PHYSICAL REVIEW E **88**, 062804 (2013)

Explaining Zipf's law via a mental lexicon

Armen E. Allahverdyan,[1,2] Weibing Deng,[1,3,4] and Q. A. Wang[1,3]

[1]*Laboratoire de Physique Statistique et Systèmes Complexes, ISMANS, 44 ave. Bartholdi, 72000 Le Mans, France*
[2]*Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia*
[3]*IMMM, UMR CNRS 6283, Université du Maine, 72085 Le Mans, France*
[4]*Complexity Science Center and Institute of Particle Physics, Hua-Zhong Normal University, Wuhan 430079, China*

(Received 21 December 2012; revised manuscript received 22 May 2013; published 3 December 2013)

Zipf's law is the major regularity of statistical linguistics that has served as a prototype for rank-frequency relations and scaling laws in natural sciences. Here we show that Zipf's law—together with its applicability for a single text and its generalizations to high and low frequencies including hapax legomena—can be derived from assuming that the words are drawn into the text with random probabilities. Their *a priori* density relates, via the Bayesian statistics, to the mental lexicon of the author who produced the text.

DOI: 10.1103/PhysRevE.88.062804 PACS number(s): 89.75.Fb, 05.65.+b, 89.75.Da

I. INTRODUCTION

Zipf's law states that in a given text the ordered and normalized frequencies $f_1 > f_2 > \cdots$ for the occurrence of the word with rank r behave as $f_r \propto r^{-\gamma}$ with $\gamma \approx 1$ [1,2]. This law applies to texts written in many natural and artificial languages. Its almost universal validity has fascinated generations of scholars, but its message is still not well understood: Is it just a consequence of simple statistical regularities [3,4] or does it reflect a deeper structure of the text [5]?

Many approaches have been proposed for deriving Zipf's law, suggesting that it can have different origins. They are divided into several groups.

(1) Some theories approach deduction of the law from certain general premises of the language [3,6–11]. This group includes the Zipf's program whereby the language trades off between maximizing the information transfer and minimizing the speaking-hearing effort (this accounts for multifunctionality and short length of the most frequent words). This program is so far also not conclusive: It is not clear whether it really reproduces Zipf's law; see Refs. [7,8] for a recent review. Another derivation of the law is based on the idea that the words organize into a hierarchical structure, where the most frequent words are the ones with a wider meaning [11].

The general problem of derivations from this group is that explaining Zipf's law for the language (and verifying it for a frequency dictionary or for a large corpus) does not yet explain the law for a concrete text, where the frequency of the same word varies widely from one text to another and is far from its value in a frequency dictionary [12]. Hence, the above derivations do not explain why Zipf's law applies to a single text.[1]

(2) The law can be derived from certain probabilistic models [4,13–16]. Albeit some of these models assume relevance for realistic text-generating processes [14,15], their *a priori* assumed probability structure is intricate; hence, the question "Why Zipf's law?" translates into "Why a specific probabilistic model?" By far the most known probabilistic model is a random text, where words are generated through random combinations of letters and the space symbol seemingly reproducing the $f_r \propto r^{-1}$ shape of the law [3,4]. But the reproduction is elusive, since the model leads to a huge redundancy—many words have the same frequency and length—absent in normal texts [17]; see also Ref. [18] in this context. A recent study outlines in detail the statistical differences between random and usual texts and reviews previous literature [19].

(3) Derivations from various generalizations of the maximum entropy method [12,20–24]. Here one employs the most noninformative (most disordered) distribution of word frequencies compatible with certain constraints. However, the choice of the entropy function to be maximized (and of relevant constraints) is neither unique nor completely clear, in contrast to the original maximum entropy method, as applied in statistical physics [25].[2] A related approach to deriving Zipf's law employs not the maximal entropy method directly but rather tools of entropy-based complexity theory [27].

Note that Zipf's law does not hold for all ranks: Rank-frequency relations deviate from Zipf's law at low and high frequencies [1,2]. The high-frequency deviation relates to functional words, while the low-frequency one indicates on *hapax legomena* the domain of rare words [1,2]. This domain is subject to a specific relation (sometimes called the Lotka's law or the second law of Zipf) that is normally discussed separately from Zipf's law proper [2,28]. However, one expects that Zipf's law does allow generalizations to high and low frequencies and that also the validity range of Zipf's law proper will come out from this generalization.

Our approach for deriving Zipf's law uses a probability model. It differs from previous models in several respects. First, it explains the law for a single text together with its limits of validity, i.e., together with the range of ranks where it holds. It also explains the rank-frequency relation for frequent (functional) words, as well as for very rare words (hapax legomena) and relates them to Zipf's law. In particular, for

[1] For example, if the word frequencies obeying Zipf's law are deduced from considerations related to the meaning of the words (as in Ref. [11]), then applicability to a single text is unclear, since the fact that the words normally have widely different frequencies in different texts requires a substantial reconsideration of the word's meaning in each text (this is not the case in real texts).

[2] In this context we note that Zipf's law can be related to modern ideas of statistical physics such as scaling and universality [26].

©2013 American Physical Society

TABLE I. Parameters of three texts: *The Age of Reason* (AR) by T. Paine, 1794 (the major source of British deism), *Thoughts on the Funding System and Its Effects* (TF) by P. Ravenstone, 1824 (economics), and *Dream Lover* (DL) by J. MacIntyre, 1987 (romance novella). Total number of words is N, the number of different words is n, $r_{\min}$ and $r_{\max}$ are, respectively, the lower and the upper ranks of the Zipfian domain, and c and γ are the fitted values of the parameters.

Texts	N	n	$r_{\min}$	$r_{\max}$	c	γ
TF	26624	2067	36	371	0.168	1.032
AR	22641	1706	32	339	0.178	1.038
DL	24990	1748	34	230	0.192	1.039

hapax legomena we propose a regularity that works better than the Lotka's law.

Second, the *a priori* structure of our model can be connected to the mental lexicon [29] of the author who produced the text. Third, the model is not *ad hoc*: It is based on the ideas of latent semantic analysis that is used successfully for text modeling [30]. The latent semantic analysis arose from a long scientific tradition (in statistics, physics, sociology, etc.), where observed effects are explained through simpler (but hidden) regularities; see, e.g., Ref. [31] for a general introduction.

This paper is organized as follows. Section II discusses the validity range of Zipf's law employing empiric data from three English texts. Section III introduces our model and discusses its basic assumptions. In Sec. IV we solve this model, while the next section, Sec. V, presents the theoretical understanding of Zipf's law. Section VI validates the model from features of the mental lexicon (no preliminary knowledge of this psycholinguistic concept is assumed). We conclude in the last section. Technical questions are discussed in the Appendices.

II. THE VALIDITY RANGE OF ZIPF'S LAW

A. Linear fitting

Below we present empirical results exemplified on three English texts (see Table I) that clarify the validity range of the law and confirm known results but also make new points that motivate the theoretical model worked out in the sequel.

For each text we extract the ordered frequencies of n different words:

$$\{f_r\}_{r=1}^{n}, \quad f_1 \geqslant \cdots \geqslant f_n, \quad \sum_{r=1}^{n} f_r = 1. \tag{1}$$

To fit $\{f_r\}_{r=1}^{n}$ to the Zipf's form $\hat{f}_r = cr^{-\gamma}$, we represent the data as $\{y_r(x_r)\}_{r=1}^{n}$, where $y_r = \ln f_r$ and $x_r = \ln r$, and fit it to the linear form $\{\hat{y}_r = \ln c - \gamma x_r\}_{r=1}^{n}$. Two unknowns $\ln c$ and γ are obtained from minimizing the sum of squared errors (see Appendix A for details),

$$S_{\rm err} = \sum_{r=1}^{n} (y_r - \hat{y}_r)^2. \tag{2}$$

Now $\min_{c,\gamma}[S_{\rm err}] = S^*_{\rm err}$ and the correlation coefficient R^2 between $\{y_r\}_{r=1}^{n}$ and $\{\hat{y}_r\}_{r=1}^{n}$ (see Appendix A) measure the fitting quality

$$S^*_{\rm err} \to 0 \quad \text{and} \quad R^2 \to 1, \tag{3}$$

indicating a good fit. We minimize $S_{\rm err}$ over c and γ for $r_{\min} \leqslant r \leqslant r_{\max}$ and find the maximal value of $r_{\max} - r_{\min}$ for which $S^*_{\rm err}$ and $1 - R^2$ are smaller than, respectively, 0.05 and 0.005. This value of $r_{\max} - r_{\min}$ also determines the final fitted values of c and γ; see Table I and Appendix A for technical details. The quality of this fitting was additionally confirmed via the Kolmogorov-Smirnov (KS) test; see Appendix A 2.

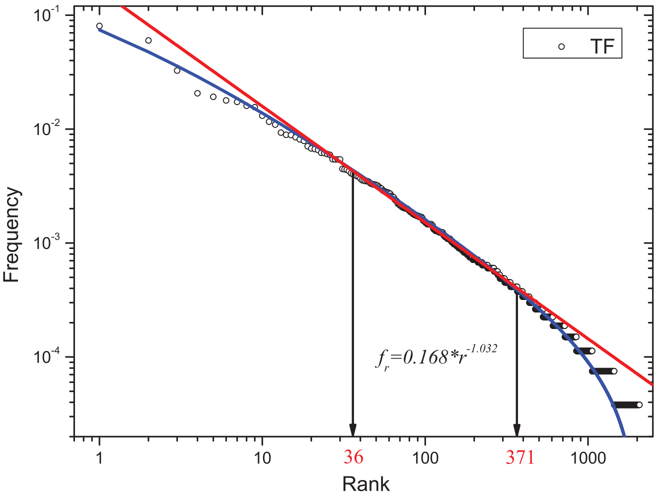

FIG. 1. (Color online) Frequency f_r vs rank for the text TF; see Table I and (1). Red (upper, straight) line: the Zipf curve $f_r = 0.168 r^{-1.032}$. Arrows indicate on the validity range of the Zipf's law. Blue (lower, curved) line: the solution of (13) and (14) for $c = 0.168$ and $n = 2067$. It coincides with the generalized Zipf law (23) for $r > r_{\min} = 36$. The stepwise behavior of f_r for $r > r_{\max}$ refers to hapax legomena.

B. Main empiric features of Zipf's law

For each text there is a specific (Zipfian) range of ranks $r \in [r_{\min}, r_{\max}]$, where the Zipf's law holds with $\gamma \approx 1$ and $c < 0.2$ [1,2]; see Table I and Fig. 1. For $r < r_{\min}$ and $r > r_{\max}$ the law is invalid, since the frequencies are below the Zipf curve. We show below how a consistent generalization of the law allows us to cover these domains as well.

Even if the same word enters into different texts it typically has quite different frequencies [12], e.g., among 83 common words in the Zipfian ranges of AR and DL (see Table I), only 12 words have approximately equal ranks and frequencies.

The pre-Zipfian $1 \leqslant r < r_{\min}$ range contains mainly function words. They serve for establishing grammatical constructions (e.g., *the, a, such, this, that, where, were*).[3] But the majority of words in the Zipfian range do have a narrow meaning (content words). A subset of those content words has a meaning that is specific for the text and can serve as its keywords [32]. Below in Sec. VI we explain why the keywords appear in the Zipfian domain.

The absolute majority of different words with ranks in $[r_{\min}, r_{\max}]$ have different frequencies. The number of different

[3] Few keywords appear also in the pre-Zipfian range, e.g., *love* and *miss* for DL and *god* and *man* for AR. Some keywords are also located in the post-Zipfian area, e.g., *eloi* for TM, but the majority of them are in the Zipfian range.

words having the same frequency is $\simeq 10$ only for $r \simeq r_{\max}$. For $r > r_{\max}$ we meet the *hapax legomena*: Words occurring only a few times in the text ($f_r N = 1,2,\ldots$ is a small integer) and many words having the same frequency f_r [2]. The effect is not described by a smooth rank-frequency relation, including Zipf's law. Hence, in this sense Zipf's law holds for as high ranks as possible (i.e., for $r > r_{\max}$ no any smooth rank-frequency relation is expected to work).

Note that the very existence of hapax legomena is a nontrivial effect, since one can easily imagine (artificial) texts, where (say) no character appear only once. The theory reviewed below in Sec. V B allows us to explain the hapax legomena range together with Zipf's law. It also predicts a generalization of Zipf's law to frequencies $r < r_{\min}$ that fits better (than Zipf's law) to the empiric data; see Fig. 1.

The minimal frequency of the Zipfian domain holds

$$f_{r_{\max}} > c/n. \tag{4}$$

We checked that this is valid not only for separate texts but also for the frequency dictionaries of English and Irish. For our texts a stronger—but less precise—relation holds,

$$f_{r_{\max}} \gtrsim \frac{1}{n}. \tag{5}$$

Hence, $f_{r_{\max}} N \gtrsim \frac{N}{n} \gg 1$; see Table I.

Thus, once the validity range $[r_{\min}, r_{\max}]$ of Zipf's law is determined via strict statistical criteria (as we did above), its linguistic meaning emerges clearly: $[1, r_{\min}]$ contains mostly functional words, while the region $[r_{\max}, n]$ carries rare words (hapax legomena). In particular, the keywords are mainly contained in the Zipfian range $[r_{\min}, r_{\max}]$.

III. INTRODUCTION TO THE MODEL

A. Necessary conditions for a model explaining Zipf's law

A model for Zipf's law is supposed to satisfy the following features.

(I) Apply to separate texts, i.e., explain how different texts can satisfy the same form of the rank-frequency relation despite the fact that the same words do *not* occur with same frequencies in the different texts; see Sec. II B.

(II) Derive the law together with its extensions for all frequencies, limits of validity, and the hapax legomena effect.

(III) Relate the law to formation of a text.

B. Main assumptions of the model

Two sources of the model are the latent semantic analysis [30], and the idea of applying ordered statistics for rank-frequency relations [9,33,34].

Our model makes four [(A)–(D)] assumptions.

(A) The *bag-of-words picture* focusses on the frequency of the words that occur in a text and neglects their mutual disposition (i.e., syntactic structure) [35]. This is a natural assumption for a theory describing word frequencies, which are invariant with respect to an arbitrary permutation of the words in a text.

Given n different words $\{w_k\}_{k=1}^n$, the joint probability for w_k to occur $\nu_k \geqslant 0$ times in a text T is multinomial,

$$\pi[\boldsymbol{\nu}|\boldsymbol{\theta}] = \frac{N!\,\theta_1^{\nu_1}\ldots\theta_n^{\nu_n}}{\nu_1!\ldots\nu_n!}, \quad \boldsymbol{\nu} = \{\nu_k\}_{k=1}^n, \quad \boldsymbol{\theta} = \{\theta_k\}_{k=1}^n, \tag{6}$$

where $N = \sum_{k=1}^n \nu_k$ is the length of the text, ν_k is the number of occurrences of w_k, and θ_k is the probability of w_k.

Hence, according to (6) the text is regarded to be a sample of word realizations drawn independently with probabilities θ_k. Note that the bag-of-words picture [together with (6)] resembles in several respects the ideal gas of statistical physics (no direct interactions between the words-particles).

The bag-of-words picture is well known in computational linguistics [35]. But for our purposes it incomplete, because it implies that each word has the same probability for different texts [recall (I)].

(B) To improve this point we make $\boldsymbol{\theta}$ a random vector [35] with a text-dependent density $P(\boldsymbol{\theta}|T)$. With this assumption the variation of the word frequencies from text to another will be explained by the randomness of the word probabilities.[4]

The above assumption is in fact basic for statistical physics of disordered systems [36]. Here certain parameters—e.g., interaction parameters for complex nuclei [36]—change widely from one sample to another. For modeling this situation one assumes that these parameters are random. The assumption agrees with experiments provided that the probability density of these parameters respect certain gross features of the system, e.g., its symmetries [36]. In our situation these gross features relate to those of the mental lexicon (for the text-producing author); see Sec. VI for details.

We now have three random objects: text T, probabilities $\boldsymbol{\theta}$, and the occurrence numbers $\boldsymbol{\nu}$. Since $\boldsymbol{\theta}$ was introduced to explain the relation of T with $\boldsymbol{\nu}$, it is natural to assume that the triple $(T,\boldsymbol{\theta},\boldsymbol{\nu})$ form a Markov chain: The text T influences the observed $\boldsymbol{\nu}$ only via $\boldsymbol{\theta}$. Then the probability $p(\boldsymbol{\nu}|T)$ of $\boldsymbol{\nu}$ in a given text T reads

$$p(\boldsymbol{\nu}|T) = \int d\boldsymbol{\theta}\, \pi[\boldsymbol{\nu}|\boldsymbol{\theta}]\, P(\boldsymbol{\theta}|T). \tag{7}$$

This form of $p(\boldsymbol{\nu}|T)$ is basic for probabilistic latent semantic analysis [30], a successful method of computational linguistics. There the density $P(\boldsymbol{\theta}|T)$ of latent variables $\boldsymbol{\theta}$ is determined from the data fitting. But we shall deduce $P(\boldsymbol{\theta}|T)$ theoretically.

(C) The text-conditioned density $P(\boldsymbol{\theta}|T)$ is generated from a prior density $P(\boldsymbol{\theta})$ via conditioning on the ordering of $\mathbf{w} = \{w_k\}_{k=1}^n$ in T,

$$P(\boldsymbol{\theta}|T) = P(\boldsymbol{\theta})\,\chi_T(\boldsymbol{\theta},\mathbf{w}) \Big/ \int d\boldsymbol{\theta}'\, P(\boldsymbol{\theta}')\,\chi_T(\boldsymbol{\theta}',\mathbf{w}). \tag{8}$$

[4] The assumption on random $\boldsymbol{\theta}$ was made within the bag-of-words picture, but it was additionally assumed that $u(\theta)$ follows the Dirichlet density $u(\theta) = \theta^{-\zeta}$ $(1 > \zeta \geqslant 0)$ [35], which is used almost exclusively for a density over probabilities. The Dirichlet density with $\zeta = 0$ was used for modeling the rank-frequency relation of letters [33]. But our purpose implies a different choice for $u(\theta)$; see (20).

Thus if different words of T are ordered as $(w_1,\dots,w_n)$ with respect to the decreasing frequency of their occurrence in T (i.e., w_1 is more frequent than w_2), then $\chi_T(\boldsymbol{\theta},\mathbf{w})$ is defined as follows: $\chi_T(\boldsymbol{\theta},\mathbf{w})=1$ if $\theta_1\geqslant\cdots\geqslant\theta_n$ and $\chi_T(\boldsymbol{\theta},\mathbf{w})=0$ otherwise.

As substantiated below in Sec. VI, $P(\boldsymbol{\theta})$ refers to the mental lexicon of the author prior to generating a concrete text.

(D) For simplicity, we assume that the probabilities θ_k are distributed identically (see Sec. VI for a partial verification of this assumption) and the dependence among them is due to $\sum_{k=1}^{n}\theta_k=1$ only,

$$P(\boldsymbol{\theta})\propto u(\theta_1)\dots u(\theta_n)\,\delta\left(\sum_{k=1}^{n}\theta_k-1\right), \tag{9}$$

where $\delta(x)$ is the δ function and the normalization ensuring $\int_0^\infty\prod_{k=1}^{n}d\theta_k\;P(\boldsymbol{\theta})=1$ is omitted.

IV. SOLUTION OF THE MODEL

The conditional probability $p_r(\nu|T)$ for the rth most frequent word w_r to occur ν times in the text T reads from (6) and (7),

$$p_r(\nu|T)=\frac{N!}{\nu!(N-\nu)!}\int_0^1 d\theta\,\theta^{\nu}(1-\theta)^{N-\nu}P_r(\theta|T), \tag{10}$$

$$P_r(t|T)=\int d\boldsymbol{\theta}\;P(\boldsymbol{\theta}|T)\delta(t-\theta_r), \tag{11}$$

where $P_r(t|T)$ is the marginal density for the probability t of w_r. For $n\gg 1$, we deduce from (8) and (9) that $P_r(t|T)$ follows the law of large numbers; see Appendix A. It is Gaussian,

$$P_r(t|T)\propto\exp\left[-\frac{n^3}{2\sigma_r^2}(t-\phi_r)^2\right], \tag{12}$$

where $\sigma_r=\mathcal{O}(1)$ [for $\phi_r=o(1)$], and the mean ϕ_r is found from two equations for two unknowns μ and ϕ_r,

$$r/n=\int_{\phi_r}^{\infty}d\theta\,u(\theta)\,e^{-\mu\theta n}\Big/\int_0^{\infty}d\theta\,u(\theta)\,e^{-\mu\theta n}, \tag{13}$$

$$\int_0^{\infty}d\theta\,\theta\,u(\theta)\,e^{-\mu\theta n}=\frac{1}{n}\int_0^{\infty}d\theta\,u(\theta)\,e^{-\mu\theta n}. \tag{14}$$

Equation (12) holds for $P_r(t|T)$ whenever its standard deviation $\sigma_r n^{-3/2}$ is much smaller than the mean ϕ_r; as checked below, this happens already for $r>10$.

The meaning of (13) and (14) is explained via the marginal density

$$P(\theta_1)=\int_0^{\infty}\prod_{k=2}^{n}d\theta_k\;P(\boldsymbol{\theta}) \tag{15}$$

$$\propto u(\theta_l)e^{-\mu\theta_l n}, \tag{16}$$

found from (9); see Appendix C for a derivation. It is seen that the meaning of (14) is that it ensures

$$\int_0^{\infty}d\theta\,\theta\;P(\theta)=\frac{1}{n}. \tag{17}$$

This relation follows from $\sum_{k=1}^{n}\theta_k=1$ and it determines μ; see Appendices B and C. It is shown there that μ emerges from the saddle-point method and is related to the Lagrange multiplier of the constraint $\delta(\sum_{k=1}^{n}\theta_k-1)$ in Eq. (9). Thus it plays the same role as the chemical potential in statistical physics. The latter enforces the conservation of the particle number, while in our situation μ enforces the conservation of probability $\sum_{k=1}^{n}\theta_k=1$.

Now one possible interpretation of (13) is that it equates the relative rank r/n to the (unconditional) probability $\int_{\phi_r}^{\infty}d\theta\;P(\theta)$ of $\theta\geqslant\phi_r$. This type of reasoning is popular in heuristic derivations of power laws (including Zipf's law) [34].

In Eq. (10), $P_r(\theta|T)$ is much more narrow peaked than $\theta^{\nu}(1-\theta)^{N-\nu}$, since $n^3\gg N\gg 1$ (see Table I). Hence, in this limit, we approximate $P_r(\theta|T)$ by the δ function $\delta(\theta-\phi_r)$ [see (12)] and get from (10)

$$p_r(\nu|T)=\frac{N!}{\nu!(N-\nu)!}\phi_r^{\nu}(1-\phi_r)^{N-\nu}. \tag{18}$$

Equation (18) is the main outcome of the model; it shows that the conditional probability $p_r(\nu|T)$ for the occurrence number ν of the word w_r has the same form (18) for different text (see I). In Eq. (18), ϕ_r is the effective probability of the word w_r. If $N\phi_r\gg 1$, $p_r(\nu|T)$ is peaked at $\nu=N\phi_r$: the frequency of a word that appears many times equals its probability (the law of large numbers). Each word of the Zipfian domain occurs at least $\nu\sim N/n\gg 1$ times; see the discussion around (4) and (5). For such words we approximate

$$f_r\equiv\nu/N\simeq\phi_r. \tag{19}$$

It is seen below that the proper Zipf's law relates via (19) to the law of large numbers.

V. RESULTS AND DISCUSSION

A. The Zipfian range

So far $u(f)$ in Eq. (9) is left unspecified. Now we postulate

$$u(f)=(n^{-1}c+f)^{-2}, \tag{20}$$

where c is related below to the prefactor of Zipf's law. The postulate (20) is explained in Sec. VI below.

For $c\lesssim 0.2$, $c\mu$ determined from (14) and (20) is small and is found from integration by parts as follows:

$$\mu\simeq c^{-1}\,e^{-\gamma_{\mathrm{E}}-\frac{1+c}{c}}, \tag{21}$$

where $\gamma_E=0.55117$ is the Euler's constant. One solves (13) for $c\mu\to 0$ as follows:

$$\frac{r}{n}=ce^{-n\phi_r\mu}/(c+n\phi_r). \tag{22}$$

For $r>r_{\min}$, $\phi_r n\mu=f_r n\mu<0.04\ll 1$; see Eq. (21) and Table I. We get from (19) and (22)

$$f_r=c(r^{-1}-n^{-1}). \tag{23}$$

This is Zipf's law generalized by the factor n^{-1} at high ranks r. This cut-off factor ensures faster (than r^{-1}) decay of f_r for large r. In the literature a cut-off factor similar to $\frac{1}{n}$ is introduced due to additional mechanisms (hence, new parameters); see Ref. [14]. In our situation the power law and cut-off come from the same mechanism.

Figure 2 illustrates the approximate solution (23) of (13), (14), and (20); it is confirmed that the approximation is reliable for $c\leqslant 0.2$.

Figure 1 shows that (23) reproduces well the empirical behavior of f_r for $r>r_{\min}$. Our derivation shows that c is the

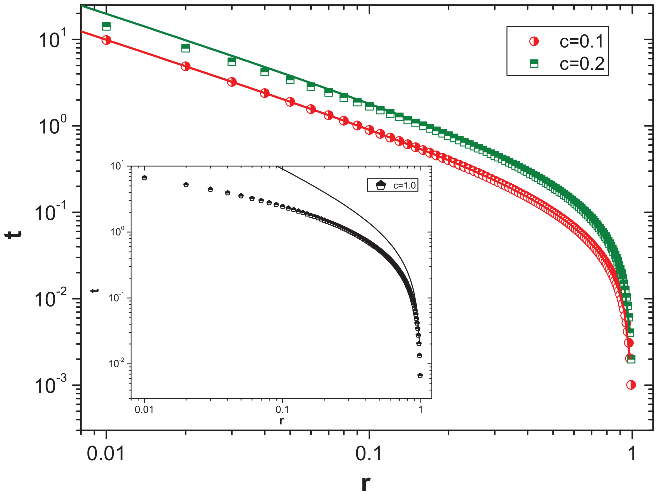

FIG. 2. (Color online) Main figure: Upper (green) curve: the generalized Zipf’s curve (23) written in the scaling form $\mathbf{t} = c[\mathbf{r}^{-1} - 1]$, where $\mathbf{t} = f_r n$, $\mathbf{r} = r/n$, and $c = 0.2$. Squares: solutions of (13), (14), and (20) for $c = 0.2$. Lower (red) curve: The same as for the upper curve but for $c = 0.1$. Rounds: Solutions of (13), (14), and (20) for $c = 0.1$. It is seen that the generalized Zipf’s curve coincide with the actual solutions. Inset: Solid curve: The same as for solid curves in the main figure but for $c = 1.0$. Dots: Solutions of (13), (14), and (20) for $c = 1.0$. Now the generalized Zipf’s curve does not coincide with the actual solution.

prefactor of Zipf’s law and that our assumption on $c < 0.2$ above (21) agrees with observations; see Table I. For $c \gg 0.2$, (13) and (14) do not predict Zipf’s law (23); see Fig. 2.[5]

When the prefactor c and the number of different words n are taken from empirical results, (13)–(20) predict the Zipfian range $[r_{\min}, r_{\max}]$, in agreement with the observed values of these quantities; see Fig. 1.

For $r < r_{\min}$, it is no longer true that $f_r n\mu \ll 1$. So the fuller expression (13) is to be used. It reproduces qualitatively the empiric behavior of f_r; see Fig. 1. We do not expect any better agreement theory and observations for $r < r_{\min}$ because the behavior of frequencies in this range is irregular.

B. Hapax legomena

TABLE II. Description of the hapax legomena for the text TF; see Table I and Eq. (24). The maximal relative error $\frac{\hat{r}_k - r_k}{r_k} = 0.0357$ is reached for $k = 6$.

r/k	1	2	3	4	5	6	7	8	9	10
r_k	1446	1061	848	722	611	529	474	437	398	370
$\hat{r}_k$	1414	1074	866	726	624	547	488	440	400	368

According to (18), the probability ϕ_r is small for $r \gg r_{\max}$ and, hence, the occurrence number $\nu \equiv f_r N$ of a words w_r is a small integer (e.g., 1 or 2) that cannot be approximated by a continuous function of r; see Eq. (18) and Fig. 1. To describe this hapax legomena range, define r_k as the rank, when $\nu \equiv f_r N$ jumps from integer k to $k + 1$. Since ϕ_r reproduces well the trend of f_r even for $r > r_{\max}$, see Fig. 1, r_k can be theoretically predicted from (23) by equating its left-hand side to k/N as follows:

$$\hat{r}_k = \left[\frac{k}{Nc} + \frac{1}{n}\right]^{-1}, \quad k = 0,1,2,\ldots \tag{24}$$

Equation (24) is exact for $k = 0$ and agrees with r_k for $k \geqslant 1$; see Table II. Hence, it describes the hapax legomena phenomenon (many words have the same small frequency).

For $k \gg Nc/n$ we deduce from (24) $\hat{r}_k - \hat{r}_{k+1} \propto k^{-2}$ for the number of words having the frequency k/N. This relation, which is a crude particular case of (24), is sometimes called the Lotka’s law or the second Zipf law [2,28].

C. Preliminary summary

Thus the theory presented in this section achieved the promises (I) and (II) of our program: though different texts can have different frequencies for same words, the frequencies of words in a given text follow Zipf’s law with the correct prefactor $c \lesssim 0.2$. Without additional fitting parameters and new mechanisms we recovered the corrected form of this law applicable for large and small frequencies (hapax legomena).

But why we would select (20) if we would not know that it reproduces Zipf’s law? Answering this question will fulfill (III).

VI. MENTAL LEXICON AND THE *A PRIORI* DENSITY

Here we explain the choice (9) and (20) for the *a priori* probability density for the probabilities $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ of different words $(w_1, \ldots, w_n)$. To avoid the awkward term “probability for probability,” we shall call $P(\boldsymbol{\theta})$ likelihood. We focus on the marginal likelihood (16) and (20),

$$P(\theta) = (n^{-1}c + \theta)^{-2} e^{-\mu n\theta}, \tag{25}$$

since $P(\theta)$ determines the rank-frequency relation (13); see (15) and (16).

A. General features of mental lexicon

Recall from Sec. III that the basic reason for the words to have random (not fixed) probabilities is that the text-producing author should be able to compose different texts, where the same word can have different frequencies. Hence, the likelihood $P(\boldsymbol{\theta})$ of random probabilities relates to the prior knowledge (or lexicon) of the text-generating author on the words. This concept of *mental lexicon*—the store of words in the long-time memory so the words are employed online for expressing thoughts via phrases and sentences—is well established in psycholinguistics [29]. Though the theory of

[5] Note that since Zipf’s law does not apply for all ranks, c is not a normalization constant, i.e., its value is not fixed from the fact that the sum of probabilities should be equal to 1. The normalization still allows us to put upper and lower bounds on c. First, recall from Fig. 1 that Zipf’s law is an upper bound for the frequencies at all ranks. This holds generally [1,2]. Then we get $c\sum_{k=1}^{n} k^{-1} \simeq c(\gamma_E + \ln n) > 1$, where $\gamma_E = 0.55117$ is Euler’s constant. Within the applicability range of Zipf’s law we have $c\sum_{k=r_{\min}}^{r_{\max}} k^{-1} \simeq c(\gamma_E + \ln \frac{r_{\max}}{r_{\min}}) < 1$. These two formulas bound c from above and below. For the text TF (see Table I) they produce $0.1218 < c < 03436$. Hence, the bounds do not explain the fact that for real texts $c < 0.2$; see Table I.

mental lexicon is yet under construction [29], some of its basic features are well-established experimentally and are employed below for explaining the choice (25).

We assume that during the conceptual planning of the text, i.e., when deciding on its topic, style, and potential audience, the author already chooses (at least approximately) two structural parameters: the number n of different words to appear there and the constant c. This is why the marginal likelihood (25) depends on the parameters c and n. We recall that c (along with n) is a structural parameter of the text (see footnote 5 in this context), e.g., because according to (4), c/n separates the Zipfian (keyword-dominated) range from the hapax legomena range (rare words).

According to (9), different words have the same marginal likelihood: the likelihood $P(\boldsymbol{\theta})$ is symmetric with respect to interchanging the words $w_1,\dots,w_n$. This feature relates to an experimental fact that words are stored in the mental lexicon in the same way [37]. The difference between them—e.g., whether the word is more familiar to the author, and/or used by him or her more frequently [38]—can be relevant during the (later) phonologization stage of speech and text production [37].

Naturally, the above symmetry holds for the *a priori* likelihood. The posterior likelihood $P(\boldsymbol{\theta}|T)$ [see (8)], the one that is conditioned over the written text, does not and should not have such a symmetry.

B. Bayesian group

Once each word w_k has to have a variable (random) probability θ_k, there should be a way for the author to change (increase or decrease) this probability, e.g., when the author decides that the word w_k is to become the keyword of the text. The ensuing relation between the probability vectors $\boldsymbol{\theta}'$ (new) and $\boldsymbol{\theta}$ (old) should be a group, since the author should be able to come back from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$, e.g., when revising the text.

One can impose two natural restrictions on this group [39]. These restrictions follow the general idea that the meaning of $\boldsymbol{\theta}$ as probabilities of certain events is conserved during the transformation.

First, the words that have strictly zero probability $\theta_k = 0$ will stay zero probability,

$$\theta_k' = 0 \text{ if and only if } \theta_k = 0. \tag{26}$$

This feature naturally means that groups operates without adding new words and without excluding the existing words.

Second, the probability mixtures are conserved: if

$$\boldsymbol{\theta} = \lambda\boldsymbol{\chi} + (1-\lambda)\boldsymbol{\eta}, \quad 0 < \lambda < 1, \tag{27}$$

where $\boldsymbol{\chi} = (\chi_1,\dots,\chi_n)$ and $\boldsymbol{\eta} = (\eta_1,\dots,\eta_n)$ are arbitrary probability vectors, and where λ is a (mixing) parameter, then

$$\boldsymbol{\theta}' = \lambda'\boldsymbol{\chi}' + (1-\lambda')\boldsymbol{\eta}', \quad 0 < \lambda' < 1. \tag{28}$$

Here primed and nonprimed probability vectors relate to each other via the sought group, while λ' and λ generally differ.[6]

The only group that (for $n \geqslant 3$) is consistent with the above two conditions [(26) and (27)] is [39]

$$\theta_k' = \frac{\tau_k\theta_k}{\sum_{l=1}^n \tau_l\theta_l}, \quad \tau_k > 0, \quad k = 1,\dots,n, \tag{29}$$

where τ_k are the group parameters. Equation (29) is a generalized Bayes formula [39].[7] It is used in the Bayesian statistics for motivating the choice of priors [39], a task related to ours.

If the author wants to increase τ_1 times the probability of the word w_1, then in Eq. (29) $\tau_1 > 1$ and $\tau_{k\geqslant 2} = 1$,

$$\theta_1' = \frac{\tau_1\theta_1}{1+(\tau_1-1)\theta_1}, \quad \theta_l' = \frac{\theta_l}{1+(\tau_1-1)\theta_1}, \quad \text{for} \quad l \geqslant 2. \tag{30}$$

The inverse of (30) is found by interchanging θ_k' with θ_k and τ_1 with τ_1^{-1}.

For Zipf's law (and assuming the limit $n \gg 1$) the relevant probabilities are small, $\theta_1' = \mathcal{O}(1/n)$; see Fig. 1. Then

$$1 + \left(\tau_1^{-1} - 1\right)\theta_1' \approx 1, \tag{31}$$

and (30) becomes the scaling transformation of one variable,

$$\theta_1' = \tau_1\theta_1, \quad \theta_l' = \theta_l, \quad l \geqslant 2. \tag{32}$$

The transformed likelihood reads from (32) and (25),

$$P'(\theta_1') = \frac{1}{\tau_1}P\left(\frac{\theta_1'}{\tau_1}\right) = \frac{1}{\tau_1}\left(\frac{c}{n} + \frac{\theta_1'}{\tau_1}\right)^{-2}, \tag{33}$$

where the factor $e^{-\mu n\theta_1'/\tau_1}$ was put to 1 [cf. (31)], since $n\theta_1' = \mathcal{O}(1)$ in the regime relevant for the Zipf's law and μ is small; see (21).

According to (32), other densities do not change $P'(\theta_l') = P(\theta_l')$ for $l \geqslant 2$.

C. The meaning of the likelihood (prior density)

Once $P(\theta)$ describes the mental lexicon, and (29) is an operation by which the text is written, we suppose that the features of $P(\theta)$ can be explained by checking its response

[6] The requirement on conserving mixtures can be explained as follows. Words have attributes (connotation, denotation, inclination, conjugation, etc.). When studying rank-frequency relations one naturally puts aside these features, i.e., one is ignorant of them. Hence, word probability f_k can be represented as a mixture $f_k = \sum_\alpha \lambda_\alpha f_{k\alpha}$, where α means the set of attributes and λ_α is a probability of those attributes. It is then natural to require that the group transformation conserves this representation over the attributes.

[7] Equation (29) becomes the Bayes formula if we relate τ_k to a conditional probability [25]. In this alternative interpretation of (29), the author has to retrieve a word w having certain specific features (i.e., it is a transitive verb) from the set of words $w_1,\dots,w_n$ having probabilities $\theta_1,\dots,\theta_n$. If we denote by $\Pr(E|w = w_k)$ the conditional probability that the word w_k displays the needed feature E, we can relate in Eq. (29) $\tau_k = \Pr(E|w = w_k)$, and (29) will describe the searching process for the word having the needed feature E.

to (29).[8] For the ratio of the new to the old likelihood of the probability θ_1' we get from (33)

$$P'(\theta_1')/P(\theta_1') = \tau_1 > 1 \quad \text{for} \quad \theta_1' \gg c\tau_1/n, \tag{34}$$

$$= \tau_1^{-1} < 1 \quad \text{for} \quad \theta_1' \ll c\tau_1/n. \tag{35}$$

The meaning of (34) and (35) is that once the author decides to increase the probability of the word w_1 by τ_1 times, this word will be τ_1 times more likely produced with the higher probabilities, and τ_1 times less likely with smaller probabilities; see (35). This is the mechanism that ensures the appearance of the keywords in the Zipfian range. It is unique to the form (25) of the marginal likelihood, which by itself is due to the form (20) of $u(\theta)$.

The feature expressed by (34) and (35) is qualitatively consistent with the fact that keywords (of a single text) tend to form clusters [40], i.e., the usage of a keyword invites its reusage on relatively short distances (no such regularity is seen for functional words).

If $P(\theta)$ is assumed to reflect the organization of the mental lexicon, then according to (34) and (35) this organization is efficient, because the decision on increasing the probability of w_1 translates to increasing the likelihood of larger values of the probability. The organization is also stable, since the likelihood at large probabilities increases right at the amount the author planned, not more.

D. The rank-frequency relation for the noninformative likelihood

Above we related the prior likelihood $P(\theta)$ to the organization of the mental lexicon. Now we would like to clarify this relation by looking at some alternative forms of the marginal likelihood. For reasons that will become apparent below, we take for such an alternative form

$$\tilde{u}(\theta) \propto (\tilde{c}n^{-1} + \theta)^{-1}, \tag{36}$$

which will produce

$$\tilde{P}(\theta) = (\tilde{c}n^{-1} + \theta)^{-1} e^{-n\theta\tilde{\mu}}. \tag{37}$$

Here $\tilde{\mu}$ is determined from

$$\int_0^\infty \frac{dy(y-1)}{\tilde{c}+y}\, e^{-\tilde{\mu}y} = 0, \tag{38}$$

by analogy to (14).

It is clear that instead of (34) and (35), we now get

$$P'(\theta_1')/P(\theta_1') = 1 \tag{39}$$

i.e., the likelihood of large probabilities does not change at all. This indicates on the lack of organization in the "mental lexicon" described by (36).

It is expected that the choice (36) does generally relate to the lack of organization or, in terms of the Bayesian statistics, to the lack of information [25,39]; see also footnote 8 in this context. This is because (36) can be considered a regularized form of the Haldane's prior likelihood $u(\theta) \propto \theta^{-1}$. There are several different arguments [including those similar to (39)] which show the Haldane's likelihood is noninformative; see Refs. [25,39] for reviews.

The rank-frequency relation generated by (37) will read by analogy to (13)

$$\frac{r}{n} = \frac{\int_{\phi_r n}^\infty \frac{dy}{\tilde{c}+y}\, e^{-\tilde{\mu}y}}{\int_0^\infty \frac{dy}{\tilde{c}+y}\, e^{-\tilde{\mu}y}}. \tag{40}$$

In the limit of a sufficiently small $\tilde{c}$, the rank-frequency relation obtained from (40) and (38) is exponential,

$$\phi_r \simeq \alpha n\, e^{-\alpha n r}, \quad \alpha = \ln(1/\tilde{c}), \tag{41}$$

instead of Zipf's law [cf. derivations (21) and (22)]. According to (41) the majority of words have negligible frequencies; hence, a small group of high-frequency words dominates the text. Intuitively, this connects well with the above statement on the lack of organization (information).

VII. CONCLUSION

We thus answer the first question asked in the introduction: Zipf's law relates to the stable and efficient organization of the mental lexicon of the text-producing author. Using the ideas of latent semantic analysis and the mental lexicon we are able to deduce the applicability of Zipf's law to a single text (cf. the fourth paragraph of Sec. I) and come up with a generalized rank-frequency relation that—besides the proper scaling regime of Zipf's law—describes hapax legomena (low-frequency, rare words), as well as the high-frequency domain (functional words).

Our derivation of Zipf's law employs ideas of Bayesian statistics, but it differs from derivations that are based on the maximum entropy method; see Refs. [12,20–24]. While the latter method looks for the most *noninformative* distribution of frequencies compatible with certain constraints, our derivation looks for a specific *informative* prior distribution of the word hidden probabilities. This difference between informational and noninformational is natural once Zipf's law is related to the mental lexicon.

Practically, our derivation of Zipf's law can motivate the usage of prior (20) in the schemes of latent semantic analysis. We expect these schemes to be more efficient for real texts, if the prior structure of the model conforms Zipf's law. The proposed methods can find applications for studying rank-frequency relations and power laws in other fields.

ACKNOWLEDGMENTS

We thank A. Galstyan and D. Manin for discussions. This work is supported by the Region des Pays de la Loire under Grant No. 2010-11967 and by the National Natural Science Foundation of China (Grants No. 10975057, No. 10635020, No. 10647125, and No. 10975062), the Programme of Introducing Talents of Discipline to Universities under

[8] Equation (29) is applied in Bayesian statistics with a similar purpose of motivating the prior likelihood [25,39]. There, however, the attention is focused on the noninformative prior likelihood that will stay invariant under (29). This is not suitable for our purpose precisely because we expect that the mental lexicon—whose organization $P(\boldsymbol{\theta})$ refers to—will somehow reflect the basic mechanism (29), i.e., $P(\theta)$ will display specific changes under (29).

Grant No. B08033, and the PHC CAI YUAN PEI Programme (LIU JIN OU [2010] No. 6050) under Grant No. 2010008104.

APPENDIX A: FITTING

1. Linear (least-squares) fitting

Here we recall the main ideas of the linear fitting method that are employed in the Sec. II of the main text.

Table I of the main text presents three texts we studied (we worked out more texts that consistently show the same applicability pattern of Zipf's law). For each text we extract the ordered frequencies of different words (the number of different words is n; the overall number of words in a text is N):

$$\{f_r\}_{r=1}^{n}, \quad f_1 \geqslant \cdots \geqslant f_n, \quad \sum_{r=1}^{n} f_r = 1. \tag{A1}$$

We should now see whether the data $\{f_r\}_{r=1}^{n}$ fits to a power law: $\hat{f}_r = cr^{-\gamma}$. We represent the data as

$$\{y_r(x_r)\}_{r=1}^{n}, \quad y_r = \ln f_r, \quad x_r = \ln r \tag{A2}$$

and fit it to the linear form $\{\hat{y}_r = \ln c - \gamma x_r\}_{r=1}^{n}$. Two unknowns $\ln c$ and γ are obtained from minimizing the sum of squared errors,

$$S_{\rm err} = \sum_{r=1}^{n} (y_r - \hat{y}_r)^2. \tag{A3}$$

It is known since Gauss that this minimization produces

$$-\gamma^* = \frac{\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}, \quad \ln c^* = \overline{y} + \gamma^* \overline{x}, \tag{A4}$$

where we defined

$$\overline{y} \equiv \frac{1}{n}\sum_{k=1}^{n} y_k, \quad \overline{x} \equiv \frac{1}{n}\sum_{k=1}^{n} x_k. \tag{A5}$$

As a measure of fitting quality one can take

$$\min_{c,\gamma}[S_{\rm err}(c,\gamma)] = S_{\rm err}(c^*,\gamma^*) \equiv S_{\rm err}^*. \tag{A6}$$

This is, however, not the only relevant quality measure. Another (more global) aspect of this quality is the coefficient of correlation between $\{y_r\}_{r=1}^{n}$ and $\{\hat{y}_r\}_{r=1}^{n}$,

$$R^2 = \frac{\left[\sum_{k=1}^{n}(y_k - \overline{y})(\hat{y}_k^* - \overline{\hat{y}^*})\right]^2}{\sum_{k=1}^{n}(y_k - \overline{y})^2 \sum_{k=1}^{n}(\hat{y}_k^* - \overline{\hat{y}^*})^2}, \tag{A7}$$

where

$$\hat{y}^* = \{\hat{y}_r^* = \ln c^* - \gamma^* x_r\}_{r=1}^{n}, \quad \overline{\hat{y}^*} \equiv \frac{1}{n}\sum_{k=1}^{n} \hat{y}_k^*. \tag{A8}$$

For the linear fitting (A4) the squared correlation coefficient is equal to the coefficient of determination,

$$R^2 = \sum_{k=1}^{n}(\hat{y}_k^* - \overline{y})^2 \Big/ \sum_{k=1}^{n}(y_k - \overline{y})^2, \tag{A9}$$

the amount of variation in the data explained by the fitting. Hence, $S_{\rm err}^* \to 0$ and $R^2 \to 1$ mean good fitting. We minimize $S_{\rm err}$ over c and γ for $r_{\rm min} \leqslant r \leqslant r_{\rm max}$ and find the maximal value of $r_{\rm max} - r_{\rm min}$ for which $S_{\rm err}^*$ and $1 - R^2$ are smaller than, respectively, 0.05 and 0.005. This value of $r_{\rm max} - r_{\rm min}$ also determines the final fitted values c^* and γ^* of c and γ, respectively; see Tables I and II and Fig. 1. Thus c^* and γ^* are found simultaneously with the validity range $[r_{\rm max}, r_{\rm max}]$ of the law. Whenever there is no risk of confusion, we, for simplicity, refer to c^* and γ^* as c and γ, respectively.

TABLE III. KS test for the numerical fitting and theoretical results. In the KS test, D and p denote the maximum difference (test statistics) and p value, respectively. D_1 and p_1 are calculated from the KS test between empiric data and numerical fitting, D_2 and p_2 are between empiric data and the theoretical result, and D_3 and p_3 are between numerical fitting and the theoretical result.

Texts	D_1	p_1	D_2	p_2	D_3	p_3
TF	0.0418	0.865	0.0365	0.939	0.0381	0.912
AR	0.0564	0.624	0.0469	0.783	0.0443	0.825
DL	0.0451	0.812	0.0421	0.865	0.0472	0.761

2. KS test for the numerical fitting and theoretical result

We wanted to have an alternative method for checking the quality of the above least-squares method and for checking to what extent our empirical data agrees with the theoretical prediction. To this end we applied the KS test to our data on the word frequencies. The empiric results on word frequencies f_r in the Zipfian range $[r_{\rm min}, r_{\rm max}]$ are fit to the power law and then also to the theoretical prediction described in Sec. IV and V. With the null hypothesis that empiric data follows the numerical fittings and/or theoretical results, we calculated the maximum differences (test statistics) D and the corresponding p values in the KS tests. From Table III one sees that all the test statistics D are quite small, while the p values are *much larger* than 0.1. We conclude that from the viewpoint of the KS test the numerical fittings and theoretical results can be used to characterize the empiric data in the Zipfian range reasonably well.

APPENDIX B: DERIVATION OF EQS. (12)–(14) OF THE MAIN TEXT.

In Eq. (11) of the main text we defined $P_r(t|T)$: the marginal density for the probability t of the word w_r. Using (8) we rewrite (11) as

$$P_r(t|T) \propto \int_0^\infty d\theta_1 \int_0^{\theta_1} d\theta_2 \int_0^{\theta_2} d\theta_3 \ldots \int_0^{\theta_{n-1}} d\theta_n \times P(\theta_1, \ldots, \theta_n)\, \delta(t - \theta_r), \tag{B1}$$

where

$$P(\boldsymbol{\theta}) \propto u(\theta_1) \ldots u(\theta_n)\, \delta\left(\sum_{k=1}^{n} \theta_k - 1\right), \tag{B2}$$

as given by (9) of the main text. Recall that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.

In Eq. (B2) we employ the Fourier representation of the δ function,

$$\delta\left(\sum_{k=1}^{n}\theta_k-1\right)=\int_{-i\infty}^{i\infty}\frac{dz}{2\pi i}\,e^{z-z\sum_{k=1}^{n}\theta_k}, \tag{B3}$$

put (B2) into (B1), and then apply integration by parts. The result reads

$$P_r(t|T)\propto u(t)\int_{-i\infty}^{i\infty}\frac{dz\,e^{z}}{2\pi i}\chi_0^{n-r}(t,z)\chi_1^{r-1}(t,z)\,e^{-tz}, \tag{B4}$$

where

$$\chi_0(t,z)\equiv\int_0^t dy\,e^{-zy}u(y),\quad \chi_1(t,z)\equiv\int_t^\infty dy\,e^{-zy}u(y).$$

The integral in Eq. (B4) will be worked out via the saddle-point method. But before that we need to fix the scales of the involved quantities. To this end, make the following changes of variables:

$$\tilde{z}=z/n,\quad \tilde{t}=tn,\quad \tilde{y}=yn,\quad \tilde{r}=r/n. \tag{B5}$$

Then $P_r(t|T)$ reads from (B4)

$$P_r(t|T)\propto u(t)\int_{-i\infty}^{i\infty}\frac{d\tilde{z}}{2\pi i}\,e^{n\varphi(\tilde{t},\tilde{z})-\tilde{t}\tilde{z}}, \tag{B6}$$

$$\varphi(\tilde{t},\tilde{z})=\tilde{z}+(1-\tilde{r})\ln\int_0^{\tilde{t}}\frac{dy\,e^{-\tilde{z}y}}{(c+y)^2}$$
$$+\left(\tilde{r}-\frac{1}{n}\right)\ln\int_{\tilde{t}}^{\infty}\frac{dy\,e^{-\tilde{z}y}}{(c+y)^2}, \tag{B7}$$

where in Eq. (B7) we already used $u(t)=(n^{-1}c+t)^{-2}$; see Eq. (20) of the main text.

If $n\gg 1$ and $0<\tilde{r}<1$ is a finite number (neither close to 1, nor to zero), the behavior of $\rho_r(t)$ in various averages, e.g., $\int dt\,t\,\rho_r(t)$, is determined by the values of $\tilde{z}=\tilde{z}_s$ and $\tilde{t}=\tilde{t}_s$ that maximize $\phi(\tilde{t},\tilde{z})$. They are found from saddle-point equations,

$$\partial_{\tilde{t}}\phi(\tilde{t}_s,\tilde{z}_s)=\partial_{\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)=0. \tag{B8}$$

After reworking the two equations (B8) we get Eqs. (13) and (14) of the main text.

Due to (B5), $\tilde{z}_s$ (that is real and positive) and $\tilde{t}_s$ stay finite for $n\gg 1$. Hence, the integration line over $\tilde{z}$ in Eq. (B6) is shifted to pass through $\tilde{z}_s$ (the saddle-point method). Now $\phi(\tilde{t},\tilde{z})$ is expanded around $\tilde{z}=\tilde{z}_s$ and $\tilde{t}=\tilde{t}_s$ [first-order terms nullify due to (B8)]:

$$\phi(\tilde{t},\tilde{z})=\phi(\tilde{t}_s,\tilde{z}_s)+\tfrac{1}{2}\partial_{\tilde{t}\tilde{t}}\phi(\tilde{t}_s,\tilde{z}_s)(\tilde{t}-\tilde{t}_s)^2, \tag{B9}$$

$$+\tfrac{1}{2}\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)(\tilde{z}-\tilde{z}_s)^2, \tag{B10}$$

$$+\partial_{\tilde{t}\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)(\tilde{t}-\tilde{t}_s)(\tilde{z}-\tilde{z}_s)+\cdots. \tag{B11}$$

Now only these terms can be retained in the integral over $\tilde{z}$. Since this integral goes over the imaginary axis, while $\tilde{z}_s$ is real, the integration contour is to be shifted to pass through $\tilde{z}_s$. For the convergence of the resulting Gaussian integral we need $\frac{1}{2}\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)>0$. Taking this Gaussian integral leads us to (up to factors that either constant or irrelevant for $n\gg 1$)

$$P_r(t|T)\propto e^{-\frac{n}{2\sigma^2}(\tilde{t}-\tilde{t}_s)^2}=e^{-\frac{n^3}{2\sigma^2}(t-\frac{\tilde{t}_s}{n})^2}, \tag{B12}$$

$$\frac{1}{\sigma^2}=\frac{[\,\partial_{\tilde{t}\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)\,]^2}{\partial_{\tilde{z}\tilde{z}}\phi(\tilde{t}_s,\tilde{z}_s)}-\partial_{\tilde{t}\tilde{t}}\phi(\tilde{t}_s,\tilde{z}_s). \tag{B13}$$

Hence $P_r(t|T)$ is approximately Gaussian, with the standard deviation $\mathcal{O}(n^{-3/2})$ much smaller than the average for $\tilde{t}_s=\mathcal{O}(1)$.

In working out (B13), we shall employ the fact that in Eq. (B7) $\tilde{z}_s=\mu$ is a small parameter; see Eq. (21) of the main text. This produces (up to smaller corrections)

$$\sigma=(c+\tilde{t}_s)\sqrt{\tilde{t}_s}. \tag{B14}$$

Equation (B12) derives (12) of the main text, while (B14) accounts for the estimate of σ_r that was presented after (12) of the main text.

APPENDIX C: DERIVATION OF EQ. (16)

The marginal probability $P(t)$ is defined from (B2) as

$$P(t)=\int d\boldsymbol{\theta}\,P(\boldsymbol{\theta})\,\delta(t-\theta_r). \tag{C1}$$

using (B2) and (B3) we obtain from (C1)

$$P(t)\propto u(t)\int_{-i\infty}^{i\infty}\frac{d\tilde{z}}{2\pi i}\,e^{n\phi(t,\tilde{z})-\tilde{t}\tilde{z}}, \tag{C2}$$

$$\phi(t,\tilde{z})=(1-t)\tilde{z}+\ln\int_0^\infty dy\,e^{-\tilde{z}y}\,(c+y)^{-2}. \tag{C3}$$

We use the saddle-point method for (C2). This produces the same saddle-point equation (B8) for $\tilde{z}_s$,

$$1=\frac{\int_0^\infty dy\,e^{-\tilde{z}_s y}\,(c+y)^{-2}}{\int_0^\infty dy\,y\,e^{-\tilde{z}_s y}\,(c+y)^{-2}}, \tag{C4}$$

provided that we note the dominant range $t\propto 1/n\ll 1$ of t in Eq. (C3). Thus

$$P(\theta)\propto u(\theta)e^{-n\theta\tilde{z}_s}. \tag{C5}$$

This validates (16) of the main text.

Likewise, one can show that the marginal density $P(\theta_1,\ldots,\theta_m)$ factorizes provided that $m\ll n$,

$$P(\theta_1,\ldots,\theta_m)\propto u(\theta_1)e^{-\mu\theta_1 n}\,\ldots\,u(\theta_m)e^{-\mu\theta_m n}. \tag{C6}$$

Equation (C6) can be established more heuristically via the exact relation $\overline{[\sum_{k=1}^{n}\theta_k]^2}=1$, where $\overline{f}$ means averaging over $P(\theta_1,\ldots,\theta_n)$. This relation predicts, together with $\overline{\theta_k}=\frac{1}{n}$, that $\overline{\theta_i\theta_j}-\overline{\theta_i}\;\overline{\theta_j}=\mathcal{O}(n^{-3})$, hence, approximate factorization.

Using (C5) with $u(\theta)=(\frac{c}{n}+\theta)^{-2}$ we note that the standard deviation $\langle(\theta-\langle\theta\rangle)^2\rangle=\frac{1}{n}\sqrt{\frac{c}{\tilde{z}_s}-1}\simeq\frac{1}{n}\sqrt{\frac{c}{\tilde{z}_s}}$ is larger than the average $\langle\theta\rangle=\int d\theta\theta\,P(\theta)=\frac{1}{n}$, since $c/\tilde{z}_s\gg 1$.

[1] R. E. Wyllys, Library Trends **30**, 53 (1981).

[2] H. Baayen, *Word Frequency Distribution* (Kluwer Academic Publishers, Dordrecht, 2001).

[3] B. Mandelbrot, *Fractal Geometry of Nature* (W. H. Freeman, New York, 1983).

[4] G. A. Miller, Am. J. Psychol. **70**, 311 (1957); W. T. Li, IEEE Inform. Theory **38**, 1842 (1992).

[5] Yu. A. Shrejder and A. A. Sharov, *Systems and Models* (Radio i Svyaz, Moscow, 1982) [in Russian]

[6] R. Ferrer-i-Cancho and R. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2003).

[7] M. Prokopenko, N. Ay, O. Obst, and D. Polani, J. Stat. Mech. (2010) P11025.

[8] R. Dickman, N. R. Moloney, and E. G. Altmann, J. Stat. Mech. (2012) P12022.

[9] V. V. Dunaev, Nauchno-Tekhnicheskaya Informatsiya **14** (1984) [in Russian]

[10] B. Corominas-Murtra, J. Fortuny, and R. V. Sole, Phys. Rev. E **83**, 036115 (2011).

[11] D. Manin, Cogn. Sci. **32**, 1075 (2008).

[12] M. V. Arapov and Yu. A. Shrejder, *Semiotics and Informatics*, Vol. 10 (VINITI, Moscow, 1978), p. 74.

[13] H. A. Simon, Biometrika **42**, 425 (1955).

[14] D. H. Zanette and M. A. Montemurro, J. Quant. Ling. **12**, 29 (2005).

[15] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[16] B. M. Hill, J. Amer. Statist. Assoc. **69**, 1017 (1974); H. S. Sichel, *ibid.* **70**, 542 (1975); G. Troll and P. beim Graben, Phys. Rev. E **57**, 1347 (1998); A. Czirok, H. E. Stanley, and T. Vicsek, *ibid.* **53**, 6371 (1996); K. E. Kechedzhi, O. V. Usatenko, and V. A. Yampol'skii, *ibid.* **72**, 6371 (2005).

[17] D. Howes, Am. J. Psyc. **81**, 269 (1968).

[18] S. Bernhardsson, S. K. Baek, and P. Minnhagen, J. Stat. Mech. (2011) P07013.

[19] R. Ferrer-i-Cancho and B. Elvevåg, PLoS ONE **5**, e9411 (2010).

[20] Yu. M. Shrejder, Prob. Peredachi Inf. **3**, 57 (1967) [in Russian].

[21] Y. Dover, Physica A **334**, 591 (2004).

[22] E. V. Vakarin and J. P. Badiali, Phys. Rev. E **74**, 036120 (2006).

[23] C.-S. Liu, Fractals **16**, 99 (2008).

[24] S. K. Baek, S. Bernhardsson, and P. Minnhagen, New J. Phys. **13**, 043004 (2011).

[25] E. T. Jaynes, IEEE Trans. Syst. Science & Cyb. **4**, 227 (1968).

[26] C. K. Hu and W. C. Kuo, *POLA Forever: Festschrift in honor of Professor S. Y. William Wang on his 70th birthday*, edited by D.-A. Ho and O. J. L. Tzeng (Academia Sinica, Taipei, 2005), Vol. 8, p. 115.

[27] S. Naranan and V. K. Balasubrahmanyan, J. Quant. Linguist. **5**, 35 (1998).

[28] A. D. Booth, Informat. Control **10**, 386 (1967); B. M. Hill, J. Am. Stat. Assoc. **65**, 1220 (1970).

[29] J. Aitchison, *Words in the Mind: Introduction to the Mental Lexicon* (Basil Blackwell Ltd, Oxford, UK, 1990).

[30] T. Hofmann, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., San Francisco, California, 1999).

[31] P. F. Lazarsfeld and N. W. Henry, *Latent Structure Analysis* (Houghton Mifflin, Boston, 1968).

[32] H. P. Luhn, IBM J. Res. Dev. **2**, 159 (1958).

[33] S. M. Gusein-Zadeh, Prob. Inform. Trans. **24**, 338 (1988).

[34] L. Pietronero, E. Tosatti, V. Tosatti, and A. Vespignani, Physica A **293**, 297 (2001); L. A. Adamic and B. A. Huberman, Glottometrics **3**, 143 (2002); R. Rousseau, *ibid.* **3**, 11 (2002).

[35] R. E. Madsen, D. Kauchak, and C. Elkan, in *Proceedings of the 22nd International Conference on Machine Learning, ICML '05* (ACM, New York, 2005), pp. 545–552.

[36] M. L. Mehta, *Random Matrices*, 3rd ed. (Elsevier Academic Press, Amsterdam, 2004).

[37] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, Behav. Brain Sci. **22**, 61 (1999); W. J. M. Levelt and A. S. Meyer, Eur. J. Cogn. Psychol. **12**, 433 (2000).

[38] C. M. McLeod and K. E. Kampe, J. Exp. Psychol. Learn. **22**, 132 (1996).

[39] M. Jaeger, Int. J. Approx. Reas. **38**, 217 (2005).

[40] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M. Somoza, Europhys. Lett. **57**, 759 (2002); P. Carpena, P. Bernaola-Galvan, M. Hackenberg, A. V. Coronado, and J. L. Oliver, Phys. Rev. E **79**, 035102 (2009).