A paradoxical property of the monkey book

J. Stat. Mech. (2011) P07013

(http://iopscience.iop.org/1742-5468/2011/07/P07013)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

A paradoxical property of the monkey book

Sebastian Bernhardsson, Seung Ki Baek and Petter Minnhagen

IceLab, Department of Physics, Umeå University, 901 87 Umeå, Sweden
E-mail: sebbeb@tp.umu.se, garuda@tp.umu.se and minnhagen@physics.umu.se

Received 14 March 2011
Accepted 29 June 2011
Published 19 July 2011

Online at stacks.iop.org/JSTAT/2011/P07013
doi:10.1088/1742-5468/2011/07/P07013

Abstract. A 'monkey book' is a book consisting of a random sequence of letters and blanks, where a group of letters surrounded by two blanks is defined as a word. We compare the statistics of the word distribution for a monkey book to real books. It is shown that the word distribution statistics for the monkey book is different and quite distinct from a typical real book. In particular, the monkey book obeys Heaps' power law to an extraordinarily good approximation, in contrast to the word distributions for real books, which deviate from Heaps' law in a characteristic way. This discrepancy is traced to the different properties of a 'spiked' distribution and its smooth envelope. The somewhat counter-intuitive conclusion is that a 'monkey book' obeys Heaps' power law precisely because its word-frequency distribution is *not* a smooth power law, contrary to the expectation based on simple mathematical arguments that if one is a power law, so is the other.

Keywords: analysis of algorithms, growth processes

J. Stat. Mech. (2011) P07013

1742-5468/11/P07013+12$33.00

Contents

1. Introduction

Words in a book occur with different frequencies. Common words like 'the' occur very frequently and constitute about 5% of the total number of written words in the book, whereas about half the different words only occur a single time [1]. The word-frequency $N(k)$ is defined as the number of words which occur k-times. The corresponding word-frequency distribution (wfd) is defined as $P(k) = N(k)/N$ where N is the total number of different words. Such a distribution is typically broad and is often called 'fat-tailed' and 'power-law' like. 'Power-law' like means that the large k-tail of the distribution to a reasonable approximation follows a power law, so that $P(k) \propto 1/k^{\gamma}$. Typically, one finds that $\gamma \leq 2$ for a real book [2]–[7]. What does this broad frequency distribution imply? Has it something to do with how the book is actually written? Or has it something to do with the evolution of the language itself? The fact that the wfd has a particular form was first associated with the empirical Zipf law for the corresponding word-rank distribution [5]–[7]. Zipf's law corresponds to $\gamma = 2$. Subsequently Simon proposed that the particular form of the wfd could be associated with a growth model, the Simon model, where the distribution of words was related to a particular stochastic way of writing a text from the beginning to the end [8]. However, a closer scrutiny of the Simon model reveals that the statistical properties implied by this model are fundamentally different from what is found in any real text [3]. Mandelbrot (at about the same time as Simon suggested his growth model) instead proposed that the language itself had evolved so as to optimize an information measure based on an estimated word cost (the more letters needed to build up a word the higher cost for the word) [9, 10]. Thus in this case the power law of the word distribution was proposed to be a reflection of an evolved property of the language itself. However, it was later pointed out by Miller in [11] that you do not need any particular language-evolution optimization to obtain a power law: a monkey randomly typing letters and blanks on a typewriter will also produce a wfd which is power-law like within a continuum approximation [11, 12]. The monkey book, hence, at least superficially, has properties in common with real books [12, 13, 15].

doi:10.1088/1742-5468/2011/07/P07013

J. Stat. Mech. (2011) P07013

The case that the relation to real books is just superficial has in particular been argued in [13].

In 1978, Heaps [16] presented another empirical law describing the relation between the number of different words, N, and the total number of words, M. Heaps' power law states that $N(M) \propto M^{\alpha}$, where α is a constant between zero and one. However, it was recently suggested that Heaps' law gives an inadequate description of this relation for real books, and that it needs to be modified so that the exponent α changes with the size of the book from $\alpha = 1$ for $M = 1$ to $\alpha = 0$ as $M \to \infty$ [2]. It was also shown that the wfd of real books, in general, can be better described by introducing an exponential cut off so that $P(K) = A\exp(-bk)k^{-\gamma}$ [3]. And, furthermore, the exponents γ and α are not independent. A simple mathematical derivation of their relation gives the result $\alpha = \gamma - 1$ [2]. This in turn means that the shape of the wfd also changes with the size of the book, so that $\gamma = 2$ for small M, but reaches the limit value $\gamma = 1$ as M goes to infinity [14]. The same analysis pointed out that the parameter b should be size dependent according to $b \approx b_0/M$ [2]. It was also shown empirically that the works of a single author follows the same $N(M)$-curve to a good approximation, which was further manifested in the meta-book concept: the $N(M)$-curve characterizing a text of an individual author is obtainable by pulling sections from the authors collective meta-book [2]. As will be further discussed below, the shape of the $N(M)$-curve is mathematically closely related to the random book transformation (RBT) [2, 3].

As mentioned above, the writing of a real book cannot be described by a growth model with history dependent elements because the statistical properties of a real book are translationally invariant [3]. This means that the statistical properties of the text are independent of where you are in the book, something which cannot be fulfilled by, e.g., the Simon model. The monkey book, on the other hand, is produced by a translational-invariant stationary process, and is thus in this respect more similar to real text. An important question of much interest is then how close the statistical properties of the monkey book really are to those of a real book. It is shown in the present work that in the context of Heaps' law the answer is somewhat paradoxical.

2. Monkey book

Imagine an alphabet with $\mathcal{A}$-letters and a typewriter with a keyboard with one key for each letter and a space bar. For a monkey randomly typing on the typewriter the chance of hitting the space bar is assumed to be q_s and the chance of hitting any of the letters is $(1 - q_s)/\mathcal{A}$. For simplicity we will here only consider the case where $q_s = 1/(\mathcal{A} + 1)$. A word is then defined as a sequence of letters surrounded by blanks. What is the resulting wfd for a text containing M words? Miller in [11] found that in the continuum limit this is in fact a power law. In the appendix we re-derive this result using an information cost method. A more standard alternative derivation can be found in [4].

We will denote the word-frequency distribution in the continuum limit by $p(k)$, and in the monkey book case it is given by

$$p(k) \propto \frac{1}{k^{\gamma}}, \tag{1}$$

J. Stat. Mech. (2011) P07013

which gives the probability to find a word with frequency k (the number of times the word appears in the text), and

$$\gamma = \frac{\ln[\mathcal{A}(\mathcal{A}+1)]}{\ln(\mathcal{A}+1)}. \tag{2}$$

Thus, $\gamma = 1$ if $\mathcal{A} = 1$ and $\gamma = 2$ in the infinite limit of $\mathcal{A}$.

2.1. Continuum approximation versus real word-frequency

The above result for $p(k)$ is an approximation of the actual (discrete) result expected from random typing. The true wfd, which would be obtained from the actual process of random typing, of the model will here be denoted as $P(k)$. What is then the relation between the power-law form of $p(k)$ and the actual probability, $P(k)$, for a word to occur k-times in the text? It is quite straightforward to let a computer take the place of a monkey and simulate monkey books [12]. Figure 1(a) gives an example for an alphabet with $\mathcal{A} = 4$ letters, and a total number of words $M = 10^6$. Such a book should have a power-law exponent of $\gamma \approx 1.86$ according to equation (2). Note that $P(k)$ for higher k consists of disjoint peaks: the peak with the highest k corresponds to the $\mathcal{A} = 4$ different one-letter words that can be formed, the next towards lower k to the $\mathcal{A}^2 = 16$ two-letter words, and so forth. Thus the power-law tail $1/k^\gamma$ in the case of a monkey book is not a smooth tail but a sequence of separated peaks as previously reported in [12, 13, 15]. The 'spikiness' of the wfd for a monkey book is a crucial feature, which leads to interesting consequences, as will be discussed below. So what is the relation to the continuum $p(k) \propto k^{-1.86}$? Plotted in log–log scale as in figure 1(a), $p(k)$ is just a straight line with the slope $-\gamma = -1.86$ (broken line in figure 1(a)). Represented in this way there is no obvious discernible relation between the separated peaks of $P(k)$ and the straight line given by $p(k)$. In order to directly see the connection one can instead compare the cumulative distributions $F(k) = \sum_{k'=k}^{M} P(k')$ and $f(k) = \sum_{k'=k}^{M} p(k') \propto 1/k^{0.86}$. In figure 1(b), $F(k)$ corresponds to the full drawn zig-zag curve and the straight broken line with slope -0.86 to the continuum approximation $f(k)$. In this plot the connection is more obvious: $f(k)$ is an envelope of $F(k)$. Figure 1(b) also illustrates that the envelope slope for the monkey book is independent of the length of the book: the two zig-zag curves correspond to $M = 10^5$ and 10^6. Both of them have the envelope slope $-\gamma = -0.86$ given by the continuum approximation $f(k)$.

To sum up: the continuum approximation $p(k) \propto 1/k^\gamma$ is very different from the actual spiked monkey book $P(k)$. However, the envelope, $f(k)$, for the cumulative wfd, $F(k)$, of the monkey book is nevertheless a power law with a slope which is independent of the size of the book.

3. Heaps' law

Heaps' law is an empirical law which states that the number of different words, N, in a book approximately increases as $N(M) \propto M^\alpha$ as a function of the total number of words, M [16], where $0 \leq \alpha \leq 1$. That is, if you read through a book, and for every new word record the number of different words you have encountered so far, you should, according to Heaps, obtain a curve described by a power law. For a random book, where the occurrences of a certain word are uniformly distributed throughout the book, like the

J. Stat. Mech. (2011) P07013

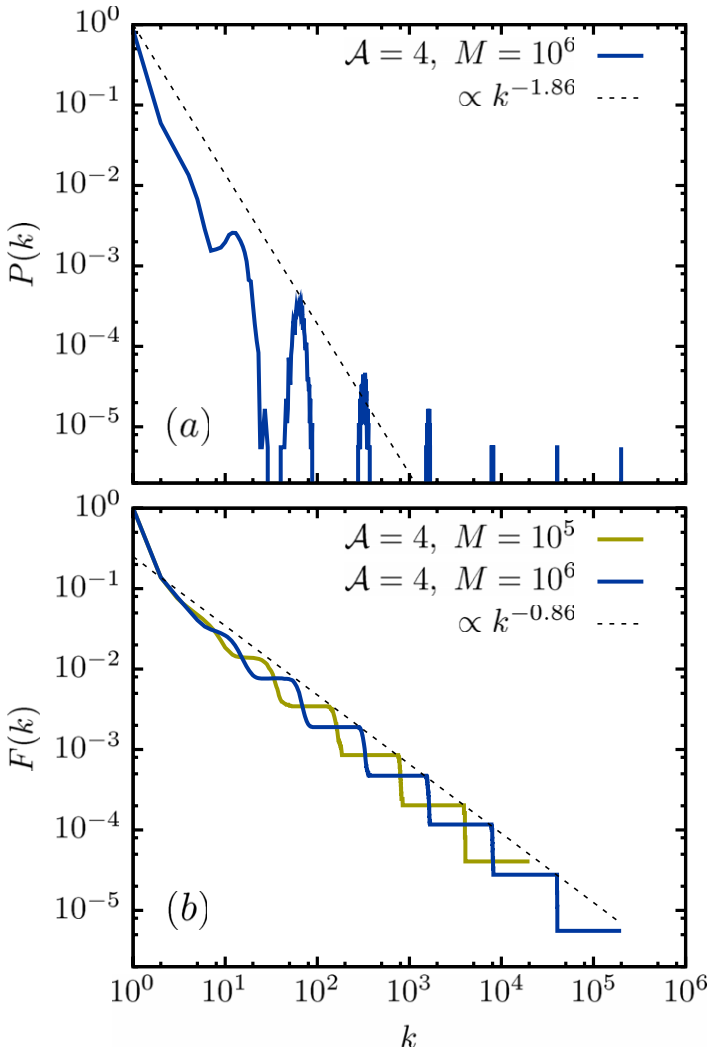

Figure 1. Word-frequency distribution for the monkey book. (a) The broken straight line corresponds to the continuum approximation $p(k) \propto k^{-\gamma}$ given by equation (2), whereas the full curve with disjoint peaks represents the real distribution $P(k)$. $P(k)$ and its continuum approximation $p(k)$ are clearly very different. (b) The corresponding cumulative distributions $f(k)$ and $F(k)$. The broken straight line corresponds to $f(k) \propto k^{-(\gamma-1)}$ and the blue zig-zag line to the corresponding real cumulative distribution $F(k)$. Note that $f(k)$ to good approximation is an envelope of the black zig-zag $F(k)$. The yellow zig-zag curve is the cumulative $F(k)$ for a tenth of the monkey book. Note that $f(k)$ still gives an equally good envelope. Thus the envelope of the cumulative $F(k)$ for a monkey book is a size-independent power law.

monkey book, there is a direct connection between $P(k)$ and the $N(M)$-curve. Suppose that such a book of size M has a wfd $P_M(k)$ created by sampling a fixed theoretical probability distribution $p(k) \propto k^{-\gamma}$, where the normalization constant is only weakly dependent on M. The number of different words for a given size is then related to M

J. Stat. Mech. (2011) P07013

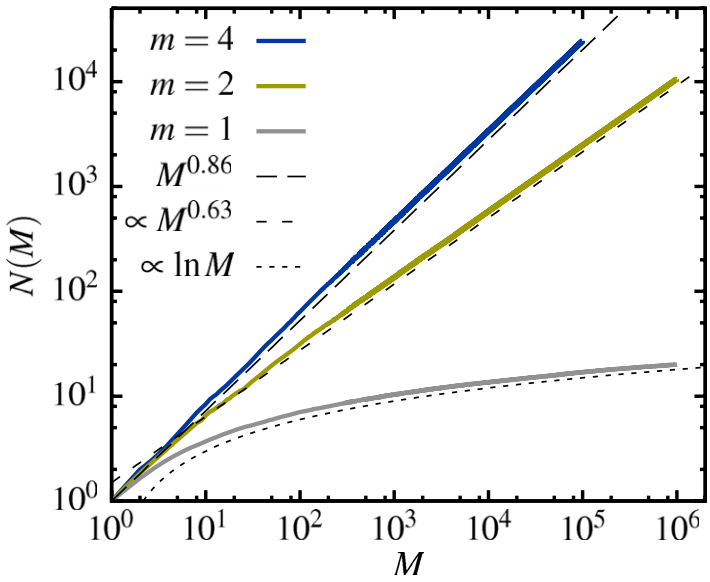

Figure 2. Heaps' law for monkey books with different sizes of the alphabet, in log–log scale. The full curves from top to bottom give the $N(M)$ for alphabets of lengths $m = 4, 2$, and 1, respectively. According to equation (7) the $N(M)$ should for $m = 4$ and 2 follow Heaps' power laws with the exponents 0.86 and 0.63, respectively, and the corresponding broken lines show that these predictions are borne out to excellent precision. For $m = 1$, equation (7) predicts that $N(M)$ instead should be proportional to $\ln M$, since $\gamma - 1 = 0$. The corresponding broken curve again shows an excellent agreement.

through the average frequency of a word, as

$$\langle k \rangle = \frac{M}{N(M)} = \sum_{k=1}^{M} kp(k). \tag{3}$$

And, since in the present case

$$\sum_{k=1}^{M} kp(k) \propto \frac{1}{2-\gamma}(M^{2-\gamma} - 1), \tag{4}$$

it follows that

$$N(M) \propto M^{\gamma-1}. \tag{5}$$

A heuristic direct way to this result is to argue that the first time for a word with frequency k to occur is inversely proportional to its frequency $\tau \propto 1/k$, so that in the time interval $[\tau, \tau + \mathrm{d}\tau]$ you introduce $n(\tau)\mathrm{d}\tau \propto (1/k^{\gamma})|\mathrm{d}k/\mathrm{d}\tau|\mathrm{d}\tau \propto \tau^{\gamma}\tau^{-2}\,\mathrm{d}\tau$ new words. Since τ is proportional to how far into the book you are, this means that $N \propto \int_0^M \tau^{\gamma-2}\,\mathrm{d}\tau \propto M^{\gamma-1}$. The conclusion from equations (3)–(5) is that the $N(M)$-curve of a random book with $P_M(k) \propto k^{-\gamma}$ should follow Heaps' law very precisely with the power-law index $\alpha = \gamma - 1$. It should be noted though that equations (3)–(5) are built on the assumption that $p(k)$ is independent of the size of the book, since only the upper limit of the summation is changing with M. However, figure 2 illustrates that this is indeed

true for monkey books by showing the $N(M)$-curve for different alphabet sizes (full drawn curves) together with the corresponding analytic solutions (broken curves). Note that, for Heaps' law, $N(M) \propto M^{\alpha}$, and the relationship $\alpha = \gamma - 1$ to hold, the full curves should be parallel to the broken curves for each alphabet size, respectively. Also, the continuum theory from equation (2) gives $\gamma = 1$ for $\mathcal{A} = 1$ (an alphabet with a single letter) which by equation (5) predicts $N \propto \ln M$, which is again in full agreement with the monkey book.

However, notwithstanding this excellent agreement, the reasoning is nevertheless flawed by a serious inconsistency: the connection to Heaps' law was here established for a random book with a continuous power-law wfd, whereas the wfd of a monkey book consists of a series of disjoint peaks. It thus seems reasonable that a random book with a wfd which is well described by a smooth power law would satisfy Heaps' law to an even greater extent. However, as we will see section 4, this reasoning is actually false.

4. Contradicting power laws

The most direct way to realize the problem is to start from a random book which has a smooth power-law wfd with an index γ. Such a book can be obtained by randomly sampling word frequencies from a continuous power-law distribution of a given γ and then placing them, separated by blanks, randomly on a line. For this 'sampled book' one can then directly obtain the $N(M)$-curve by reading through the book, as described earlier. Figure 3(a) gives an example of an $N(M)$-curve for a sampled book with $\gamma = 1.86$, $N = 10^5$ and $M = 10^6$. The resulting wfd is shown in figure 3(b).

It is immediately clear from figure 3(a) that a sampled book with a power-law wfd does not have an $N(M)$-curve which follows Heaps' law, $N(M) \propto M^{\alpha}$ (it deviates from the straight line in the figure). This is thus in contrast to the result of the derivation given by equations (3)–(5), and the monkey book which does obey Heaps' law with $\alpha = \gamma - 1$, as seen from figure 2. This must mean that the monkey book obeys Heaps' law because of its spiky form of the wfd. That is, we have derived the relation between the wfd and Heaps' law for a continuous wfd. Despite this fact, as we have just seen, this relation only holds for the monkey book, and not the sampled book which actually has a continuous wfd.

The core of this paradoxical behaviour lies in the fact that the derived form of the $N(M)$-curve requires a size-independent wfd, which according to figure 1(b) is true for the monkey book. However, it is not true for a sampled book, as seen from figure 3(b). The reason is that a random book is always subject to well-defined statistical properties. One of these properties is that the $P_M(k)$ transforms according to the RBT (random book transformation) when dividing it into parts [2, 1]: the probability for a word that appears k' times in the full book of size M to appear k times in a smaller section of size M' can be expressed in binomial coefficients. Let $P_M(k')$ and $P_{M'}(k)$ be two column matrices with elements numerated by k' and k, then

$$\boldsymbol{P}_{M'}(k) = C \sum_{k'=k}^{M} \boldsymbol{A}_{kk'} \boldsymbol{P}_M(k') \tag{6}$$

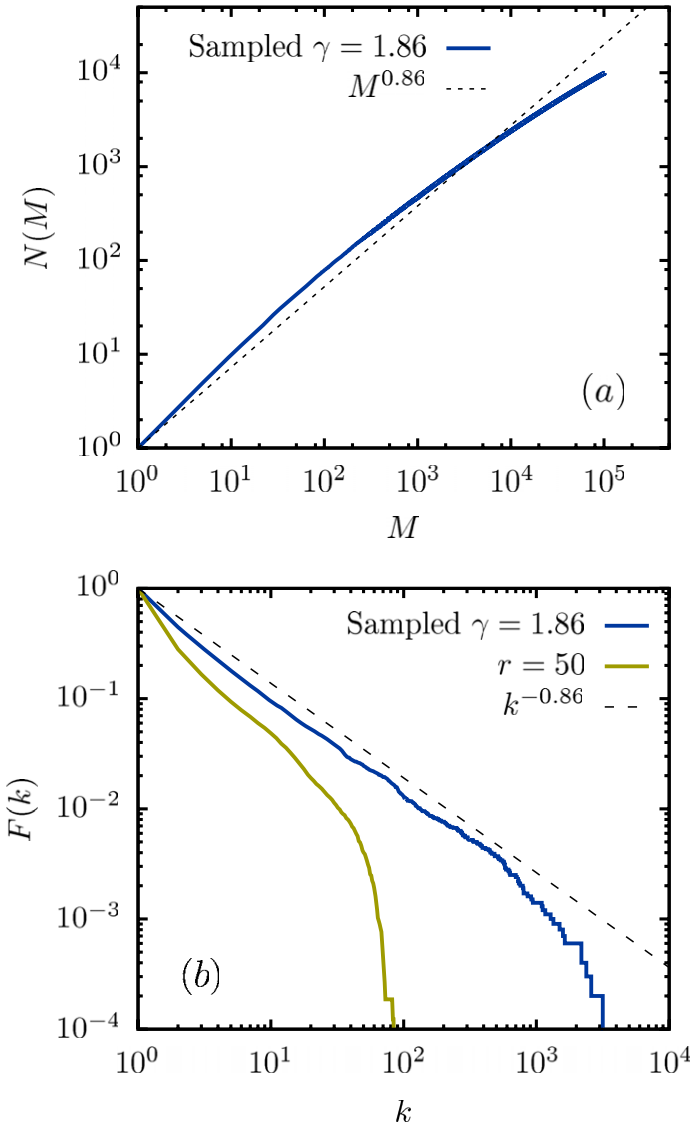

Figure 3. Results for a 'sampled book' of length M described by a smooth power-law wfd $P(k) \propto k^{-\gamma}$. (a) The full drawn curve is the real $N(M)$ whereas the broken straight line is the Heaps' power-law prediction from equation (7). Since the real $N(M)$-curve is bent, it is clear that a power-law wfd does not give a power law $N(M)$. (b) Illustrates that the wfd obtained for a part of the full book containing M' words where $r = M/M'$ has a different functional form than P_M. The curves show the cumulative distributions $F(k) = \sum_{k'=k}^{M} P(k')$ for the full random book $M = 10^6$ and $M' = 5000$, respectively.

where $A_{kk'}$ is the triangular matrix with the elements

$$A_{kk'} = (r-1)^{k'-k} \frac{1}{r^{k'}} \binom{k'}{k} \tag{7}$$

and $r = M/M'$ is the ratio of the book sizes. The normalization factor C is

$$C = \frac{1}{1 - \sum_{k'=1}((M-M')/M)^{k'} P_M(k')}. \tag{8}$$

Suppose that $P_M(k)$ is a power law with an index γ. The requirement for the corresponding random book to obey Heaps' law is then that $P_M(k)$ under the RBT-transformation remains a power law with the same index γ. However, the RBT-transformation does not leave invariant a power law with an index $\gamma > 1$ [2, 3]. This fact is illustrated in figure 3(b), which shows that a power law $P_M(k)$ changes its functional form when describing a smaller part of the book. This change of the functional form is the reason for why the $N(M)$-curve in figure 3(a) does not obey Heaps' law. The implication of this is that a random book which is well described by the continuum approximation $P(k) \propto 1/k^\gamma$ can never have an $N(M)$-curve of the Heaps' law form $N(M) \propto M^\alpha$.

The explanation for the size invariance of the monkey book can be found in the derivation presented in the appendix. Since the frequency of each word is exponential in the length of the word, a discrete size-invariant property of the book is naturally introduced. This discreteness is responsible for the disjoint peaks shown in figure 1(a), and it is easy to realize that non-overlapping Gaussian peaks will transform into new Gaussian peaks with conserved relative amplitudes, thus resulting in a size-independent envelope.

In figures 4(a) and (b) we compare the result for a power law $P_M(k)$ in figures 3(a) and (b) to the real book *Moby Dick* by Herman Melville. Figure 4(a) shows the $N(M)$ for $M \approx 212\,000$ both for the real book and for a randomized version (where the words in the real book are randomly re-distributed throughout the book) [3]. As seen, the $N(M)$-curves for the real and randomized books are very close and very reminiscent of the pure power-law case in figure 3(a): real and random books, as well as power-law books, deviate from Heaps' law in the same way. In figure 4(b) we show that the reason is the same: the form of the wfd changes with the size of the book in similar ways. The result for the real book is not a property solely found in *Moby Dick*, but has previously been shown to be a ubiquitous feature of novels [2].

5. Conclusions

We have shown that the $N(M)$-curve for a monkey book obeys Heaps' power-law form $N(M) \propto M^\alpha$ very precisely. This is in contrast to real and randomized real books, as well as sampled books with word-frequency distributions (wfd) which are well described by smooth power laws: all of these have $N(M)$-curves which deviate from Heaps' law in similar ways. In addition, we discussed the incompatibility of simultaneous power-law forms of the wfd and the $N(M)$-curves (Heaps' law). This led to the somewhat counter-intuitive conclusion that Heaps' power law requires a wfd which is *not* a smooth power law! We have argued that the reason for this inconsistency is that the simple derivation that leads to Heaps' law when starting from a power-law wfd assumes that the functional form is size independent when sectioning down the book to smaller sizes. However, it is shown, using the random book transformation (RBT), that this assumption is in fact not true for real or randomized books, nor for a sampled power-law book. In contrast, a monkey book, which has a spiked and disjoint wfd, possesses an invariance under this transformation. It is shown that this invariance is a direct consequence of the discreteness in the frequencies of words due to the discreteness in the lengths of the words (see appendix).

J. Stat. Mech. (2011) P07013

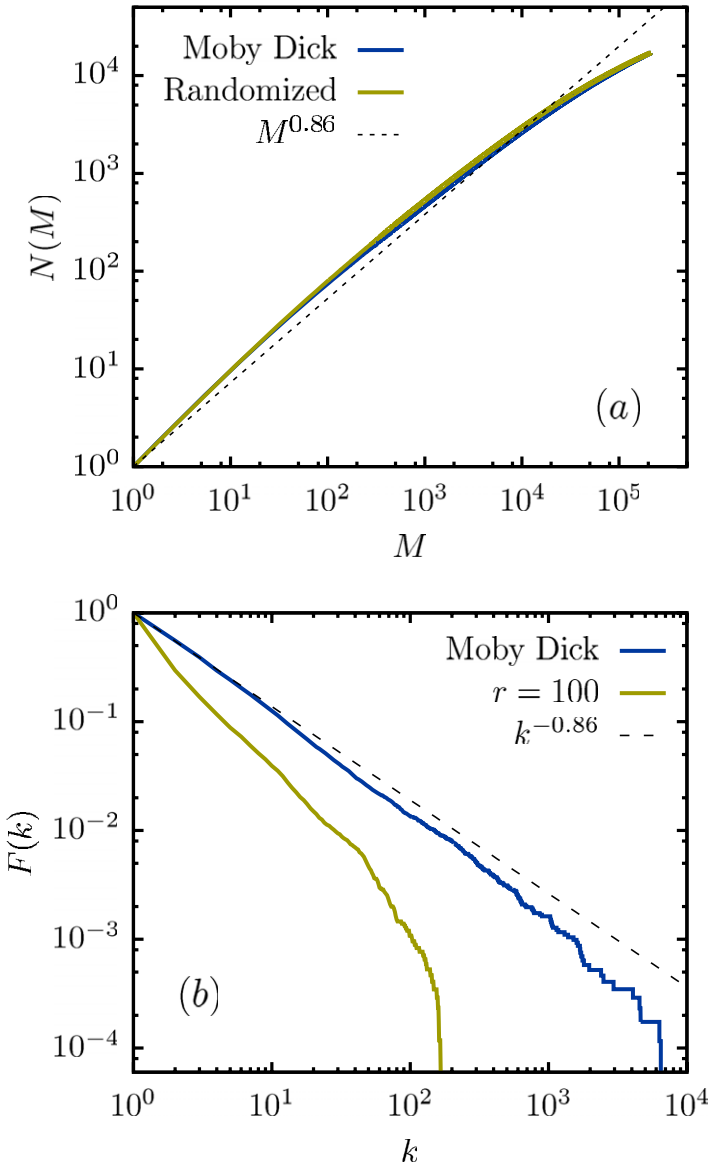

Figure 4. Comparison with a real book. (a) $N(M)$-curves for *Moby Dick* (dark curve) and for the randomized *Moby Dick* (light curve) together with a power law (straight broken line). The real and random versions of *Moby Dick* have to an excellent approximation the same $N(M)$ and this $N(M)$-curve is not a power law. Note the striking similarity with figure 3(a). (b) The change in the cumulative distribution $F(k)$ with text length for *Moby Dick*. The dark curve corresponds to the full length $M_{\text{tot}} \approx 212\,000$ words and the light curve to $M' \approx 2000$ ($r = M/M' = 100$). The change in the functional form of the wfd is very similar to that for the power-law book shown in figure 3(b).

Appendix. The information cost method

Let us imagine a monkey typing on a keyboard with $\mathcal{A}$ letters and a space bar, where the chance for typing a space is q_s and for any of the letters is $(1 - q_s)/\mathcal{A}$. A text produced by this monkey has a certain information content given by the entropy of the letter

doi:10.1088/1742-5468/2011/07/P07013

J. Stat. Mech. (2011) P07013

configurations produced by the monkey. These configurations result in a word-frequency distribution (wfd) $P(k)$ and the corresponding entropy $S = -\sum_k P(k) \ln P(k)$ gives a measure of the information associated with this frequency distribution. The most likely $P(k)$ corresponds to the maximum of S under the appropriate constraints. This can be formulated in terms of the maximum mutual information of the probability function $P(k)$ and the *a priori* given probability, p_i, for hitting a letter or a space bar [17]. Equivalently it can be phrased as the minimum of the constrained entropy $H[p_i|P(k)]$ [14, 17]. Here, we use the latter formulation because $H[p_i|P(k)]$ has a simple interpretation: it corresponds to the minimum information loss, or cost [14]. Consequently, the minimum cost $P(k)$ gives the most likely wdf for a monkey book.

Since the wfd in the continuum approximation is different from the real distribution $P(k)$, we will call the former $p(k)$. Let k be the frequency with which a specific word occurs in a text and let the corresponding probability distribution be $p(k)\,\mathrm{d}k$. This means that $p(k)\,\mathrm{d}k$ is the probability that a word belongs to the frequency interval $[k, k + \mathrm{d}k]$. The entropy associated with the probability distribution $p(k)$ is $S = -\sum_k p(k) \ln p(k)$ (where $\sum_k$ implies an integral whenever the index is a continuous variable). Let $M(l)\,\mathrm{d}l$ be the number of words in the word letter length interval $[l, l + \mathrm{d}l]$. This means that the number of words in the frequency interval $[k, k + \mathrm{d}k]$ is $M(l)(\mathrm{d}l/\mathrm{d}k)\,\mathrm{d}k$ because all words of a given length l occur with the same frequency. The number of distinct words in the same interval is $n(k)\,\mathrm{d}k = Np(k)\,\mathrm{d}k$, which means that $(M(l)/n(k))(\mathrm{d}l/\mathrm{d}k)$ is the degeneracy of a word with frequency k. The information loss due to this degeneracy is $\ln((M(l)/n(k))(\mathrm{d}l/\mathrm{d}k)) = \ln(M(l)\mathrm{d}l/\mathrm{d}k) - \ln p(k) + \text{const}$. The average information loss is given by

$$I_{\text{cost}} = \sum p(k)[-\ln p(k) + \ln(M(l)\,\mathrm{d}l/\mathrm{d}k)] \tag{A.1}$$

and this is the appropriate information cost associated with the words: the $p(k)$ which minimizes this cost corresponds to the most likely $p(k)$ [14]. The next step is to express $M(l)$ and $\mathrm{d}l/\mathrm{d}k$ in terms of the two basic probability distributions, $p(k)$ and the probability for hitting the keys: $M(l)$ is just $M(l) \sim \mathcal{A}^l$. The frequency k for a world containing l letters is

$$k \sim \left(\frac{1 - q_{\text{s}}}{\mathcal{A}}\right)^l q_{\text{s}}. \tag{A.2}$$

Thus $k \sim \exp(al)$ with $a = \ln(1 - q_{\text{s}}) - \ln \mathcal{A}$ so that $\mathrm{d}k/\mathrm{d}l = ka$ and, consequently, $I_{\text{loss}} = -\sum p(k) \ln p(k) + \sum p(k)[\ln \mathcal{A}^l - \ln ka]$. Furthermore, $\ln(\mathcal{A}^l/ka) = l \ln \mathcal{A} - \ln k - \ln a$ and from equation (A.2) one gets $l = \ln(k/q_{\text{s}})/\ln(1 - q_{\text{s}})/\mathcal{A})$ from which it follows that $\ln(\mathcal{A}^l/ka) = (-1 + (\ln \mathcal{A})/(\ln(1 - q_{\text{s}}) - \ln \mathcal{A})) \ln k + \text{const}$. Thus the most likely distribution $p(k)$ corresponds to the minimum of the information word cost

$$I_{\text{cost}} = -\sum p(k) \ln p(k) + \sum p(k) \ln k^{-\gamma} \tag{A.3}$$

with

$$\gamma = \frac{2 \ln \mathcal{A} - \ln(1 - q_{\text{s}})}{\ln \mathcal{A} - \ln(1 - q_{\text{s}})}. \tag{A.4}$$

Variational calculus then gives $\ln(p(k)k^{\gamma}) = \text{const}$, so that

$$p(k) \propto k^{-\gamma}. \tag{A.5}$$

J. Stat. Mech. (2011) P07013

Note that the total number of words M only enters this estimate though the normalization condition. This means that the continuum approximation $p(k) \propto (1/k^{\gamma})$ for the monkey book is independent of how many words M it contains. Thus if you start from a monkey book with M words and you randomly pick a fraction of these M words, then this smaller book will also a have a wfd which in the continuum limit follows the same power law. This is a consequence of the fact that the frequency k for a word of length l is always given by equation (A.2) irrespective of the book size. It is this specific monkey book constraint which makes I_{cost} in equation (A.1) M-invariant and hence forces the continuum $p(k)$ to always follow the same power law. The crucial point to realize is that the very same constraint forces the real $P(k)$ to have a 'peaky' structure. One should also note that if one started from a book consisting of M words randomly drawn from the continuum $p(k)$ then a randomly drawn fraction from this book would no longer follow the original power law.

References

[1] Baayen R H, 2001 *Word Frequency Distributions* (Dordrecht: Kluwer Academic)
[2] Bernhardsson S, Correa da Rocha L E and Minnhagen P, *The meta book and size-dependent properties of written language*, 2009 *New J. Phys.* **11** 123015
[3] Bernhardsson S, Correa da Rocha L E and Minnhagen P, *Size dependent word frequencies and translational invariance of books*, 2010 *Physica* A **389** 330
[4] Newman M E J, *Power laws, Pareto distributions and Zipf's law*, 2005 *Contem. Phys.* **46** 323
[5] Zipf G, 1932 *Selective Studies and the Principle of Relative Frequency in Language* (Cambridge, MA: Harvard University Press)
[6] Zipf G, 1935 *The Psycho-Biology of Language: An Introduction to Dynamic Philology* (Boston, MA: Mifflin Company)
[7] Zipf G, 1949 *Human Bevavior and the Principle of Least Effort* (Reading, MA: Addison-Wesley)
[8] Simon H, *On a class of skew distribution functions*, 1955 *Biometrika* **42** 425
[9] Mandelbrot B, 1953 *An informational theory of the statistical structure of languages* (Woburn, MA: Butterworth)
[10] Mitzenmacher M, *A brief history of generative models for power law and lognormal distributions*, 2003 *Internet Math.* **1** 226
[11] Miller G A, *Some effects of intermittance silence*, 1957 *Am. J. Psychol.* **70** 311
[12] Li W, *Random texts exhibit Zipf's-law-like word frequency distribution*, 1992 *IEEE Trans. Inf. Theory* **38** 1842
[13] Ferrer-i-Cancho R and Elvevåg B, *Random texts do not exhibit the real Zipf's law-like rank distribution.*, 2010 *PLoS One* **5** e9411
[14] Baek S K, Bernhardsson S and Minnhagen P, *Zipf's law unzipped*, 2011 *New J. Phys.* **13** 043004
[15] Conrad B and Mitzenmacher M, *Power laws for monkeys typing randomly: the case of unequal probabilities*, 2004 *IEEE Trans. Inf. Theory* **50** 1403
[16] Heap H S, 1978 *Information Retrieval: Computational and Theoretical Aspects* (New York: Academic)
[17] Cover T M and Thomas J A, 2006 *Elements of Information Theory* (New York: Wiley)

J. Stat. Mech. (2011) P07013