that grows exponentially, in a loose sense, is 'killed' (or observed once) 'randomly', the distribution of the killed (observed) state will follow power laws in one or both tails. Consider the simple case of deterministic exponential growth $X(t) = e^{\mu t}$ killed at a random time T which is exponentially distributed with parameter ν. The killed state $\bar{X} = e^{\mu T}$ has the probability density function $f_{\bar{X}}(x) = (\nu/\mu)x^{-\nu/\mu-1}$ for $x > 1$, giving power-law behavior over its full range. We shall consider four stochastic processes all exhibiting exponential or geometric growth in expectation. We denote expectation by E, and freely use the conditional expectation identity $E[f(X)] = E_Y\{E[f(X)|Y]\}$.

Geometric Brownian motion (GBM): where dB_t has a normal distribution with mean 0 and variance dt,

$$dX = \mu X dt + \sigma X dB_t, \quad E(X_t|X_0) = X_0 \exp(\mu t). \quad (1)$$

Discrete multiplicative process: with $\{Z_n\}$ independent identically distributed random variables with mean μ,

$$X_{n+1} = Z_n X_n, \quad E(X_n|X_0) = X_0\mu^n. \quad (2)$$

Homogeneous birth-and-death process: with birth and death rates λ and δ,

$$\begin{aligned} P(X_{t+h} = n+1|X_t = n) &= \lambda nh + o(h), \\ P(X_{t+h} = n-1|X_t = n) &= \delta nh + o(h), \\ P(X_{t+h} = n|X_t = n) &= 1 - (\lambda+\delta)nh + o(h), \\ E(X_t|X_0) &= X_0 e^{(\lambda-\delta)t}. \end{aligned} \quad (3)$$

Galton-Watson branching process: with $\{Z_i\}$ independent identically distributed random variables representing numbers of offspring and $E(Z_i) = \mu$,

$$X_{n+1} = Z_1 + Z_2 + \cdots + Z_{X_n}, \; E(X_n|X_0) = X_0\mu^n. \quad (4)$$

We consider the killed state of these processes when killing occurs at random. For the discrete-time processes (2) and (4), we assume that the discrete hazard is constant: if the process has not been killed by time $n-1$, the probability of it being killed at time n is a constant p, independent of n, giving the geometric distribution for the generation number N corresponding to the killed state:

$$P(N = n) = p(1-p)^{n-1}, \; n = 1, 2, 3, \ldots ; \quad (5)$$

we exclude killing in the zeroth generation. For the continuous-time processes (1) and (3), we assume a constant hazard rate ν, giving the exponential distribution

$$P(\text{killed at time} \geq t) = e^{-\nu t}. \quad (6)$$

To find the distribution of the killed state we use the *moment generating function* (mgf) $E[e^{Xs}]$ for the continuous state processes (1) and (2), and the *probability generating function* (pgf) $E[s^X]$ for the discrete state processes (3) and (4). The pgf of the geometric distribution (5) is

$$E[s^N] = ps[1-(1-p)s]^{-1}. \quad (7)$$

A number of plots can reveal power-law behavior in an empirical size distribution. If there is lower-tail power-law behavior, a plot of the empirical cumulative distribution function (cdf) on logarithmic axes should be close to linear at its lower end. Equivalently a logarithmic plot of the (ascending) rank against size should be close to linear at the lower end. For upper-tail power-law behavior a logarithmic plot of the empirical survivor function (or complementary cdf) should be linear at its upper end, as should be a logarithmic plot of descending rank against size. One can also look for linearity in a plot of frequencies against size on logarithmic axes. However many sizes (especially extreme ones) will not occur at all or occur only once, and so some binning will probably be required.

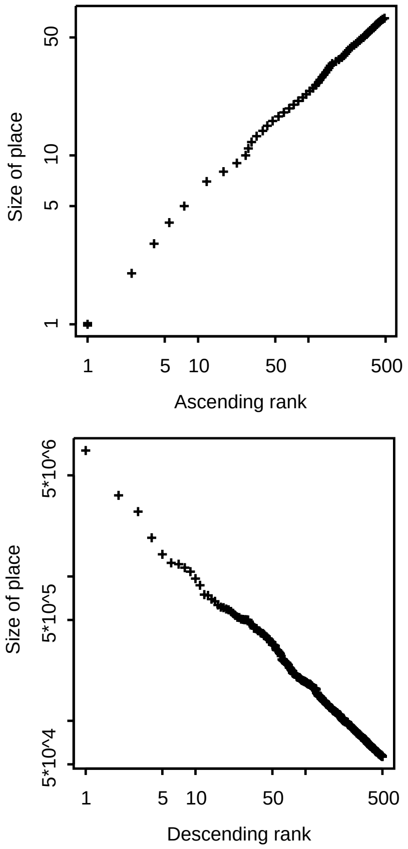

FIG. 1: Distribution of the sizes of 19,399 places in the U.S.A in 1999. Logarithmic rank-size plots of the smallest 500 places (upper panel) and the largest 500 places (lower panel) suggest power-law behavior in both tails of the distribution.

Killed geometric Brownian motion. Let $\bar{X}$ be the killed state and let $\bar{Y} = \log \bar{X}$. The mgf of $\bar{Y}$ is $E[\exp(\bar{Y}s)] = E_T\{E[\exp(Y_T s)|T]\}$. Since $Y_T - Y_0 = \log(X_T/X_0)$ has the normal distribution $\mathrm{N}((\mu - \sigma^2/2)T, \sigma\sqrt{T})$ [3], we find $E[\exp(Y_T s)|T] = \exp[Y_0 s + (\mu - \sigma^2/2)Ts + \frac{1}{2}\sigma^2 Ts]$ and

$$E[\exp(\bar{Y}s)] = e^{Y_0 s}\alpha\beta(\alpha - s)^{-1}(\beta + s)^{-1}, \qquad (8)$$

where $Y_0 = \log X_0$, while α and $-\beta$ ($\alpha, \beta > 0$) are the roots of the quadratic $\frac{1}{2}\sigma^2 s^2 + (\mu - \frac{1}{2}\sigma^2)s - \nu = 0$. Equation (8) is the mgf of the probability density function $f_{\bar{Y}}(y) = A\{e^{-\alpha(y-Y_0)}H(y - Y_0) + e^{\beta(y-Y_0)}H(Y_0 - y)\}$, an asymmetric Laplace distribution; $A = \alpha\beta/(\alpha + \beta)$ and the Heaviside function H is positive if its argument is positive, and zero if its argument is negative. It follows that the distribution of $\bar{X}$ is a *two-sided Pareto* or *double Pareto* distribution, with density

$$f_{\bar{X}}(x) = \begin{cases} AX_0^{-\beta}x^{\beta-1} & \text{if } x \leq X_0, \\ AX_0^{\alpha}x^{-\alpha-1} & \text{if } x > X_0, \end{cases} \qquad (9)$$

which exhibits power-law behavior in both tails.

Equation (9) has been used to explain the upper-tail power-law phenomenon observed for incomes (Pareto's law of incomes) [4] and city sizes (the rank-size law) [5].

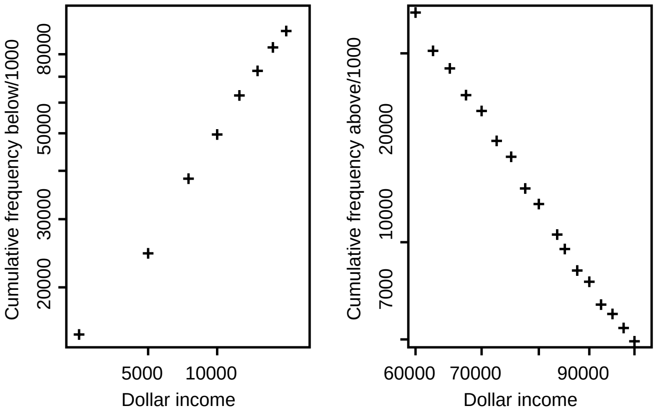

FIG. 2: Distribution of the total money income of 216 million people in U.S.A in 2000. The left and right panels show (on logarithmic axes) the cumulative frequency distributions (binned) in the lower and upper tails respectively (unbinned cumulative frequency plots, like those in Fig. 1, are not available since income data is only published in binned form). The plots suggest power-law behavior in both tails.

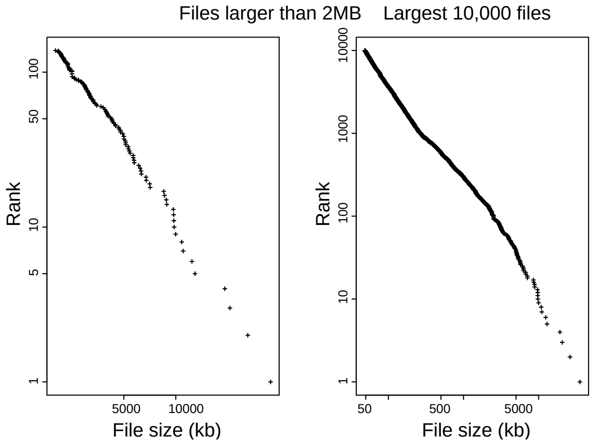

FIG. 3: Logarithmic rank-size plots for the upper tail of the distribution of 734,814 http response sizes at the University of North Carolina at Chapel Hill Main Link [10].

It is assumed that both individual incomes and settlement sizes evolve as GBM, while the time that an individual has been earning and the time a settlement has been in existence can reasonably be modelled as having exponential distributions, based on the assumption that the workforce and the population of settlements are growing at a fixed rate. The predicted lower-tail power-law behavior has been shown to occur in the empirical distributions of both incomes and human settlement sizes, facts previously unrecognized. Figures 1 and 2 illustrate empirical power-law behavior in both tails of settlement size and income distributions, using recent U.S. data [6].

Killed discrete multiplicative process. We distinguish three cases for the multiplicative process $X_{n+1} = Z_n X_n$:
(a) monotonic increasing, when $P(Z_n > 1)) = 1$;
(b) monotonic decreasing, when $P(Z_n < 1) = 1$;

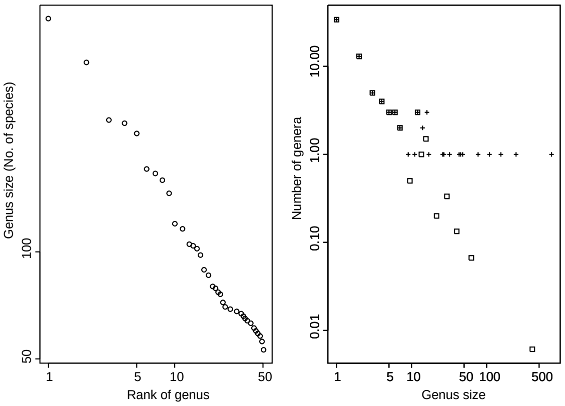

FIG. 4: Logarithmic plots of the observed sizes of 1829 genera of North American vascular plants. The left-hand panel is a rank-size plot for the largest 50 genera. The straight-line behavior suggests upper-tail power-law behavior in the genus size distribution. The right-hand panel shows frequency against genus size (crosses) along with a similar plot with binning (boxes), so that each bin contains at least two genera.

(c) bidirectional, when $P(Z_n > 1) > 0$, $P(Z_n < 1) > 0$. Again let $\bar{Y} = \log \bar{X}$, so that $\bar{Y} = \log X_0 + \sum_{j=1}^{N} U_j$, where N is distributed geometrically with parameter p in accord with Eq. (5), and $U_j = \log Z_j$. For brevity we write the mgf of each U_j as $E[\exp(sU_j)] = M_U(s)$. It follows that $E[\exp(\bar{Y}s)] = X_0^s E_N[M_U(s)^N]$ and we recognize $E_N[M_U(s)^N]$ as the pgf (7) of N evaluated at $M_U(s)$. Writing for brevity $q = 1 - p$, we have

$$E[\exp(\bar{Y}s)] = X_0{}^s spM_U(s)[1 - qM_U(s)]^{-1}.$$

The tail behavior of the density of $\bar{Y}$ is determined by singularities in the mgf from solutions of $M_U(s) = 1/q$. Since $M_U(0) = 1$ and $M_U''(s) > 0$, provided that the mgf of U exists in the neighbourhood of $s = 0$ [7], real zeros of $M_U(s) - 1/q$ are simple zeros. There are three cases.

(a) $U_j > 0$ (i.e. $Z_j > 1$) with probability 1, so the process $\{X_n\}$ is increasing. Because $M_U(s)$ is increasing with $M_U(s) \to \infty$ as $s \to \infty$ and $M_U(s) \to 0$ as $s \to -\infty$, there is a unique simple zero of $M_U(s) - 1/q$ at $s^+ > 0$, giving a simple pole of $E[\exp(\bar{Y}s)]$ at s^+ and so an upper power-law tail in the distribution of the killed state $\bar{X}$.

(b) $U_j < 0$ (i.e. $Z_j < 1$) with probability 1, so the process $\{X_n\}$ is decreasing. Because $M_U(s)$ is decreasing with $M_U(s) \to 0$ as $s \to \infty$ and $M_U(s) \to \infty$ as $s \to -\infty$ and there is unique simple zero of $M_U(s) - 1/q$ at $s^- < 0$, giving a simple pole of $E[\exp(\bar{Y}s)]$ at s^- and so a lower power-law tail in the distribution of the killed state $\bar{X}$.

(c) $P(U_j > 0) > 0$ and $P(U_j < 0) > 0$, so the process $\{X_n\}$ can both increase and decrease. Here $M_U(s) \to \infty$ as s tends to ∞ or $-\infty$. As $M_U(s)$ is convex, $E[\exp(\bar{Y}s)]$ has two isolated singularities of opposite sign, both simple poles, and $\bar{X}$ has power-law behavior in both tails.

A multiplicative model for growth in file sizes coupled with a model which yielded geometric killing was used by Huberman and Adamic [8] to explain upper-tail power-law behavior in the size (number of pages) of World-Wide Web sites. Mitzenmacher [9] used a similar model (with Z_n assumed to have a lognormal distribution) to explain the phenomenon of power-law behavior in both tails of the distribution of the size of computer files. Sample data with the characteristic power-law rank-size property is shown in Figure 3.

Killed birth-and-death process. Let $\bar{X}$ be the value of X at the time of killing. It can be shown [11] that in the case $\lambda > \delta$ the distributions of $\bar{X}$ can exhibit power-law behavior in the upper tail. Precisely, as $n \to \infty$,

$$P(\bar{X} = n) \sim k_1 n^{-[1+\nu/(\lambda-\delta)]} \text{ for } \lambda > \delta. \tag{10}$$

We have 'stretched exponential' behavior in the case $\lambda = \delta$, and $P(\bar{X} = n) \sim k_2(\lambda/\delta)^n n^{-[1+\nu/(\delta-\lambda)]}$ for $\lambda < \delta$. Reed and Hughes [11] used this model to explain the distribution of the size (number of species) of live biological genera. It is assumed that species are created from existing species by speciations which occur independently and at random; and that species likewise suffer individual extinctions independently and at random. Thus the evolution of the number of living species can be represented by a birth-and-death process (*e.g.* [12]). Genera are assumed to be created in a similar fashion to species [13], so that the time since origination of a live genus is exponentially distributed and the current size of such a genus is that of a randomly killed birth-and-death process (*i.e.* of $\bar{X}$). Figure 4 shows a logarithmic rank-size plot for the largest 50 genera of North American vascular plants; and a frequency plot of all 1829 living genera of such plants. The rank-size plot is approximately linear, as is the frequency plot for larger genera, consistent with upper-tail power-law behavior as predicted by the model.

Let $\bar{Y}$ be the number of population elements that have existed up until the time of killing. It can be shown [14] in the case $\lambda > \delta + \nu$ that $\bar{Y}$ also exhibits upper-tail power-law behavior. That is, as $m \to \infty$

$$P(\bar{Y} = m) \sim c_1 m^{-[1+\nu/(\lambda-\delta)]} \quad \text{for } \lambda > \delta + \nu. \tag{11}$$

This result may explain upper-tail power-law behavior in the size distribution of extinct fossil taxa (*e.g.* [15]), with killing corresponding to a cataclysmic extinction event (when the whole taxon is destroyed), with such events assumed to occur in a Poisson process with rate ν.

The upper-tail power-law behavior of the distribution of the size of living genera has long been known. Yule [13] developed the eponymous *Yule distribution* to fit such data,using essentially the above model with $\delta = 0$. He assumed that the size of a genus was that of a killed pure birth (or Yule) process. That similar behavior occurs for the distribution of extinct taxa has been observed [15].

The randomly killed birth and death process may also provide a better model for power-law distributions in gene family and protein family size distributions [16]. Assume that the size of a gene family evolves as homogeneous birth and death death process [17] so that new genes in the family can arise from existing ones independently at random, and similarly may be lost. If gene

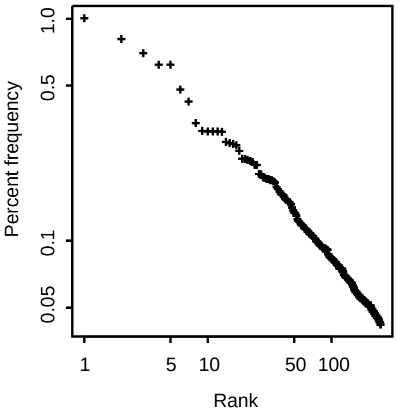

FIG. 5: Distribution of family names in the 1990 U.S. Census. We show a logarithmic plot of rank against frequency for the most common 250 names. The upper left point corresponds to the most common name.

families evolve in a Yule process then time in existence of family is exponentially distributed, and a power-law tail results, in the same way as in the taxon model above.

Killed Galton–Watson branching process. The pgf for the number X_n of individuals in the nth generation of a Galton–Watson branching process $X_{n+1} = Z_1 + Z_2 + \cdots + Z_{X_n}$, started with one individual for the zeroth generation, is given [18] by $G_n(s) = G_{n-1}(g(s))$. Here $g(s) = G_1(s)$ is the pgf for the number of offspring of an individual. The pgf $G(s) = E(s^{\bar{X}})$ of the state $\bar{X}$ of the branching process killed on the production of the Nth generation according to the geometric distribution (5) is $G(s) = \sum_{n=1}^{\infty} G_n(s)p(1-p)^{n-1}$, and it satisfies the functional equation $G(s) = pg(s) + (1-p)G(g(s))$. Functional equations of this kind were encountered in stochastic processes by Hughes, Shlesinger and Montroll [19], who observed a close analogy with real-space renormalization methods and antecedents in the theory of nondifferentiable functions, and noncontinuable analytic functions. By analysing the singular behavior of the solutions of the functional equation, we have argued elsewhere [20] that if the offspring distribution has finite mean $\mu = g'(1) = E(Z_j)$, the dominant behavior $P(\bar{X} = m) \sim R(m)m^{-1-\kappa}$ will be found as $m \to \infty$, where $\kappa = \log[(1-p)^{-1}]/\log\mu$ and $R(m)$ has log-periodic oscillations, that is, $R(m)$ is periodic in $\log m$ with period $\log\mu$. The existence of the oscillations can be rigorously proved [20] when the offspring distribution is geometric, but the oscillations are of very small amplitude. Recently, Gluzman and Sornette [21] have reviewed the existence of log-periodic oscillations mirroring underlying scale hierarchies in several areas of physics.

Since Galton proposed the branching process as a model for family names (and Watson partially solved the problem of the probability of extinction of a name), we have investigated the applicability of the killed branching process model to the size distribution of names (under the hypothesis that new names can enter either *via* immigration in a Poisson process, or from a mutation of any existing name, which can occur with constant probability). Figure 5 shows the rank-size plot for US surnames [22]. The closeness of the points to a straight line (corresponding to power-law behavior) is impressive. Similar plots of data [23] for Isle of Man surnames in 1881 and Chinese family names show the same linearity. Grouped frequency plots also provide evidence of power law tails.

[1] P. Bak, C. Tang, and K. Wiesenfeld, (1987) *Phys. Rev. Lett.* **59**, 381 (1987).

[2] M. Newman, *Nature* **405**, 412 (2000).

[3] See, for example, Chapter 3 of X. Mao, *Stochastic Differential Equations* (Horwood, Chichester, 1997).

[4] W.J. Reed *Econ. Lett.* **74**, 15 (2001).

[5] W.J. Reed, *J. Regional Sci.* **42**, 1 (2002).

[6] For data, see ferret.bls.census.gov/macro/032001/perinc/new01_001.htm.

[7] If U has no mgf, use the characteristic function $E(e^{itU})$. For a symmetric stable density of order $\beta < 2$, we have $E(e^{itU}) = \exp(-b|t|^{\beta})$, and $P(0 < \bar{X} < x) \sim C_1/|\log x|^{\beta}$ as $x \to 0$, while $P(\bar{X} > x) \sim C_2/(\log x)^{\beta}$ as $x \to \infty$.

[8] B.A. Huberman and L.A. Adamic *Nature* **401**, 131 (1999); L.A. Adamic and B.A. Huberman *Comm. ACM* **44**, 55 (2001).

[9] M. Mitzenmacher, preprint, www.eecs.harvard.edu/˜michaelm/NEWWORK/papers.html (2001).

[10] Data for 7 days (April 2001) courtesy of Dr J.S. Marron, www.orie.cornell.edu/˜marron/OR778NetworkData/OR778home.html.

[11] W.J. Reed and B.D. Hughes, *J. Theor. Biol.*, in press (2002).

[12] D.M. Raup, *Paleobiology* **11**, 45 (1982).

[13] G.U. Yule, *Phil. Trans. Roy. Soc. Lond. B* **213**, 21 (1924).

[14] B.D. Hughes and W.J. Reed (2001), preprint.

[15] e.g. by B. Burlando, *J. Theor. Biol.* **163**, 161 (1993). A branching process model was proposed by J. Chu and C. Adami, *Proc. Nat. Acad. Sci. (U.S.A.)* **96**, 15017 (1999).

[16] M.A. Huynen and E. van Nimwegen, *Mol. Biol. Evol.* **15**, 583 (1998); J. Qian, N.M. Luscombe and M. Gerstein, *J. Mol. Biol.* **313**, 673 (2001).

[17] cf. J.S. Bader, arxiv.org/abs/physics/9908032 (1999).

[18] T.E. Harris, *The Theory of Branching Processes* (Springer, Berlin, 1963).

[19] B.D. Hughes, M.F. Shlesinger and E.W. Montroll, *Proc. Nat. Acad. Sci. (U.S.A.)* **78** 3287 (1981); M.F. Shlesinger and B.D. Hughes, *Physica* **109A**, 597 (1981).

[20] W.J. Reed and B.D. Hughes, submitted to *Physica A*.

[21] D. Sornette, *Phys. Rep.***297**, 239 (1998); S. Gluzman and D. Sornette, *Phys. Rev. E* **65**, 036142 (2002).

[22] See www.census.gov/ftp/pub/genealogy/www/ freqnames.html.

[23] See www.geocities.com/hao510/namefreq and www.ee.surrey.ac.uk/Contrib/manx/famhist/fnames/sn1881.htm.