# Two Universality Properties Associated with the Monkey Model of Zipf's Law

**Richard Perline [1,*] and Ron Perline [2]**

[1]    Independent Researcher, 34–50 80th Street, Jackson Heights, New York, NY 11372, USA
[2]    Department of Mathematics, Drexel University, Korman Center at 33rd and Market Streets,
       Philadelphia, PA 19104, USA; rperline@math.drexel.edu
[*]    Correspondence: richperline@gmail.com; Tel.: +1-718-541-5315

**Abstract:** The distribution of word probabilities in the monkey model of Zipf's law is associated with two universality properties: (1) the exponent in the approximate power law approaches $-1$ as the alphabet size increases and the letter probabilities are specified as the spacings from a random division of the unit interval for any distribution with a bounded density function on $[0,1]$; and (2), on a logarithmic scale the version of the model with a finite word length cutoff and unequal letter probabilities is approximately normally distributed in the part of the distribution away from the tails. The first property is proved using a remarkably general limit theorem from Shao and Hahn for the logarithm of sample spacings constructed on $[0,1]$ and the second property follows from Anscombe's central limit theorem for a random number of independent and identically distributed (i.i.d.) random variables. The finite word length model leads to a hybrid Zipf-lognormal mixture distribution closely related to work in other areas.

## 1. Introduction

In his popular expository article on universality examples in mathematics and physics Tao [1] mentions Zipf's empirical power law of word frequencies as lacking any "convincing explanation for how the law comes about and why it is universal". By *universality* he means the idea where systems follow certain macroscopic laws that are largely independent of their microscopic details. In this article we drop down many levels from natural language to discuss *two* universality properties associated with the toy monkey model of Zipf's law. Both of these properties were previously considered in Perline [2], but the presentation there was incompletely developed, and here we clarify and greatly expand upon these two ideas by showing: (1) how the model displays a nearly universal tendency towards a $-1$ exponent in its approximate power law behavior; and (2) the significance of the central limit theorem (CLT) for a random number of i.i.d. variables in the model with a finite word length cutoff. We will sometimes refer to the case with a finite word length cutoff as *Monkey Twitter*, as explained in Section 3. Our analysis regarding the very general tendency towards a $-1$ exponent when letter probabilities are drawn as a sample from any of an enormous class of probability distributions is a new result in the literature on this topic. Our application of the CLT with a random number of summands was first introduced with respect to the monkey model in [2], and we want to emphasize here that this is a significant aspect of the structure of the model.

Our paper is laid out as follows. In Section 2 we prove the strong tendency towards an approximate $-1$ exponent under very broad conditions by means of a limit theorem for the logarithms

of random spacing due to Shao and Hahn [3]. An equivalent but relatively longer proof of this, still based on [3], has been given in [4]. In Section 3, we explain the underlying lognormal structure of the *central part* of the distribution of word probabilities in the case of a *finite word length* cutoff and show how it leads to a hybrid Zipf-lognormal distribution (called a lognormal-Pareto distribution in [2]). These are universality properties in exactly Tao's sense because the distribution of the word probabilities (the macro behavior) is within very broad bounds independent of the details of how the letter probabilities for the monkey's typewriter are selected (the micro behavior). In Section 4, we discuss the multiple connections between these results and well-known research from other areas where Pareto–Zipf type distributions have been investigated. In the remainder of this section we sketch some historical background.

It has been known for many years that the monkey-at-the-typewriter scheme for generating random "words" conforms approximately to an inverse power law of word frequencies and therefore roughly mimics the approximate inverse power form of Zipf's [5] statistical regularity for many natural languages [6]. *However, Zipf's empirical word frequency law not only approximates a power distribution, but as he emphasized, it also very frequently exhibits an exponent in the neighborhood of* −1. That is, letting $f_1 \geq f_2 \geq \ldots f_r \cdots \geq \ldots$ represent the observed ranked word frequencies, Zipf found $f_r \approx Cr^{-1}$, with $C > 0$ a constant depending on the sample size. Many qualifications have been raised regarding this approximation (including the issue of the divergence of the harmonic series); yet as I.J. Good [7] commented: "The Zipf law is unreliable but it is often a good enough approximation to demand an explanation". Figure 1 illustrates the fascinating universal character of this word frequency law using the text from four different authors writing in four different European languages in four different centuries. The plots are shown in the log-log form that Zipf employed, and they resemble countless other examples that researchers have studied since his time. The raw text files for the four books used in the graphs were downloaded from the Public Gutenberg databank [8].
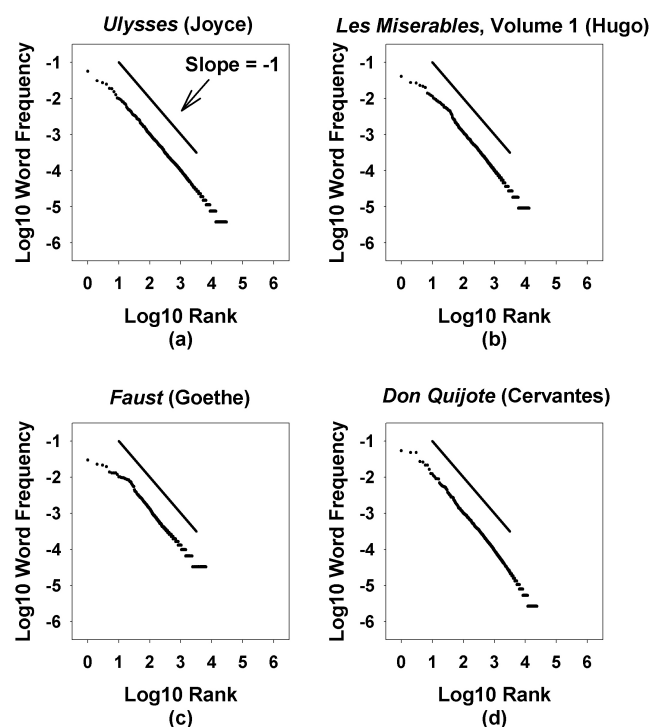


**Figure 1.** Log-log plots of relative word frequencies by rank for four authors writing in four different European languages in four different centuries. The approximate −1 slope in all the graphs is an iconic feature of Zipf's word frequency law.

The monkey model has a convoluted history. The model can be thought of as actually embedded within the combinatoric logic of Mandelbrot's early work providing an information-theoretic explanation of Zipf's law. For example, in Mandelbrot [9] there is an appendix entitled "Combinatorial Derivation of the Law $p_r = Pr^{-B}$" (his notation) that effectively specifies a monkey model, including a Markov version, but is not explicitly stated as such. However, his derivation there is informal, and the first completely rigorous analysis of the general monkey model with independently combined letters was given only surprisingly recently by Conrad and Michenmacher [10]. Their analysis utilizes analytic number theory and is, as a result, somewhat complicated—a point they note at the end of their article. A simpler analysis using only elementary methods based on the Pascal pyramid has now been given by Bochkarev and Lerner [11]. They have also analyzed the more general Markov problem [12] and hidden Markov models [13]. Edwards, Foxall and Perkins [14] have provided a directly relevant analysis in the context of scaling properties for paths on graphs explaining how the Markov variation can generate both an approximate power law or a weaker scaling law, depending on the nature of the transition matrix.

A clarifying and important milestone in the history of this topic came from the cognitive psychologist, G.A. Miller [15,16]. Miller used the simple model with equal letter probabilities and an independence assumption to give a heuristic analysis showing that this version generates a distribution of word probabilities that yields a step function approximation to an inverse power law such that the probability $P_r$ of the $r$<sup>th</sup> largest word probability is (in a sense that needs to be made precise) roughly of the form $P_r \approx Cr^{-\beta}$, for $-\beta < -1$. *In addition*, he made the very interesting empirical observation that by using a keyboard of $K = 26$ letters and one space character, with the space character having a probability of 0.18 (about what is seen in empirical studies of English text), the value of $-\beta$ in this case is approximately $-1.06$, very close to the nearly $-1$ value in samples of natural language text as illustrated in Figure 1. In his simplified model, it turns out that $-\beta = -1 + \ln(1-s)/\ln K$, where $0 < s < 1$ is the space probability, so that $-\beta$ approaches $-1$ from below as $K$ increases. (Natural logarithms are denoted $\ln(x)$ and logarithms with any other radix will be explicitly indicated, as in $\log_K(x)$. Note that Miller's $-\beta$ involves a a *ratio* of logarithms, so that it is invariant with respect to the radix used in the numerator and denominator.) *In this article we give a broad generalization of Miller's observation to the case of unequal letter probabilities.* In light of our present results given in Section 2, the application of sample spacings from a uniform distribution in [2] to study the $-1$ behavior using an asymptotic regression line should now be viewed as just a step in a direction that ultimately led us to the Shao and Hahn limit used here.

## 2. Proof that $-\beta$ Tends Towards $-1$ from the Asymptotics of Log-Spacings

For our analysis we specify a keyboard with an alphabet of $K \geq 2$ letters $\{L_1 \ldots, L_K\}$ and a space character $S$. The letter characters have non-zero probabilities $q_1, q_2, \ldots, q_K$. The space character has probability $s$, so that $\sum_{i=1}^{K} q_i + s = 1$. A word is defined as any sequence of non-space letters terminating with a space. A word $W$ of exactly $n$ letters is a string such as $W = L_{i_1} L_{i_2} \ldots L_{i_n} S$ and has a probability of the form $P(W) = P = q_{i_1} q_{i_2} \ldots q_{i_n} s$ because letters are struck independently. The space character with no preceding letter character will be considered a word of length zero. The rank ordered sequence of descending word probabilities in the ensemble of all possible words is written $P_1 = s > P_2 \geq \ldots P_r \geq \ldots$ ($P_1 = s$ is always the first and largest word probability). We break ties for words with equal probabilities by alphabetical ordering, so that each word probability has a unique rank $r$.

Conrad and Mitzenmacher [10] give a carefully constructed definition of power law behavior in the context of the monkey model as the situation where there exist $C_1, C_2, \beta$ such that the inequality

$$C_1 r^{-\beta} \leq P_r \leq C_2 r^{-\beta} \tag{1}$$

holds for sufficiently large $r$. It turns out that $\beta > 1$. A few comments about this definition are in order. First, note that there will always be many words with tied probability values. For example,

the words $L_1 L_2 S$ and $L_2 L_1 S$ have the same probability $q_1 q_2 s$. Because of this, the word probabilities in this model could never *exactly* conform to a perfect inverse power law $P_r = Cr^{-\beta}$. The asymptotic behavior for $P_r$ is more subtle than this and highlights the significance of Conrad and Mitzenmacher's analysis. They are able to give a stronger result when at least two letter probabilities have an irrational log-ratio $\ln q_i / \ln q_j$. For this case, they show that there exists a constant $C$ such that $P_r \sim Cr^{-\beta}$, where $\sim$ denotes asymptotic equivalence, $\lim_{r \to \infty} P_r / (Cr^{-\beta}) = 1$. For our purposes here, we will simply refer to the Conrad and Mitzenmacher results as generally implying that $P_r$ has *approximate power law behavior*. Note that inequality (1) implies the weaker asymptotic condition $\ln P_r \sim -\beta \ln r$ as $r \to \infty$. We'll call this *asymptotic log-log linear* behavior and observe that on a log-log scale, $\ln C_1$ and $\ln C_2$ are asymptotically negligible so that only the exponent $-\beta$ becomes significant. As can be seen in Figure 2b–d, to be explained shortly, monkey model word probabilities generated from unequal letter probabilities in the manner described below produce graphs distinctly log-log linear.

Conrad and Mitzenmacher derive the exponent value $-\beta$ but do not express it in what we consider its simplest form. Following Bochkarev and Lerner [11], the parameter $\beta$ can be represented as the solution to

$$q_1^{1/\beta} + q_2^{1/\beta} + \cdots + q_K^{1/\beta} = 1, \tag{2}$$

with $\beta > 1$. Bochkarev and Lerner [11] indicate $\ln p(r) - \ln r / \gamma$ in the inequality for their Theorem 1, but what was evidently intended is $\ln p(r) + \ln r / \gamma$. Their $1/\gamma$ is equal to our $\beta$. In the Miller model with equal letter probabilities, $q_1 = \cdots = q_K = (1-s)/K$, so from Equation (2), $\beta$ in this case is found to be $1 - \ln(1-s) / \ln K$. In the Fibonacci example given by Conrad and Mitzenmacher [10] and in Mitzenmacher [17] they use $K = 2$ letters with probabilities $q_1, q_2 = q_1^2$ and $q_1 < (-1 + \sqrt{5})/2$ so that $q_1 + q_2 < 1$. Then $\beta$ is the solution to $q_1^{1/\beta} + q_1^{2/\beta} = 1$, giving

$$\beta \quad = \quad \frac{\ln q_1}{\ln((-1 + \sqrt{5})/2)}. \tag{3}$$

Conditions leading to $-\beta \approx -1$ were not explored in [10,11]. To understand how this behavior can emerge, we define spacings through a random division of the unit interval and then state the Shao-Hahn limit law. Let $X_1, X_2, \ldots, X_{K-1}$ be a sample of $K-1$ i.i.d. random variables drawn from a distribution on $[0,1]$ with a bounded density function $h(x) \leq M$. Shao and Hahn present the conditions for this limit in a more general way that reduces to this simpler statement when a density function exists. Write the order statistics of the sample as $X_{1:K-1} \geq X_{2:K-1} \cdots \geq X_{K-1:K-1}$. The $K$ *spacings* $D_i$ are defined as the differences between the successive order statistics: $D_1 = 1 - X_{1:K-1}$, $D_i = X_{i-1:K-1} - X_{i:K-1}$ for $2 \leq i \leq K-1$ and $D_K = X_{K-1:K-1}$. We'll refer to this as a *generalized broken stick process*. By Shao and Hahn [3] Corollary 3.6, we have

$$\frac{1}{K} \sum_{i=1}^{K} \ln(KD_i) \xrightarrow{a.s.} - \int_0^1 h(x) \ln h(x) dx - \lambda, \tag{4}$$

as $K \to \infty$ and where *a.s.* signifies *almost sure convergence*, $\lambda = 0.577\ldots$ is the Euler constant and $-\int_0^1 h(x) \ln h(x) dx$ is the differential entropy of $h(x)$. Clearly,

$$\frac{1}{K} \sum_{i=1}^{K} \ln(KD_i) = \ln K + \frac{1}{K} \sum_{i=1}^{K} \ln D_i, \tag{5}$$

so dividing through by $\ln K$ gives

$$\frac{\ln K}{\ln K} + \frac{1}{K} \frac{\sum_{i=1}^{K} \ln D_i}{\ln K}$$
$$\xrightarrow{a.s.} \frac{- \int_0^1 h(x) \ln h(x) dx}{\ln K} - \frac{\lambda}{\ln K}, \tag{6}$$

as $K \to \infty$. The right side of the limit in (6) goes to 0 because $-\int_0^1 h(x) \ln h(x)dx / \ln K$ goes to 0 by the boundedness of the density $h(x)$. Expressing logarithms with a radix $= K$ then leads to the limit

$$\frac{\sum_{i=1}^K \log_K D_i}{K} \xrightarrow{a.s} -1 \text{ as } K \to \infty. \tag{7}$$

Our *universality property* for $\beta$ will now follow almost immediately from this. We use sample spacings to populate the $K$ letter probabilities for the monkey keyboard. Since $s$ is the probability of the space character, define $q_i = (1-s)D_i$ $(1 \leq i \leq K)$ so that $\sum_{i=1}^K q_i = 1-s$. Let $\overline{m}_K = \sum_{i=1}^K \log_K q_i / K$. Then from the limit (7), we have

$$\begin{aligned}
\overline{m}_K &= \sum_{i=1}^K \frac{\log_K q_i}{K} = \sum_{i=1}^K \frac{\log_K((1-s)D_i)}{K} \\
&= \log_K(1-s) + \frac{\sum_{i=1}^K \log_K D_i}{K} \xrightarrow{a.s} -1
\end{aligned} \tag{8}$$

as $K \to \infty$. Observe that $\log_K(1-s) = \ln(1-s) / \ln K \to 0$. Since $\overline{m}_K \xrightarrow{a.s} -1$ from (8) and since $-\beta < -1$, showing that $\overline{m}_K \leq -\beta$ will prove the argument that $-\beta \approx -1$ for finite but sufficiently large $K$. To see that this is the case, note:

$$\sum_{i=1}^K \frac{\log_K q_i^{1/\beta}}{K} = \log_K \left[ \left( q_1^{1/\beta} q_2^{1/\beta} \cdots q_K^{1/\beta} \right)^{1/K} \right] \tag{9}$$

$$\leq \log_K \left( \frac{q_1^{1/\beta} + q_2^{1/\beta} \cdots + q_K^{1/\beta}}{K} \right) \tag{10}$$

$$= \log_K \frac{1}{K} = -1, \tag{11}$$

where the inequality in Equation (10) follows from the geometric-arithmetic mean inequality. Therefore, from

$$\sum_{i=1}^K \frac{\log_K q_i^{1/\beta}}{K} = \frac{1}{\beta} \sum_{i=1}^K \frac{\log_K q_i}{K} = \frac{1}{\beta} \overline{m}_K \leq -1, \tag{12}$$

it has now been established that $\overline{m}_K \leq -\beta$. In the special case of Miller's model using all equal letter probabilities, $\overline{m}_K = -\beta$.

We want to complement this analytic result for $-\beta$ using a natural and insightful approximation Prof. David Mumford kindly brought to our attention after looking at a preprint of this article. Consider the function $p(x) = \sum_{i=1}^K q_i^{1/x}$ for $x > 0$. Then $p'(x) = -\sum_{i=1}^K q^{1/x} \ln q_i / x^2$ and a first order approximation from a Taylor expansion about $x = 1$ gives $p(x) \approx p(1) + (x-1)p'(1) = 1 - s + (x-1) \left( -\sum_{i=1}^K q_i \ln q_i \right)$. Since $\beta$ satisfies $p(\beta) = 1$, it follows that $1 = p(\beta) \approx 1 - s + (\beta-1)(-\sum_{i=1}^K q_i \ln q_i)$, giving $\beta \approx 1 + s/(-\sum_{i=1}^K q_i \ln q_i)$. If the term $-\sum_{i=1}^K q_i \ln q_i$ is sufficiently large, then $\beta \approx 1$. For the class of random variables that we considered in partitioning $[0, 1-s]$ to populate the letter probabilities, this will be the case for large $K$.

Figure 2 presents graphical results using simulations of sample spacings from several distributions to illustrate our result for $\beta$. In Figure 2a we plot $\log_{10} P_r$ by $\log_{10} r$ for the Miller model with exactly the parameter values he used, *i.e.*, $K = 26$ letters, a space probability $s = 0.18$ and equal letter probabilities $q_1 = \cdots = q_{26} = (1-0.18)/26$. The graph is based on the ranks $1 \leq r \leq 475,255 = \sum_{i=0}^4 26^j$, corresponding to all words of length $\leq 4$ non-space letters. Figure 2b through 2d are based on generating $K = 26$ letter probabilities from three different continuous distributions with bounded densities $h(x)$ on $[0, 1]$ using a generalized broken stick method to obtain the sample spacings. Again $s = 0.18$ was used for the probability of the space character and the letter probabilities were populated with values from the spacings with $q_i = (1-s)D_i$, $1 \leq i \leq 26$, so that in each case, $\sum_{i=1}^{26} q_i = 0.82$, as in Miller's example. For each graph in Figure 2b–d, we generated the
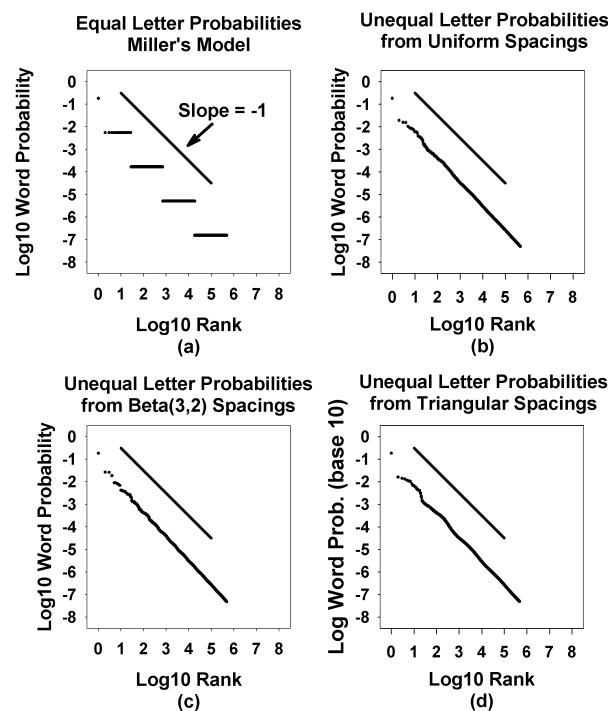
largest 475,255 word probability values in order to match the Miller example of Figure 2a. The three continuous distributions, all defined on $[0, 1]$, are:

- a uniform distribution with density $h(x) = 1$;
- a beta $B(3, 2)$ distribution with density

$$h(x) = \frac{\Gamma(3+2)}{\Gamma(3)\Gamma(2)} x^{3-1}(1-x)^{2-1};\tag{13}$$

- a triangular distribution with density

$$h(x) = \begin{cases} 4x, & \text{if } 0 \le 1/2 \\ 4(1-x) & \text{if } 1/2 \le x \le 1. \end{cases}\tag{14}$$



**Figure 2.** Log-log plots of monkey word probabilities by rank showing the asymptotic tendency towards a $-1$ exponent (slope on the log-log scales) using four different distributions to generate letter probabilities: equal probabilities in (**a**) and a generalized broken stick process in (**b–d**). $K = 26$ letters were used in all cases and the largest $\sum_{i=0}^{4} 26^i = 475,255$ word probabilities are displayed.

The graphs in Figure 2b–d illustrate our theoretical derivation, but we also note that the very linear plots indicate what appears to be an almost "immediate" convergence to approximate power law behavior exceeding what might be expected from the asymptotics of Conrad-Mitzenmacher and Bochkarov-Lerner.

## 3. Monkey Twitter: Anscombe's CLT for the Model with a Finite Word Length Cutoff

The discussion of the finite word length version of the monkey model in [2] was incomplete because it did not explicitly provide the normalizing constants for the application of Anscombe's CLT for a random number of i.i.d. variables, as we do here. We also want to explain more clearly the nature of the hybrid Zipf-lognormal distribution that results from a finite length cutoff when letter probabilities are not identical.

The focus in Section 2 on the infinite ensemble of word probabilities $\mathbf{P}_\infty$ *ranked* to give $P_1 > P_2 \ge P_3 \ldots$ has actually obscured the significant *hierarchical structure* of the monkey

model—what Mandelbrot (p. 345, [18]), called a *lexicographic tree*—which only becomes evident when word length is considered. To see this, write $\mathbf{P}_{\text{len}=n}$ for the multiset of all word probabilities for monkey words of exactly length $n$. There are $K^n$ probabilities in $\mathbf{P}_{\text{len}=n}$ having a total sum of $(q_1 + q_2 + \cdots + q_K)^n s = (1 - s)^n s$ and an average value of $((1 - s)/K)^n s$, declining geometrically with $n$. Now taking word length into account, Mandelbrot's lexicographic tree structure becomes clear with the simple case of $K = 2$ letters: the root node of the tree has a probability of $s$ (for the "space" word of length 0); it has two branches to the next level consisting of nodes for words of length 1 with probabilities $q_1 s$ and $q_2 s$; these each branch out to the next level with four nodes having probability values $q_1^2 s, q_1 q_2 s, q_2 q_1 s, q_2^2 s$ and so on.

The formation of word probabilities as the product of letter probabilities immediately suggests lognormal structure (normal on a log scale). A simple way to exhibit the underlying approximate lognormal character of the finite multisets of word probabilities for all words of exactly length $n$ and those of length $\leq n$ is to construct their natural product probability spaces using the basic counting measure for letter probability values. For this analysis it makes no difference how the letter probabilities $q_i$ have been specified. We only require that they are positive values and at least two of them are different from each other.

The relevance of the CLT is first seen by representing any probability $P \in \mathbf{P}_{\text{len}=n}$ as a product of i.i.d. random variables times the constant $s$: $P = X_1 X_2 \ldots X_n s$, where each $X_i$ takes on one of the letter probability values $q_1, q_2, \ldots, q_K$ with probability $1/K$ (*i.e.*, we use the natural counting measure to construct a probability space on $\mathbf{P}_{\text{len}=n}$). Let $\mu_1 = \sum_{i=1}^K \ln q_i / K$ and $\sigma_1^2 = \sum_{i=1}^K (\ln q_i - \mu_1)^2 / K$. Assume that $\sigma_1^2 > 0$, *i.e.*, the letter probabilities are not all equal. Then $\mu_1 + \ln s$ is the mean and $\sigma_1^2$ is the variance of $\ln P$ for $P \in \mathbf{P}_{\text{len}=1}$, and it follows that $n\mu_1 + \ln s$ and $n\sigma_1^2$ are the respective mean and variance of $\ln P = S_n = \sum_{i=1}^n \ln X_i + \ln s$ for $P \in \mathbf{P}_{\text{len}=n}$. Therefore, $(S_n - n\mu_1)/\sqrt{n\sigma_1^2}$ is asymptotically normally distributed $N(0, 1)$ (the term $\ln s$ is rendered negligible in the asymptics) so that for sufficiently large $n$, $P$ itself will be approximately lognormal $LN(n\mu_1, n\sigma_1^2)$. This approximate normality for the log-probabilities of words of fixed length is quite obvious and was noted in passing by Mandelbrot (p. 210, [19]). However, he missed a *stronger* observation that the $N_n = \sum_{i=0}^n K^i$ word probabilities for $P \in \mathbf{P}_{\text{len}\leq n} = \cup_{i=0}^n \mathbf{P}_{\text{len}=i}$ have a distribution that behaves in its central part away from the tails very much like $P \in \mathbf{P}_{\text{len}=n}$. That is, a version of the CLT can be applied to words of length $\leq n$, not just to words of length $n$, and the two distributions, *in one sense*, are very close to each other.

To show this, we make use of CLT results that pertain to a *random number* of i.i.d. random variables. In the preceding paragraph, we explained how for $P \in \mathbf{P}_{\text{len}=n}$, $\ln P$ could be regarded as the sum $S_n$ of a *fixed* number of i.i.d. random variables (plus a constant). Now we will apply an extended version of this same logic to the multiset of all words of length $\leq n$. For $P \in \mathbf{P}_{\text{len}\leq n}$, there are $K^j$ word of length $j$, $0 \leq j \leq n$. Think of $\ln P$ as represented by a sum $\ln X_1 + \ln X_2 + \cdots + \ln X_{R_n} + \ln s$, where $0 \leq R_n \leq n$ is the number of letters in the word. As before, we regard the $X_i$ as random variables taking on the values $q_1, q_2, \ldots, q_K$, each with probability $1/K$. Now we will *also* let $R_n$ be a random variable assuming the values $0, 1, \ldots, n$. Define the distribution of $R_n$ in terms of the natural counting measure for word length: since there are $K^j$ words of length $j$ and a total of $N_n = \sum_{j=0}^n K^i$ words in all, then $\text{Prob}\{R_n = j\} = K^j / N_n$, $0 \leq j \leq n$.

We first state Anscombe's generalization of the CLT (Theorem 3.1, p.15, [20]) before applying it. Let $T_n = \sum_{i=1}^n Y_i$ be the sum of $n$ i.i.d. random variables with mean 0 and variance $\sigma^2$, so that $\frac{T_n}{\sigma\sqrt{n}}$ converges in distribution to $N(0, 1)$. Let $\nu_n$ be an integer random variable such that $\nu_n/n \to \theta$ in probability as $n \to \infty$ ($0 < \theta < \infty$). Then by Anscombe the sum $T_\nu = \sum_{i=1}^\nu Y_i$ normalized as $\frac{T_\nu}{\sigma\sqrt{n\theta}}$ also converges in distribution to $N(0, 1)$, just as $T_n$ does. Note the presence of $\theta$ in the denominator in the case for $T_\nu$.

It is now clear how to apply this theorem to $S_{R_n}$ as we have constructed it. We will not repeat the calculations of [2] here, but it is easily shown that $R_n/n \to \theta = 1$ in probability as $n \to \infty$. Because this limiting constant $\theta$ is 1, Anscombe's generalization can be applied with the *identical* normalizing

constants $n\mu_1$ and $\sqrt{n\sigma_1^2}$ as used above with $\ln P$ for $P \in \mathbf{P}_{\text{len}=n}$. Consequently, it is *also* true that for $P \in \mathbf{P}_{\text{len}\leq n}$, the normalized sum $(S_{R_n} - n\mu_1)/\sqrt{n\sigma_1^2}$ has an asymptotic $N(0,1)$ distribution. In other words the two random sums $S_n$ and $S_{R_n}$ behave so similarly in their centers that the word probabilities in $\mathbf{P}_{\text{len}=n}$ and $\mathbf{P}_{\text{len}\leq n}$ are *both* approximately $LN(n\mu_1, n\sigma_1^2)$. (Again, the constant $\ln s$ can be ignored in dealing with the asymptotics for both $S_n$ and $S_{R_n}$.) However, it is essential to remark that the two distributions have very different upper tail behavior. In this regard, Le Cam's [21] comment that French mathematicians use the term "central" referring to the CLT "because it refers to the center of the distribution as opposed to its tails" is particularly relevant.

The behavior of $P \in \mathbf{P}_{\text{len}\leq n}$ is illustrated in the graphs of Figure 3a,b. The plot in Figure 3a was generated just as in Figure 2b except that a finite length cutoff of $n \leq 4$ letters has been applied. To make this clear, in Figure 2b the plot shows the top 475,255 word probabilities in $\mathbf{P}_\infty$ generated using $K = 26$ letters derived from uniform spacings. In contrast, using the same letter probabilities, the plot in Figure 3a shows the 475,255 word probabilities generated with word length $\leq 4$ letters, *i.e.*, all the values of $P \in \mathbf{P}_{\text{len}\leq 4}$. Word length in the case of *unequal* letter probabilities is certainly correlated, *but not perfectly*, with word probability: words of shorter length will, *on average*, have a higher probability than those of longer length, but except in Miller's degenerate case, there will always be reversals where a longer word will have a higher probability than a shorter word. In natural languages, Zipf [5] underscored that "the length of a word tends to bear an inverse relationship to its relative frequency", which he called the *Law of Abbreviation*. In many respects, this is the starting point for his *Principle of Least Effort*. However, writing $P_{1:N_n} > P_{2:N_n} \cdots \geq P_{N_n:N_n}$ for the ranked values of $\mathbf{P}_{\text{len}\leq n}$, it should be evident that for any rank $r$, $P_{r:N_n} \leq P_r$ and that $P_{r:N_n} = P_r$ when $n$ is sufficiently large. In short, $\mathbf{P}_{\text{len}\leq n}$ inherits its upper tail approximate power law behavior from $\mathbf{P}_\infty$, which is illustrated by comparing the top part of the curve in Figure 3a with the corresponding part of Figure 2b.
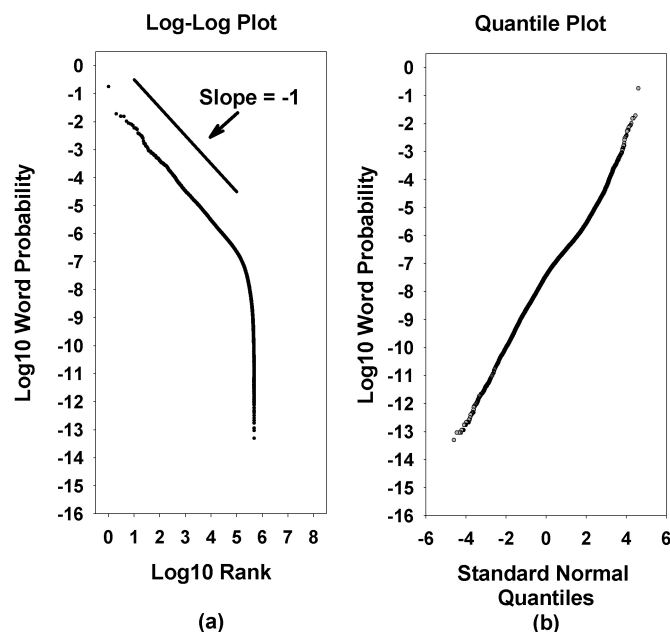


**Figure 3.** Figure 3a is a log-log plot of monkey word probabilities by rank for all words of length $\leq 4$ non-space characters using letter probabilities from uniform spacings. The linear upper tail coincides closely with the previous Figure 2b, but the approximate power law clearly breaks down. Figure 3b shows the same word probabilities in a standard normal quantile plot. Its rough linearity confirms an approximate Gaussian fit over the whole distribution even though the upper tail is an approximate power law as seen on the left in Figure 3a. We refer to this distribution as a Zipf-lognormal hybrid, and it has many connections to other distributions discussed in the statistical literature.

Figure 3b uses the same data points from Figure 3a graphed as a normal quantile plot, and its roughly linear appearance for the logarithm of word probabilities for $P \in \mathbf{P}_{\mathrm{len} \leq \mathrm{n}}$ conforms to an approximate Guassian fit, although certainly the bending in the upper half of the distribution departs a bit from the linear trend of the lower half. The fact that a distribution can have a power law tail and lognormal central part and *still look lognormal over essentially its entire range* may seem surprising. The discussion in [22] about *power law mimicry* in the upper tail of lognormal distributions helps to explain why this can happen. We will call the distribution for $\mathbf{P}_{\mathrm{len} \leq \mathrm{n}}$ a Zipf-lognormal distribution.

The monkey model with a fixed word length cutoff can prove useful as a motivating idea. In the next section, we will discuss how models with the same branching tree structure have been proposed many times in the past, typically in settings where something like a finite word length is natural to consider. For the moment, think of the social networking service, Twitter, which allows members to exchange messages limited to at most 140 characters. Define *Monkey Twitter* with a finite limit of $n + 1$ characters. For convenience, in any implementation of a Monkey Twitter random experiment we will assume that monkeys would always fill up their allotted message space of $n + 1$ characters. Monkey words still require a terminating space character, so it is possible for a monkey to type (at most) one "non-word", which can vary in length from 1 to $n + 1$ characters, and will always be the last part of a message string. Non-words are discarded, and the probabilities for the legitimate monkey words, varying in length from 0 to $n$ non-space characters plus a terminating space character, will correspond to the values in $\mathbf{P}_{\mathrm{len} \leq \mathrm{n}}$.

## 4. Connections to Other Work

The subject of power law distributions is a vast topic of research reaching across diverse scientific disciplines [17,22–24]. Here we provide a brief sketch of how our results connect to some other work and how they can be considered in a much more general light.

The monkey model viewed in terms of its hierarchical tree structure is the starting point for understanding its general nature. Though Miller did not explicitly present his model as a branching tree, several researchers at almost the same time in the mid-to-late 1950s were highlighting this form using the equivalent of equal letter probabilities to motivate the occurrence of empirical power law distribution in other areas. For example, Beckman [25] introduced this idea in connection with the power law of city populations within a country ("Auerbach's law" [26]). Using essentially the same logic as Miller, but in a completely different setting, Beckman assumed that a given community will have $K$ satellite communities, each with a constant decreasing fraction of the population at a higher level. That is, if $A$ is the population of the largest city and $0 < p < 1$, there will be $K$ nearby satellite communities with population $pA$, $K^2$ smaller communities nearby to these with population $p^2 A$, and so on. Mandelbrot (p. 226, [27]) had an apt expression for this pattern: he described it as "compensation between two exponentials" because the number of observations at each level increases geometrically while the mean value decreases geometrically. Beckman then went on to note that if instead of using a constant decreasing fraction $p$, one used a random variable $X$ on $(0, 1)$, the population at the $n^{\mathrm{th}}$ level down would be a random variable of the form $X_1 X_2 \cdots X_n A$, leading to an approximate lognormal distribution of populations at this level. This corresponds to our discussion of monkey word probabilities in $\mathbf{P}_{\mathrm{len} = \mathrm{n}}$, although Beckman was not aware of the still stronger statement of lognormality for the probabilities of words of length $\leq n$ in $\mathbf{P}_{\mathrm{len} \leq \mathrm{n}}$ that we demonstrated using Anscombe's CLT.

Many more examples of a branching tree structure essentially equivalent to Miller's monkey model with equiprobable letters have been proposed over the years to motivate the occurrence of a huge variety of empirical power law distributions, including such size distributions as lake areas, island areas ("Korcak's law"), river lengths, etc. Indeed, Mandelbrot's [18] classic book on fractals is a rich source of these. This equiprobability case is so simple to analyze that it has been invoked over and over again to illustrate how power law behavior can arise. However, the more complicated case using unequal letter probabilities (or proportions) *and a finite word length cutoff* is what really

underscores the close analogy between the monkey Zipf-lognormal distribution and other mixture models exhibiting hybrid power law tails and approximately lognormal central ranges.

Empirical distributions of this form have appeared with great frequency. In fact there is a long history of researchers, including Pareto, first discovering what appears to be an empirical power law over an entire distribution because they start out by looking at only the largest observations (cities, corporations, islands, *etc.*). However, when they extend their measurements to the part of the distribution below the upper tail, *which is always more difficult to observe*, the power law behavior typically breaks down. Several examples of this are chronicled in [22]. The power law for city sizes noted above is a perfect illustration. Auerbach in 1913 [26] looked at the 94 largest German cities from a 1910 census and showed a good power law fit, but because he did not include the thousands of smaller communities, he may not have been aware that this relationship breaks down. Modern studies of the *full* distribution of communities, such as Eeckhout's [28] analysis of the populations of 25,359 places from U.S. Census 2000, prove with high confidence that the *bulk* of the community populations fit an approximate lognormal distribution and that the power law behavior is confined to the upper tail.

Montroll and Shlesinger [29,30] explained how to generate a Pareto-lognormal hybrid for income distributions by using a hierarchical mixture model of lognormal distributions. This is motivated from several angles, including the notion that higher classes amplify their income by organizing in such a way as to benefit from the efforts of others. Reed and Hughes [31] have provided a far-ranging framework for understanding these hybrid distributions across the entire spectrum of disciplines where they have been discovered: "physics, biology, geography, economics, insurance, lexicography, internet ecology, *etc.*" In [31] he and Hughes give a condensed summary 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 work encompasses: (1) geometric Brownian motion (GBM); (2) discrete multiplicative processes; (3) homogeneous birth and death processes; and (4) Galton–Watson branching processes. In one of many variations on this theme, in his GBM model [32] Reed specifies a stochastic process that uses lognormal distributions varying continuously in time with an exponential mixing distribution and shows that this generates a "Double Pareto-Lognormal Distribution". This has an asymptotic Pareto upper tail and a central part approximately lognormal; in addition, it exhibits interesting asymptotic behavior in the it lower tail where it is characterized by a direct (rather than an inverse) power law. Reed has gone to great effort to demonstrate the high quality of the fit of this distribution for all ranges of values (low, middle, high) across a wide variety of size distributions such as particle and oil field sizes [32], incomes [33], internet file sizes and the sizes of biological general [31].

In today's nomenclature, the term *Zipf's law* has come to mean any empirical distribution with an approximate power law form and an exponent close to the value $-1$, not just the word frequency law. (To be clear here, when we refer to power law distributions, we also mean the hybrids with power law tails that we have been considering.) The subset of power laws with this restricted exponent value is surprisingly large and includes the distribution of firm sizes [34], city sizes [35], the famous Gutenberg–Richter law for earthquake magnitudes [36] and many others.

Gabaix [35] pointed out that while it has long been known that stochastic growth processes could generate power laws, "these studies stopped short of explaining why the Zipf exponent should be 1." To address this question, he has given a theoretical explanation of the genesis of a $-1$ exponent for the Zipf law for city populations, but in fact, his approach can be regarded more generally. His key idea is a variation of Gibrat's [37] *Law of Proportion Effect*: using a fixed number of normalized city population quantities that sum to 1 and assuming growth processes (expressed as percentages) randomly distributed as i.i.d. variables with a common mean and variance, he proves that a steady state will satisfy Zipf's law. This proof requires an additional strong assumption in order to reach a steady state, namely, a lower bound for the (normalized) population size. Gabaix goes on to discuss relaxed versions of his model and how it fits into the larger context of similar work done by others.

In their monograph on Zipf's law Saichev *et al.* [38] modify and extend the core Gabaix idea in numerous ways that render it more realistic. For concreteness, they present their work using the terminology of financial markets and corporate asset values, but they are very clear on how their models are relevant to a broad range of "physical, biological, sociological and other applications". As with Reed, GBM plays an important role in their investigations, which incorporate birth and death processes, acquisitions and mergers, the subtleties of finite-size effects and other features. Notably, they focus on the sets of conditions that lead to an approximate $-1$ exponent.

In both Harrëmoes and Topsøe [39] and Corominas-Murtra and Solé [40] $-1$ emerges as a consequence of certain information theoretic ideas pertaining to the entropy function under growth assumptions. In [39] the focus is on language and the expansion of vocabulary size while simultaneously maintaining finite entropy. The perspective in [40] is broader, but still based on the idea of systems with an expanding number of possible states "characterized by special features on the behavior of the entropy".

There is another topic area of statistical research that impacts on our work. Natural language word frequency distributions have been characterized as *Large Number of Rare Events* (LNRE) distributions because when collections of text are examined, no matter how big, there are always a large number of words that occur very infrequently in the sample [41]. LNRE behavior indicates a tip-of-the-iceberg phenomenon resulting from a sampling bias that captures the most common words, but necessarily misses a vast quantity of rarely used words. This is a classic and much studied problem encountered in species abundance surveys in ecological research, where a key question becomes estimating the size of the *zero abundance class*—the typically great number of species not observed in a sample because of their rarity [42].

To get an idea of the significance of this issue, consider that Figure 3a graphs the parent distribution of word probabilities for the Zipf-lognormal distribution, and not the *sample* frequencies as would be obtained from actually carrying out the Monkey Twitter experiment. Simple visual inspection of Figure 3a indicates that the linear part of the graph (*i.e.*, the power law tail) holds for about 5 logarithmic decades or about the first 100,000 word probabilities out of the total of 475,255 probabilities plotted there. However, these 100,000 probabilities comprise 97.4% of the total probability mass of all 475,255 values. Intuitively, it should be clear that sampling from this Zipf-lognormal population distribution will be dominated by the Zipf part, not the lognormal part. Unless the sample size was astronomically large, so that the large number of low probability words showed up, the underlying structure of the parent distribution would not reveal itself. To carry this idea still further, imagine the situation if the word length cutoff was on the order of $n = 140$ characters, such as with real Twitter. No experiment could be run within any realistic time frame to ever hope to obtain a sample sufficiently large to uncover the true hybrid character of the population distribution—the sample would always appear as a Zipf law, not a Zipf-lognormal law. This kind of *visibility bias* has been a constant and recurring theme in all areas where Pareto-Zipf type distributions have been studied [2,22]. We believe its significance has been poorly appreciated in relation to this topic.

Indeed, there are many challenging questions in connection with the sampling properties of variations of the monkey model and how they compare to natural language text. Ferrer-i-Concho and Gavaldà [43] show that even using the simple Miller model, there is a need for careful analysis to properly evaluate the deviations between actual frequency histograms from finite samples and their theoretical expected values. The analysis in [44] demonstrates that the monkey model of independently combined characters, even when the number of characters and their probabilities are populated with values corresponding to specific real language target text samples, yields sample distributions that depart in systematic ways from that of the target text. Bernhardsson *et al.* [45] and Yan and Minnhagen [46] have highlighted how real language corpora have frequency distributions whose characteristics vary with the size of the sample and do not correspond to sampling properties of the monkey model. As just one example, in [46] it is argued that an approximate power law

form such as Zipf's law cannot account for the fact that the scaling exponent typically increases with increasing sample sizes in real language. Their Figure 2d shows the clear fluctuations in the exponent as a function of text length.

We conclude by emphasizing how the monkey monkey fits into a much larger conceptual scheme than appears at first glance. For example, by viewing the model as a branching tree—Mandelbrot's lexicographic tree—the self-similar character of its construction becomes clear. In fact, the model can be construed as a fractal object with a fractal dimension of $1/\beta$, where $-\beta < -1$ is the exponent studied in Section 2 ([18,47]). Our application of random spacings is *also* very much in the spirit of the enormously fruitful study of *random matrix theory* and a class of stochastic systems referred to as the *KPZ universality class*. Along these lines Borodin and Gorin [48] have discussed a variety of probabilistic systems "that can be analyzed by essentially algebraic method", yet are applicable to a broad array of topics. Miller's model, analyzable with simple algebra, appears to us as an elementary example perhaps falling into this category.

## 5. Conclusions

Sample spacings provide a natural way to populate the letter probabilities in the monkey model through a random division of the unit interval. The Shao and Hahn asymptotic limit law for the logarithms of spacings then leads to the result that the exponent in the approximate power law of word frequencies will tend towards a $-1$ value under broad conditions as the number of letters in the alphabet increases. A first order approximation for $\beta$ provides a good intuitive understanding of this behavior. The monkey model can also be viewed as a branching tree structure. In that light, Anscombe's CLT reveals an underlying lognormal central part of the word frequency distribution when a finite word length cutoff is imposed. The resulting hybrid Zipf-lognormal distribution has many connections to other work. The visibility bias inherent in sampling from this distribution is similar to what has been noted as a characteristic of many different empirical power laws.

**Author Contributions:** The data analysis and graphics were done by Richard Perline. Both authors jointly wrote the paper and have approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Tao, T. E pluribus unum: From complexity, universality. In *The Best Writing on Mathematics 2013*; Pitici, M., Ed.; Princeton University Press: Princeton, NJ, USA, 2014; pp. 32–46.
2. Perline, R. Zipf's law, the central limit theorem and the random division of the unit inteval. *Phys. Rev. E* **1996**, *54*, 220–223.
3. Shao, Y.; Hahn, M.G. Limit theorems for the logarithm of sample spacings. *Stat. Probab. Lett.* **1995**, *24*, 121–132.
4. Perline, R. The random division of the unit interval and the approximate $-1$ exponent in the monkey-at-the-typewriter model of Zipf's law. *Stat. Probab. Lett.* **2015**, submitted.
5. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Cambridge, MA, USA, 1949.
6. Bell, T.C.; Cleary, J.G.; Witten, I.H. *Text Compression*; Prentice Hall: Bergen County, NJ, USA, 1990.
7. Good, I.J. Statistics of language: Introduction. In *Encyclopedia of Linguistics, Information and Control*; Meetham, A.R., Ed.; Pergamon Press: New York, NY, USA, 1969; pp. 567–581.
8. Hart, M.S. Project Gutenberg. Available online: http://www.gutenberg.org/ (accessed on 6 June 2014).
9. Mandelbrot, B.B. On recurrent noise limiting coding. In *Information Networks, the Brooklyn Polytechnic Institute Symposium*; Weber, E., Ed.; Interscience: New York, NY, USA, 1955; pp. 205–221.

10. Conrad, B.; Mitzenmacher, M. Power laws for monkeys typing randomly: The case of unequal letter probabilities. *IEEE Trans. Inf. Theory* **2004**, *50*, 1403–1414.

11. Bochkarev, V.V.; Lerner, E.Y. The Zipf law for random texts with unequal probabilities of occurrence of letters and the Pascal pyramid. *Russ. Math.* **2012**, *56*, 25–27

12. Bochkarev, V.V.; Lerner, E.Y. Strong power and subexponential laws for an ordered list of trajectories of a Markov chain. *Electron. J. Linear Algebra* **2014**, *27*, 534–556.

13. Bochkarev, V.V.; Lerner, E.Y. Zipf exponent of trajectory distribution in the hidden Markov model. *J. Phys. Conf. Ser.* **2014**, *490*, 012008.

14. Edwards, R.; Foxall, E.; Perkins, T.J. Scaling properties of paths on graphs. *Electron. J. Linear Algebra* **2012**, *23*, 966–988.

15. Miller, G.A. Some effects of intermittent silence. *Am. J. Psychiatry* **1957**, *70*, 311–314.

16. Miller, G.A.; Chomsky, N. Finitary Models of Language Users. In *Handbook of Mathematical Psychology*; Luce, R.D., Bush, R.R., Galanter, E., Eds.; Wiley: New York, NY, USA, 1963; Volume 2, pp. 419–491.

17. Mitzenmacher, M. A brief history of generative models for power law and lognormal distributions. *Internet Math.* **2003**, *1*, 226–251.

18. Mandelbrot, B.B. *The Fractal Geometry of Nature*; W.H. Freeman and Company: New York, NY, USA, 1983.

19. Mandelbrot, B.B. On the theory of word frequencies and on related markovian models of discourse. In *Structure of Language and Its Mathematical Aspects: Proceedings of Symposia on Applied Mathematics Volume XII*; Jakobson, R., Ed.; American Mathematical Society: Providence, RI, USA, 1961; pp. 190–219.

20. Gut, A. *Stopped Random Walks: Limit Theorems and Applications*; Springer-Verlag: New York, NY, USA, 1988.

21. Le Cam, L. The central limit theorem around 1935. *Stat. Sci.* **1986**, *1*, 78–96.

22. Perline, R. Strong, weak and false inverse power laws. *Stat. Sci.* **2005**, *20*, 68–88.

23. Clauset, A.; Shalizi, C.R.; Newman, M.E. Power law distributions in empirical data. *SIAM Rev.* **2009**, *51*, 661–703.

24. Arnold, B. *Pareto Distributions*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2015.

25. Beckman, M.J. City hierarchies and the distribution of city sizes. *Econ. Dev. Cult. Chang.* **1958**, *6*, 243–248.

26. Auerbach, F. Das Gesetz der Bevölkerungskonzentration. *Petermanns Geogr. Mitteilungen*, **1913**, *59*, 74–76.

27. Mandelbrot, B.B. *Fractals and Scaling in Finance: Discontinuity, Concentration, Risk Selecta Volume E*; Springer: New York, NY, USA, 1997; pp. 219–251.

28. Eeckhout, J. Gibrat's law for (all) cities. *Am. Econ. Rev.* **2004**, *94*, 1429–1451.

29. Montroll, E.; Shlesinger, E. On $1/f$ noise and other distributions with long tails. *Proc. Natl. Acad. Sci. USA* **1982**, *79*, 3380–3383.

30. Montroll, E.; Shlesinger, E. Maximum entropy formalism, fractals, scaling phenomena, and $1/f$ noise: A tale of tails. *J. Stat. Phys.* **1983**, *32*, 209–230.

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

32. Reed, W.J. On Pareto's law and the determinants of Pareto exponents. *J. Income Distrib.* **2004**, *13*, 1–2.

33. Reed, W.J. The double Pareto-lognormal distribution—A new parametric model for size distributions. *Commun. Stat.* **2004**, *33*, 1733–1753.

34. Axtell, R.L. Zipf distribution of U.S. firm sizes. *Science* **2001**, *293*, 1818–1820.

35. Gabaix, X. Zipf's law for cites: An explanation. *Q. J. Econ.* **1999**, *114*, 739–767.

36. Kagan, Y.Y. Universality of the seismic moment-frequency relations. *Pure Appl. Geophys.* **1999**, *155*, 537–573.

37. Gibrat, R. *Les Inegalites Economiques*; Libraire du Recueil Sirey: Paris, France, 1931. (In French)

38. Saichev, A.; Malevergne, Y.; Sornette, D. *Theory of Zipf's Law and Beyond*; Springer-Verlag: Berlin, Germany, 2010.

39. Harrëmoes, P.; Topsøoe, F. Maximum entropy fundamentals. *Entropy* **2001**, *3*, 191–226.

40. Corominas-Murtra, B.; Solé, R. Universality of Zipf's law. *Phys. Rev. E* **2010**, *82*, 011102.

41. Baayen, R.H. *Word Frequency Distributions*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.

42. Bunge, J.; Fitzpatrick, M. Estimating the number of species: A review. *J. Am. Stat. Assoc.* **1993**, *88*, 364–373.

43. Ferrer-i-Concho, R.; Gavaldà, R. The frequency spectrum of finite samples from the intermittent silence process. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 837–843.

44. Ferrer-i-Cancho, R.; Elvevåg, B. Random texts do not exhibit the real Zipf's law-like rank distribution. *PLoS One* **2010**, *5*, e9411.

45. Bernhardsson, S.; Baek, S.K.; Minnhagen, P. A paradoxical property of the monkey book. *J. Stat. Mech. Theory Exp.* **2011**, *7*, P07013.

46. Yan, X.; Minnhagen, P. Randomness versus specifics for word-frequency distributions. *Phys. A* **2015**, *444*, 5828–5837.

47. Schroeder, M. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*; W.H. Freeman and Company: New York, NY, USA, 1991.

48. Borodin, A.; Gorin, V. Lectures on Integrable Probability. **2015**, arXiv:1212.3351v2.