# Maximal Diversity and Zipf's Law

Onofrio Mazzarisi,[1,2,*] Amanda de Azevedo-Lopes,[3] Jeferson J. Arenzon,[3,4] and Federico Corberi[2,5]
[1]*Max Planck Institute for Mathematics in the Sciences, Inselstrae 22, 04103 Leipzig, Germany*
[2]*Dipartimento di Fisica "E. R. Caianiello," Università di Salerno, via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy*
[3]*Instituto de Física, Universidade Federal do Rio Grande do Sul, CP 15051, 91501-970 Porto Alegre RS, Brazil*
[4]*Instituto Nacional de Ciência e Tecnologia–Sistemas Complexos, 22290-180 Rio de Janeiro RJ, Brazil*
[5]*INFN, Gruppo Collegato di Salerno, and CNISM, Unità di Salerno, Università di Salerno,*
*via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy*

(Received 16 March 2021; revised 18 June 2021; accepted 12 August 2021; published 16 September 2021)

Zipf's law describes the empirical size distribution of the components of many systems in natural and social sciences and humanities. We show, by solving a statistical model, that Zipf's law co-occurs with the maximization of the diversity of the component sizes. The law ruling the increase of such diversity with the total dimension of the system is derived and its relation with Heaps's law is discussed. As an example, we show that our analytical results compare very well with linguistics and population datasets.

DOI: 10.1103/PhysRevLett.127.128301

Diversity is a central concept in ecology, economics, information theory, and other natural and social sciences. It can be quantified by diversity indices [1,2], such as (species) richness, the Gini-Simpson index, or Boltzmann-Shannon entropy, which characterize the system under study from different angles. Loosely understanding the term, high diversity may represent an advantage in terms of resilience and performance. This is the case, for instance, in ecology, where well differentiated ecosystems are often (see, e.g., Ref. [3] for the debate on this topic) considered to be more stable [4–6], and in economy as well: strong countries have a well diversified production [7].

In most cases diversity is hindered by limiting factors. For an ecosystem the amount of energy and chemical components available does not allow an unbounded increase of the population. Similarly, the number of different items produced by an economy is limited by its strength. The diversity drift is therefore a complex optimization process.

Elaborating on that, in this Letter we take the aforementioned restrictions into account and, among the possible measures of diversity [1] we consider the richness index $D$, which turns out to be particularly suited for a quantitative description of such optimization tendency in many complex systems. Richness is a quantity that counts the number of different types which are present in a collection of items. For instance, the set of integers $\{3,7,1,9,0,1\}$ is richer than $\{3,2,3,7,7,2\}$, because there are five different figures in the former and only three in the latter. Every diversity measure can be rephrased in terms of Rényi [8] (or, equivalently, Tsallis [9]) entropies [see Ref. [1] and Supplemental Material (SM) [10]]. Notice, however, that the index $D$ alone is insensitive to the abundance of each type but only to their presence or absence.

We consider situations where types can be identified by quantitative labels $s$, as in the example above. $D$ is the richness of the collection of entities $\{s_1, \ldots, s_N\}$, with arbitrary $N$, but subjected to the additive constraint $S = \sum_{n=1}^{N} s_n$. Here, $s_n$ represents the portion of the total resource $S$ assigned to the $n$th entity of the ensemble, i.e., its size. Entities can be cities [12] of a country with total population $S$, distinct words [13] occurring with absolute frequencies $\{s_n\}$ in a book of size $S$ or genes [14] expressed with abundances $\{s_n\}$ where $S$ is the total number of proteins synthesized in a cell.

These systems are instances where the Zipf's law [15,16] is observed to hold. Other well known examples include [17] GDP of nations [18], firm sizes [19], species in taxa [20], and fragmentation processes [21]. If ranked according to their size $s$, components obey Zipf's law when

$$s(r) \propto r^{-a}, \tag{1}$$

where $r$ is the rank, with $a \simeq 1$. A representation in terms of the distribution of sizes [16,22] $p(s) \propto s^{-\tau}$, with $\tau = 1 + a^{-1}$, is better suited to our purposes. To explain Zipfian behavior many generative mechanisms have been proposed [23–29] and it has also been framed in a broader statistical perspective [30–32]. For instance, it has been shown to be associated to maximally informative samples in modeling complex systems [31,33].

In this Letter, we show that the maximization of the diversity index $D$ and the occurrence of Zipf's law in the distribution of the component sizes $\{s_n\}$ are naturally related. This is achieved by deriving, in a statistical model, a *diversity law* that can be used to estimate the index $D$ of distributions of empirical data. We put our results to the test showing remarkable agreement with data for quantitative linguistics, taken from the Project Gutenberg English texts database [34], and for urbanistics from the GeoNames database [35]. Finally, within our approach we also recover in a simple way the expression of Heaps's law [36,37] and discuss its relation with the diversity law. The fact that specifically $D$, among the possible diversity measures, is extremized, indicates the prominent role played by this quantity in the many and diverse natural phenomena described by the Zipf's law and represents a different and perhaps profitable rationalization for its occurrence.

*The model.*—Consider sets of independent and identically distributed integer random variables $\{s_n\}$, sampled from a generic probability distribution $p(s)$. We call $s_n$ the size of the $n$th component (or entity). $p(s)$ will be denoted as the *bare* distribution, since the effective (*dressed*) distribution of the $s_n$ is shaped by the presence of a global constraint $\sum_{n=1}^N s_n = S$, where $S$ is the total dimension of the system. $N$ is the fluctuating number of entities that, according to the particular extraction of the $\{s_n\}$, is needed to fulfill the constraint. The probability of a particular configuration $\mathcal{C} \equiv [\{s_1, \dots s_N\}; N]$ is given by

$$p_S(\{s_1, \dots, s_N\}; N) = \frac{1}{Z_S} \prod_{n=1}^{N} p(s_n) \delta_{\sum_{n=1}^{N} s_n, S}, \tag{2}$$

where the constraint is enforced by the Kronecker delta. The quantities $Z_S = \sum_{N=1}^{\infty} Z_S(N)$ and

$$Z_S(N) \equiv \sum_{s_1=1}^{S} \sum_{s_2=1}^{S} \cdots \sum_{s_N=1}^{S} \prod_{n=1}^{N} p(s_n) \delta_{\sum_{n=1}^{N} s_n, S} \tag{3}$$

play the role of partition functions in an ensemble where $N$ is fluctuating or fixed, respectively. One obtains the probability of having a number $N$ of entities as $p_S(N) = Z_S(N)/Z_S$. The dressed probability of observing a size $s$ can be written using Eq. (3) as

$$p_S(s) = \frac{p(s)}{\sum_{N=1}^{\infty} N Z_S(N)} \sum_{N=1}^{\infty} N Z_{S-s}(N-1), \tag{4}$$

where the factor $N$ appears because we do not distinguish among components.

If $t_s$ is the number of times the value $s \in [1, S]$ is found in a given configuration $\mathcal{C}$, the diversity index $D$ (hereafter also referred to as simply *diversity*) is defined as

$$D = \sum_{s=1}^{S} (1 - \delta_{t_s,0}), \tag{5}$$

namely the number of different values assumed by the entities. The probability $p_S(D)$ of observing a certain value of $D$ is formally given in the Supplemental Material [10].

We are interested in highly diverse configurations, therefore we consider power law bare probability distributions, which grant access to a wide range of sizes,

$$p(s) = \frac{s^{-\tau}}{\Lambda(\tau, S)}; \quad \text{for } 1 \leq s \leq S, \tag{6}$$

and $p(s) = 0$ otherwise. The normalization $\Lambda(\tau, S) = \zeta(\tau) - \zeta(\tau, S+1)$ is a generalized harmonic number and can be written in terms of the Riemann and Hurwitz zeta functions, $\zeta(x)$ and $\zeta(x, y)$ respectively.

Our goal is to compute the average diversity $\langle D \rangle_S$ and the value of $\tau$ which maximizes it (see Fig. 1). Given the complicated expression of $p_S(D)$, we directly determine $\langle D \rangle_S$ as follows. We split the range of sizes into $s \leq s^*$ and $s > s^*$ [38], where $s^*$ is defined by $\langle N \rangle_S \, p_S(s^*) = 1$; these two sectors contribute to $\langle D \rangle_S$ as

$$\langle D \rangle_S \simeq s^* + \langle N \rangle_S \sum_{s=s^*}^{S} p_S(s). \tag{7}$$

Indeed, given an average number of entities $\langle N \rangle_S$, there is at least one of them for each size $s \leq s^*$, contributing to the first term on the right-hand side of Eq. (7). The second term is the average number of entities with $s > s^*$. Since these are represented at most once this also corresponds to their contribution to $\langle D \rangle_S$.

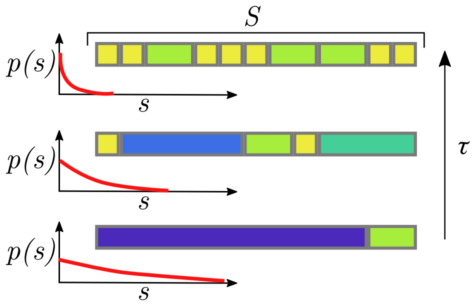

FIG. 1. Pictorial representation of the problem. Power laws, $p(s) \sim s^{-\tau}$, are sketched with an increasing exponent $\tau$ (from bottom to top) alongside with relative typical realizations $\{s_n\}$. Entities of the same size are depicted as blocks of the same color and in all the cases they add up to $S$, the total length of the bar. For large values of $\tau$, most of the entities have small and similar sizes, resulting in a poor diversity $D$. In the other limit, small $\tau$, large sizes do get more probable but the total number of entities required to fill $S$ is smaller. Consequently, diversity is again small. The diversity is expected to be maximal for an intermediate value of $\tau$.

With Eq. (7), the evaluation of $\langle D\rangle_S$ only depends on the knowledge of $\langle N\rangle_S$ and $p_S(s)$. These quantities can be computed numerically with an exact recursive method, as discussed in the Supplemental Material [10]. For an analytical treatment of the problem it is possible to approximate the dressed probability distribution with the bare one, i.e., $p_S(s) \simeq p(s)$ (see the Supplemental Material [10]). This simplification leads to an asymptotic expression for $\langle D\rangle_S$ which is accurate for large $S$. The average component size reads $\langle s\rangle_S = \sum_{s=1}^{S} s p_S(s) \simeq \sum_{s=1}^{S} s p(s) = \Lambda(\tau-1, S)/\Lambda(\tau, S)$, from which $\langle N\rangle_S$ can be obtained as $\langle N\rangle_S \simeq S/\langle s\rangle_S$. Using $\Lambda(x, S) \simeq \zeta(x) + S^{1-x}/(1-x)$ for $x \neq 0$, 1, $\Lambda(1, S) \simeq \ln S$, and $\Lambda(0, S) \simeq S$, valid for large $S$, we obtain

$$\langle N\rangle_S \simeq \begin{cases} (2-\tau)/(1-\tau); & \text{for } \tau < 1 \\ \ln\ S; & \text{for } \tau = 1 \\ \zeta(\tau)(2-\tau)S^{\tau-1}; & \text{for } 1 < \tau < 2 \\ \zeta(2)S/\ln\ S; & \text{for } \tau = 2 \\ \zeta(\tau)S/\zeta(\tau-1); & \text{for } \tau > 2, \end{cases} \tag{8}$$

which is in excellent agreement with the exact determination, see the Supplemental Material [10]. From the definition $\langle N\rangle_S p_S(s^*) = 1$, we obtain $s^*(\tau, S) \simeq [S/\Lambda(\tau-1, S)]^{1/\tau}$ and, substituting in Eq. (7), one arrives at the sought-after result for the average diversity: $\langle D\rangle_S \simeq s^* + (s^*)^\tau[\zeta(\tau, s^*) - \zeta(\tau, S+1)]$. Approximating the Riemann zeta function by $\zeta(x) \simeq (x-1)^{-1} + \gamma$, where $\gamma \simeq 0.577$ is the Euler constant, we can write

$$s^*(\tau, S) \simeq S^{1/\tau}[\gamma + (S^{2-\tau} - 1)/(2-\tau)]^{-1/\tau}, \tag{9}$$

$$\langle D\rangle_S \simeq \frac{\tau s^* - (s^*)^\tau S^{1-\tau}}{\tau - 1}, \tag{10}$$

where the appropriate limits for $\tau = 1$ and 2 are taken.

This determination of $\langle D\rangle_S$ is portrayed in Fig. 2 and compared with the outcome of numerical simulations finding a very good agreement. For large $S$, the leading contribution to Eq. (10) is

$$\langle D\rangle_S \simeq \begin{cases} (2-\tau)/(1-\tau); & \text{for } \tau < 1 \\ \ln\ S; & \text{for } \tau = 1 \\ \frac{\tau(2-\tau)^{1/\tau}}{\tau-1} S^{1-1/\tau}; & \text{for } 1 < \tau < 2 \\ 2(S/\ln\ S)^{1/2}; & \text{for } \tau = 2 \\ \frac{\tau}{\tau-1}\left[\frac{S}{\gamma+(\tau-2)^{-1}}\right]^{1/\tau}; & \text{for } \tau > 2. \end{cases} \tag{11}$$

One has $\langle D\rangle_S \sim S^{\alpha(\tau)}$ with $\alpha(\tau) = 0$ for $\tau < 1$, $\alpha(\tau) = 1 - 1/\tau$ for $1 < \tau < 2$, and $\alpha(\tau) = 1/\tau$ for $\tau > 2$, see inset of Fig 2. In conclusion, for large $S$, $\langle D\rangle_S$ presents a pronounced peak at $\tau = 2$. This behavior is due to the competition between the abundance of entities $\langle N\rangle_S$, favored by large $\tau$, and the diversity of their sizes which instead is enhanced by small $\tau$, as shown in Fig. 1. We remark that the upper bound obtained by considering the deterministic partition $S \simeq 1 + 2 + \cdots + D$ with $D \sim S^{1/2}$ overpowers the $\tau = 2$ case only by a logarithmic factor.

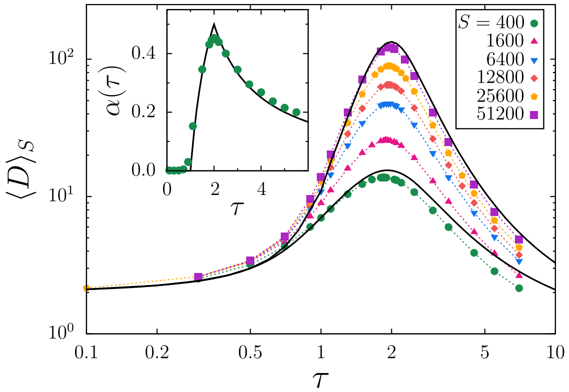

FIG. 2. Average diversity $\langle D\rangle_S$, obtained through numerical simulations for various sizes $S$ (see key), dashed lines are guides to the eye. Entities are extracted from the bare distribution Eq. (6) and the statistics is restricted over configurations respecting the global constraint. Results are averaged over $10^4 - 10^6$ configurations. Solid lines (shown only for the extreme sizes), are the analytical solutions given by Eq. (10). Inset: The exponent $\alpha(\tau)$, defined below Eq. (11), as a function of $\tau$. Solid line is the analytical result, dots are fits from the simulation data.

Let us mention that, although we explicitly solved the model for power law distributions, which yield maximum diversity, our calculations can be straightforwardly generalized to different $p(s)$. For instance, in the case of algebraic distributions with a lower cutoff, a case often representative of real situations [39], one recovers similar results provided that the cutoff is independent of $S$ (see the Supplemental Material [10]).

We also stress that, as shown in the Supplemental Material [10], among the possible measures of diversity usually considered in the literature, $D$ is the only one to be maximized in connection with Zipf's law.

We notice also that the model considered here is related to the random allocation model [40] where the resource $S$ is distributed among an *assigned* number $N$ of components. The diversity properties of such a model, however, are very different and, in particular, the special role played by $\tau = 2$ is missing. This is briefly discussed in the Supplemental Material [10].

*Diversity, Zipf's, and Heaps's laws.*—Since the diversity is determined once an empirical distribution of sizes is given, we can use $\langle D\rangle_S$ given in Eq. (11) to estimate the diversity index $D$ of power law distributed empirical data, regardless of the mechanism whereby they are produced. If this assumption holds, on the basis of our analytical arguments, one can conclude that if a system displays

Zipf's law ($\tau \simeq 2$) it is at the edge of maximal diversity and vice versa.

As a first example we consider quantitative linguistics, the field in which Zipf's law has been originally observed in almost every human language [13,41–43]. The regime of validity of the law in this context [44], its deviations [45], and the underlying mechanism(s) are still a matter of dispute. Nonetheless, large scale studies have been performed in order to validate that. For example, Moreno-Sánchez *et al.* [43] considered a very large set of English books (more than 30 000) from the Project Gutenberg database. They checked how well some simple, one-parameter forms of the Zipf's law describe these data on the whole interval of frequencies, finding very good agreement with a distribution of exponents centered on $\tau \simeq 2$.

We use the filtered data of Ref. [43] and, for each book, measure the diversity index $D$. The total number of words a book contains is its total size $S$, the number of distinct words is the number of entities, $N$, and the size $s$ of each entity is its absolute frequency, i.e., how many times that word appears. The diversity $D$ is therefore the number of *different frequencies* a given text displays. The result of this analysis is shown in Fig. 3 along with Eq. (11) for $\tau = 2$. Notice that there are no free parameters in the plot. The agreement between our theoretical prediction and the experimental points is consistent with the results reported in Ref. [43] showing that a great deal of the books have $\tau$ close to 2.

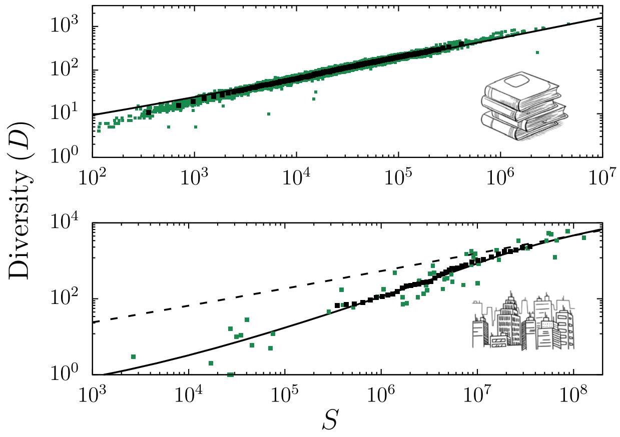

FIG. 3. Top panel: diversity index $D$ evaluated from the data of Ref. [43]. Each green point is one of the more than 30 000 English books in the Project Gutenberg database (accessed July 2014), while the black squares are a running average over 20 points. The solid line is the result $\langle D \rangle_S = 2(S/\ln S)^{1/2}$, from Eq. (11) for $\tau = 2$, which corresponds to maximal diversity. Bottom panel: diversity index using data from the GeoNames database [35] for cities. Each green point is a European country and the black squares are the corresponding running average. The solid line is Eq. (18) of the Supplemental Material [10], where the presence of a lower cutoff $s_L$ is taken into account. $s_L$ is estimated from the average of the smallest city in each country ($s_L \simeq 1313$), see Supplemental Material [10]. The dashed line is the behavior $\langle D \rangle_S = 2(S/\ln S)^{1/2}$, which is approached only asymptotically.

As a second example, we consider how the total population $S$ of a country is distributed among its cities. We use data for European countries from the GeoNames database [35], for which Simini and James [46] showed that the size $s$ of cities closely follows a Zipf's distribution ($\tau \simeq 2.02$). The diversity index $D$ is shown in Fig. 3 (bottom panel). Since cities cannot be smaller than a certain lower cutoff $s_L$, the analytical prediction to compare with is Eq. (18) of the Supplemental Material [10] (solid line). Despite the noisy character of the data, there is a very good agreement between the data and our theory.

The content of Eq. (8) is Heaps's law, which gives an estimate of the number of components of a system of total size $S$ given that the empirical size distribution follows a power law with exponent $\tau$. Our expression of the law for $\tau > 1$ is in accordance with Ref. [37] and complements the result with the cases with $\tau \leq 1$ and with the appropriate prefactors. Heaps's law is expected to hold for systems which are robust in the statistics of their component [$p_S(s)$ in our notation] at varying $S$ [37,39]. This is captured in our approach, where Eq. (8) is only arrived at using distributions which have the same form for any $S$ [the same applies to Eq. (11)].

In our approach, Heaps's law (8) and the diversity law (11) imply each other, encoding dependencies on the system size on equal footings. However, notably, the latter naturally selects the exponent $\tau = 2$ as a special one. Moreover, our analysis of the Gutenberg dataset shows that the diversity law is obeyed up to the largest sizes considered ($S \simeq 10^7$), whereas it is known [47] that strong deviations from Heaps's law are caused by the finiteness of the vocabulary. Therefore, at least in the context of language, the diversity law appears more robust and this suggests that its use could be more suited to interpret the size dependence of empirical data.

*Discussion.*—The partition of a finite resource $S$ among constituents informs numerous systems in diverse fields of science and humanities. In this Letter, by solving a paradigmatic statistical model, we have shown that a maximally diverse partition is accompanied by Zipf's law. Such co-occurrence is a general property of the empirical distribution, holding irrespectively of the specific mechanisms at work in generating Zipfian behavior in given systems.

Diversity and information are fundamental concepts for the description of complex statistical systems whose formalization led to the definition of a coherent set of quantitative measures, Boltzmann-Shannon entropy above all. Our results show that in the case of system obeying Zipf's law an important role is played by one of such measures, the index $D$. When framed in terms of extremization of appropriate cost functions, problems are endowed with a complementary description and can be approached with new strategies. Our study suggests that, in some instances where Zipf's law is empirically observed, promoting diversity to the role of a driving force could provide further theoretical insights towards a deeper and more general comprehension.

O. M. is indebted with I. A. Hatton, M. Smerlak, and A. Zadorin for numerous and insightful discussions and acknowledges the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the German Federal Ministry of Education and Research for providing funding for this work. A. A. L. and J. J. A. thank Salerno University for hospitality. A. A. L acknowledges the Brazilian funding agency CAPES in the framework of the Capes-PrInt program (Grant No. 88887.466912/2019-00). J. J. A. thanks the Brazilian funding agency CNPq (Grant No. 308927/2017-6). The authors thank S. Bora for the drawings in Fig. 3.

*mazzaris@mis.mpg.de

[1] L. Jost, Oikos **113**, 363 (2006).

[2] H. Tuomisto, Oecologia **164**, 853 (2010).

[3] A. R. Ives and S. R. Carpenter, Science **317**, 58 (2007).

[4] C. S. Elton, *The Ecology of Invasions by Animals and Plants* (Methuen & Co. Ltd., London, United Kingdom, 1958).

[5] D. Tilman, P. B. Reich, and J. M. Knops, Nature (London) **441**, 629 (2006).

[6] F. Arese Lucini, F. Morone, M. S. Tomassone, and H. A. Makse, PLoS One **15**, e0228692 (2020).

[7] A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero, Sci. Rep. **2**, 723 (2012).

[8] A. Rényi *et al.*, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (The Regents of the University of California, Berkeley, 1961), Vol. 1.

[9] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).

[10] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.128301 for an account of Rényi entropies, their connection with diversity indices and arguments for studying specifically the diversity index $D$ considered in this paper based on numerical simulations, an explicit expression for the probability distribution of the diversity $p_S(D)$, an exact computation of the dressed probability distribution $p_S(s)$ and $p_S(N)$, motivations for the approximation $p_S(s) \simeq p(s)$, the case of power law bare distributions with a lower cutoff $s_L$ blue and details of the analysis of population datasets, and an account of the behaviour of diversity in the random allocation model, which includes Ref. [11]

[11] F. Corberi, Phys. Rev. E **95**, 032136 (2017).

[12] X. Gabaix, Q. J. Econ. **114**, 739 (1999).

[13] S. T. Piantadosi, Psychon. Bull. Rev. **21**, 1112 (2014).

[14] C. Furusawa and K. Kaneko, Phys. Rev. Lett. **90**, 088102 (2003).

[15] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA, 1949).

[16] M. E. J. Newman, Contemp. Phys. **46**, 323 (2005).

[17] A. Clauset, C. R. Shalizi, and M. E. Newman, SIAM Rev. **51**, 661 (2009).

[18] M. Cristelli, M. Batty, and L. Pietronero, Sci. Rep. **2**, 812 (2012).

[19] R. L. Axtell, Science **293**, 1818 (2001).

[20] J. C. Willis and G. U. Yule, Nature (London) **109**, 177 (1922).

[21] L. Oddershede, P. Dimon, and J. Bohr, Phys. Rev. Lett. **71**, 3107 (1993).

[22] A. Corral, I. Serra, and R. Ferrer-i-Cancho, Phys. Rev. E **102**, 052113 (2020).

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

[24] M. Levy and S. Solomon, Int. J. Mod. Phys. C **07**, 595 (1996).

[25] M. Marsili and Y.-C. Zhang, Phys. Rev. Lett. **80**, 2741 (1998).

[26] R. Ferrer-i-Cancho and R. V. Solé, Proc. Natl. Acad. Sci. U.S.A. **100**, 788 (2003).

[27] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz, Sci. Rep. **4**, 5890 (2014).

[28] B. Corominas-Murtra, R. Hanel, and S. Thurner, Proc. Natl. Acad. Sci. U.S.A. **112**, 5348 (2015).

[29] A. Mazzolini, M. Gherardi, M. Caselle, M. Cosentino Lagomarsino, and M. Osella, Phys. Rev. X **8**, 021023 (2018).

[30] T. Mora and W. Bialek, J. Stat. Phys. **144**, 268 (2011).

[31] M. Marsili, I. Mastromatteo, and Y. Roudi, J. Stat. Mech. (2013) P09003.

[32] D. J. Schwab, I. Nemenman, and P. Mehta, Phys. Rev. Lett. **113**, 068102 (2014).

[33] R. J. Cubero, J. Jo, M. Marsili, Y. Roudi, and J. Song, J. Stat. Mech. (2019) 063402.

[34] Project Gutenberg, www.gutenberg.org.

[35] GeoNames, www.geonames.org.

[36] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, Inc., Orlando, FL, 1978).

[37] L. Lü, Z.-K. Zhang, and T. Zhou, PLoS One **5**, e14139 (2010).

[38] A. de Azevedo-Lopes, A. R. de la Rocha, P. M. C. de Oliveira, and J. J. Arenzon, Phys. Rev. E **101**, 012108 (2020).

[39] G. De Marzo, A. Gabrielli, A. Zaccaria, and L. Pietronero, Phys. Rev. Research **3**, 013084 (2021).

[40] C. Godrèche, J. Stat. Mech. (2019) 063207.

[41] E. U. Condon, Science **67**, 300 (1928).

[42] M. Gerlach and E. G. Altmann, New J. Phys. **16**, 113010 (2014).

[43] I. Moreno-Sánchez, F. Font-Clos, and A. Corral, PLoS One **11**, e0147073 (2016).

[44] F. Font-Clos, G. Boleda, and A. Corral, New J. Phys. **15**, 093033 (2013).

[45] R. Ferrer-i-Cancho, Eur. Phys. J. B **44**, 249 (2005).

[46] F. Simini and C. James, EPJ Data Sci. **8**, 24 (2019).

[47] L. Lü, Z.-K. Zhang, and T. Zhou, Sci. Rep. **3**, 1082 (2013).