The distinct flavors of Zipf's law and its maximum likelihood fitting: rank-size and size-distribution representations

Álvaro Corral,[1, 2, 3, 4] Isabel Serra,[1] and Ramon Ferrer-i-Cancho[5]

[1]Centre de Recerca Matemàtica, Edifici C,
Campus Bellaterra, E-08193 Barcelona, Spain

[2]Departament de Matemàtiques, Facultat de Ciències,
Universitat Autònoma de Barcelona, E-08193 Barcelona, Spain

[3]Barcelona Graduate School of Mathematics, Edifici C,
Campus Bellaterra, E-08193 Barcelona, Spain

[4]Complexity Science Hub Vienna, Josefstädter Straße 39, 1080 Vienna, Austria

[5]Complexity and Quantitative Linguistics Lab,
Departament de Ciències de la Computació,
Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

Abstract

In the last years, researchers have realized the difficulties of fitting power-law distributions properly. These difficulties are higher in Zipfian systems, due to the discreteness of the variables and to the existence of two representations for these systems, i.e., two versions depending on the random variable to fit: rank or size. The discreteness implies that a power law in one of the representations is not a power law in the other, and *vice versa*. We generate synthetic power laws in both representations and apply a state-of-the-art fitting method to each of the two random variables. The method (based on maximum-likelihood plus a goodness-of-fit test) does not fit the whole distribution but the tail, understood as the part of a distribution above a cut-off that separates non-power-law behavior from power-law behavior. We find that, no matter which random variable is power-law distributed, using the rank as the random variable is problematic for fitting, in general (although it may work in some limit cases). One of the difficulties comes from recovering the "hidden" true ranks from the empirical ranks. On the contrary, the representation in terms of the distribution of sizes allows one to recover the true exponent (with some small bias when the underlying size distribution is a power law only asymptotically).

Keywords: Statistical inference; Scaling in socio-economic systems; Zipf's law.

I. INTRODUCTION

Power-law distributions are supposed to show up in the statistics of many complex systems [1–6]. However, the fitting of power-law distributions, or more precisely, power-law tails, is a delicate issue [7–9]. Traditionally, statistical physicists and complex-systems scientists have treated power laws with little statistical rigor, using the well-known fact that a power law in a double-logarithmic plot looks like a straight line. In the last 15 years, several researchers have advocated for the use of maximum-likelihood methods, which have solid statistical foundations [10–12] and are not difficult to implement, at least for continuous pure power-law distributions.

Nevertheless, real data do not show power-law behavior for the whole range of the random variable, but only (if any) above a certain (lower) cut-off, i.e., for the tail. The critical issue is then to determine where the power-law tail starts. Clauset and coauthors [12] proposed a recipe that consists in examining all possible values of the cut-off, selecting the one that yields the minimum value of the Kolmogorov-Smirnov statistic. Although this method has been widely used since its introduction, several authors have criticized it, on the basis of several problems: (i) The method is *ad hoc* (there is no justification on why the true value of the cut-off has to have the property of minimizing the Kolmogorov-Smirnov statistic [13]). (ii) The method cannot be successfully extended to (upper-) truncated power-law distributions. (iii) The methods fails in some simple "controlled experiments" (with several simulated data sets with true power-law tails [14–16]). In consequence, subsequent authors have proposed different alternative methods of power-law-tail fitting [13, 16, 17].

Zipfian systems constitute a very special but important area in complex-systems science where power-law fitting is particularly involved, due to the existence of two possible random variables to fit, both of discrete (non-continuous) nature. Since its apparent first discovery in linguistics by Estoup, more than 100 years ago [18], and subsequent rediscovery by Condon [19] (and finally by Zipf [20]), Zipf's law has emerged as a paradigm in the statistics of social, economic, cognitive, biological, and technological processes. Indeed, this law, of empirical nature, has been proposed to hold in many different complex systems [5, 21–25].

Zipfian systems can be described in the following simple way. Let us consider a system composed by some entities, which we call types, and that each of these types can be characterized by a certain discrete "size"; further, each elementary unit of size will constitute a token. If the system is a text, each appearance of a word is a token corresponding to the type given by the word itself; then, the size of a type will be its number of appearances (i.e., its absolute frequency) [26–28]. For a country, the tokens will be its citizens, whereas the types can be the cities where they live [29], or their family names [30], etc.; the size of each type will be the population associated to it. Another possibility is that the tokens are monetary units (let us say, richness translated into one-dollar pieces) and the types are the persons owning that money; the measure of the "size" of one person will be, in these terms, his/her richness (for other examples see, e.g., Ref. [31] and Table I).

Zipf's law deals with how these tokens are distributed into types. Counting the number of tokens that correspond to any type allows one not only to define the size n of the types but also their rank r, which is the position of each type when these are ordered by decreasing size. Then, $n(r)$ is the number of tokens of the type with rank r. If several types have the same size (which is common for low sizes), the assignation of distinct ranks to them is arbitrary. For instance, in the book *Ulysses* (by James Joyce), the rank $r = 1$ corresponds to type *the* (as this is the most common word type); for the population of USA cities, $r = 1$ corresponds to New York; and for USA wealth, the person ranked first is William Henry Gates III; the size of these types is $n(r = 1) = 14,934$ appearances, 8,622,698 inhabitants, and 95×10^9 dollars, respectively (at the time of writing this article for the USA cities and the USA wealth).

The dependence between size n and rank r (necessarily non-increasing) yields the rank-size (or size-rank) or rank-frequency relation, and a first form of Zipf's law holds when both n and r are related through a decreasing power law, i.e.,

$$n(r) \propto \frac{1}{r^\alpha}, \qquad (1)$$

with exponent α more or less close to one, and the symbol "$\propto$" denoting proportionality (not asymptotic behavior). This formulation will be referred to as Zipf's law for types, as it is obtained from the statistics of types (counting their repetitions, i.e., their tokens).

An equivalent description of this sort of systems is possible in terms of the distribution of sizes (or distribution of frequencies). For that, one counts not only the repetitions of each type (i.e., its size n) but also the repetitions of each size, i.e., one counts the number of types with a given size n [27, 30]. In probabilistic terms, this means that the size of the types is considered as the random variable (that is the target of the statistics). Then, $f(n)$, the

TABLE I: Diverse examples of Zipfian systems, together with their corresponding tokens and types. Note that we do not claim that a Zipfian system has to fulfill Zipf's law (in any of the two forms considered in this article). The horizontal lines roughly separate different sorts of Zipfian systems: Repetitions of identical tokens (1st and 2nd cases, up to row 5); gathering of different tokens into common groups (3rd case); network-like systems (4th); and systems whose tokens are composed by merging simpler tokens (5th). RW denotes a random walk.

Sort of system	System/discipline	Token	Type
1st	Texts	word occurrence	word itself
	Musical pieces	chord occurrence	chord itself
	Bibliometry	citation	paper cited
	RW in networks	visit to a site	site
2nd	Cells	count of molecule	molecule itself
3rd	Sociology	individual believer	religion
	Demography	individual inhabitant	city
	Demography	individual	family name
	Bibliometry	paper authored	author
	Bibliometry	papers on a topic	journal
	Ecology	individual insect	hosting plant
	Economics	employee	company
	Economics	pieces of $1 (monetary unit)	individual wealth-holder
4th	Networks	link	node
	Telephony	call	customer
	Internet	connection	computer
5th	RW in networks	transition occurrence	transition itself
	Texts	bigram occurrence	bigram itself

probability mass function of n, is the quantity of interest. Zipf's law for sizes holds when

$$f(n) \propto \frac{1}{n^\gamma}, \tag{2}$$

i.e., when $f(n)$ follows a power law.

Many authors have argued or assumed that both forms of Zipf's law, Eq. (1) and Eq. (2), are equivalent [5, 22, 30], with a well-known relation between their exponents given by

$$\gamma = 1 + \frac{1}{\alpha}, \tag{3}$$

but the equivalence only holds exactly asymptotically, for large n [32]. Note that Zipf's law is expected to hold for large sizes (in texts, going deep into less common words, a different power-law regime appears, although the previous relation still holds empirically [33]). We will turn to this important issue below.

In the second version of the law, it may seem strange that one needs to perform double statistics – first the statistics of types, counting tokens to obtain the size of every rank (i.e., of every type), and then the statistics of sizes, counting types, to obtain the number of types of a given size. This is indeed the case of words, where the frequency obtained counting tokens plays the role of the random variable, which needs to be counted also, within this framework. In contrast, in other systems where Zipf's law is supposed to hold, such as cities, Zipf's law can be obtained directly from the statistics of the sizes n of the studied entities, which are usually precomputed (we do not need to go city by city counting all their inhabitants, as a census usually does that work for us). Nevertheless, this does not constitute a fundamental difference between both kinds of systems, and the only difference is in the way the data are usually available (see Ref. [34] for a counterexample).

In order to dispel any misunderstanding, it is useful to clarify what $n(r)$ and $f(n)$ mean in practice. If one picks a token randomly from a system (e.g., a person from the census), the probability that it corresponds to the type (to the city) with rank r is given by $n(r)/L$ (where L is the total number of tokens, i.e., the sum of the sizes of all types).

Knowledge of $n(r)$ allows one to build a *directory* (a list of cities, a dictionary...) with the sizes $n(r)$ of all types; then, if one picks a type uniformly at random from the directory (a city from the list of cities), the probability that it has size n in the system is $f(n)$. There is still a third distribution, given by $nf(n)/\langle n \rangle$ (if $\langle n \rangle$, the mean of n, is finite), which represents the probability that, if one picks a token at random from the system (a person from the census), it corresponds to any type (any city) of size n. As this latter distribution is directly related to $f(n)$, it will not be considered in this article.

In this article, we address the question about the best approach to verify the concordance of empirical data with Zipf's law, taking into account that Zipf's law can be understood either as a power-law relation between rank and size, Eq. (1), or as a power-law distribution of sizes, Eq. (2). By power law we specifically mean non-truncated power laws, i.e., power laws without an upper truncation or upper cut-off. These are "genuine" or pure power laws, in the sense that they fulfill the peculiar properties of scale invariance and divergence of moments, in contrast to truncated power laws, which will be left to future research. Let us stress that although pure power-law tails may be considered an idealization, they constitute an excellent model for many systems, in the sense that rigorous statistical testing cannot refute their validity (e.g., [17, 29, 35]).

In addition, note that, despite the fact that the two representations of Zipfian systems are equivalent (one can obtain any of the two from the information contained in the other) a power law in one representation is not a power law in the other (and thus we can refer to the different "flavors" of each power-law representation); this will become clearer throughout the article. Our distinction between the two representations holds also for other distributions that have more parameters than power laws (e.g., Ref. [36]); this will be left for future research too.

Taking for granted that the most suitable way to fit a power-law probability distribution (or any other "well behaved" distribution) [10–12] is maximum likelihood (ML) estimation [37] (see also Appendix A), we will apply this method both to the rank-size relation $n(r)$ and to the distribution of sizes $f(n)$ for simulated systems. We will use a state-of-the art fitting procedure [13] adapted to discrete distributions, which, in addition to ML estimation, also incorporates a goodness-of-fit test (the testing is necessary in order to evaluate the goodness or badness of the ML fit; ML estimation does not provide goodness of fit). The fitting procedure is aimed at improving Clauset et al.'s well-known method [12]. Note that, although Clauset et al.'s method applies only to non-truncated power laws, our alternative [13] can be applied to both non-truncated and truncated power laws; nevertheless, as stated, the object of the present article are non-truncated power laws.

As the two definitions of Zipf's law (for types and for sizes) are not really equivalent, we simulate random systems for the two versions of the law (and, as mentioned, we study each system both using $n(r)$ and $f(n)$). This yields four different cases of study, which are further doubled when one distinguishes between continuous and discrete distributions. In quantitative linguistics, the overwhelming majority of research has focused on rank as the random variable [19, 26, 30, 31, 38–44] (some exceptions are Refs. [33, 45–48]). Here we argue that the alternative track of fitting the distribution of sizes has some clear advantages over ranks and is much more appropriate if one wishes to benefit from the virtues of maximum-likelihood estimation. Note that Clauset et al. [12] also fitted the distribution of sizes, but without explicit mention of its alternatives and the possible problems.

The remainder of the article is organized as follows. Section II presents the frameworks used for the (probabilistic) description of Zipf's law and explains why the two representations of this law are not equivalent. Section III deals with the problem of recovering the true parameters from simulations of the two versions of the law using maximum likelihood estimation; the advantages of maximum-likelihood estimation and its practical application are briefly explained. Section III discusses the results highlighting the advantages of the distribution of sizes. Finally, the complete fitting procedure for the discrete case, including how to simulate Zipf's law (which is not totally straightforward), is explained in Appendices A and B. This article can be considered a complement or another option to the approach of Ref. [49].

II. DIFFERENCES BETWEEN ZIPF'S LAWS FOR TYPES AND FOR SIZES

In order to proceed, we need two useful quantities: the number of tokens L, also referred to as size of the system, and the number of types V (vocabulary for a text). These are empirical quantities related through

$$L = \sum_{r=1}^{V} n(r).$$

Of course, $V \le L$, and in any non-trivial case, $V < L$. It is important to mention that in our analysis we will not consider data outside the power-law range (to be determined by the fitting procedure), and therefore, V and L do not correspond to the complete empirical data but to a restricted, truncated data set. For the complete (total) data set, we will use the notation V_{tot} and L_{tot}.

A. Zipf's law for types

Let us now assume that Zipf's law for types, Eq. (1), holds empirically,

$$n(r) = \frac{A}{r^{\alpha}}, \tag{4}$$

with A being an appropriate normalization constant scaling linearly with system size. It turns out that Zipf's law for types can be written as

$$S(n) = \frac{B}{n^\beta}, \tag{5}$$

with B some constant ($B = A^\beta/V$), $\beta = 1/\alpha$ and $S(n)$ the complementary cumulative distribution of the size, also called survivor function, i.e., the probability that the size of a type equals or exceeds a particular value n; formally, $S(n) = \text{Prob}\,[\text{size} \geq n]$. Indeed, by definition, $S(n)$ can be estimated as $S(n) = r/V$ [5, 50], and if n is not too low the estimated $S(n)$ will be very close to the true $S(n)$; then, the inversion of Eq. (4) leads directly to Eq. (5) [22, 30]. When several types have the same size, the r used in the calculation of $S(n)$ has to be the one with the largest value among those types (due to our definition of $S(n)$, which contains the "$\geq$" inequality). We can obtain the corresponding probability mass function of n as $f(n) = S(n) - S(n+1)$ (from the definition $f(n) = \text{Prob}\,[\text{size} = n]$); then,

$$f(n) = B\left(\frac{1}{n^\beta} - \frac{1}{(n+1)^\beta}\right) = \frac{B}{n^\beta}\left[1 - \left(\frac{1}{1+1/n}\right)^\beta\right].$$

Using the binomial theorem and the geometrical series we can obtain the asymptotic behavior of $f(n)$,

$$f(n) \simeq \frac{B}{n^\beta}\left[1 - \left(1 - \frac{\beta}{n} + \frac{\beta(\beta+1)}{2n^2} + \cdots\right)\right] = \frac{\beta B}{n^{\beta+1}}\left(1 - \frac{\beta+1}{2n} + \cdots\right).$$

A similar result has been obtained for $\beta = 1$ in Refs. [26, 51]. Thus, this simple example shows that Zipf's law for types, Eq. (4), leads to a power-law distribution $f(n) \propto 1/n^\gamma$ only for infinitely large n, with

$$\gamma = 1 + \beta = 1 + \frac{1}{\alpha}. \tag{6}$$

This relation between exponents is well-known.

What we have shown here is that a pure discrete power-law form for $n(r)$, Eq. (4), does not lead to a pure power law for the probability mass function of the size, $f(n)$, in the sense that although the power law is fulfilled exactly for the rank-size relation, it will not hold exactly for $f(n)$, only asymptotically. This issue has received very little attention in the literature, as one is usually interested in the fulfillment of Zipf's law in an almost qualitative sense, for instance just by plotting the logarithm of either $n(r)$, $S(n)$, or $f(n)$ versus $\log r$ or $\log n$ and obtaining something reminiscent of a straight line in some part of the plot (then, the distinction between a pure and an asymptotic power law becomes diluted).

A notable exception is provided by Mandelbrot [32], who calculated $f(n)$ when L_{tot} tokens are drawn randomly and independently of each other from Zipf's law for types, Eq. (4), with an infinite population ($V \to \infty$). We explain this construction in Sec. III B. In contrast to the previous case, Zipf's law in the form of Eq. (4) is supposed to hold not only for a single empirical sample of the system but for the underlying population. Mandelbrot's result, for large L_{tot}, was

$$f(n) = \frac{\beta A^\beta}{V_{tot}} \frac{\Gamma(n-\beta)}{\Gamma(n+1)}, \tag{7}$$

for $n = 1, 2, \ldots,$ with Γ the gamma function [52]. For large n, the quotient of gamma functions tends to $1/n^{\beta+1}$ (using Stirling's approximation), and again, one gets a power law for $f(n)$ only asymptotically. A normalized version of Mandelbrot's $f(n)$ is

$$f(n) = \frac{\beta}{\Gamma(1-\beta)} \frac{\Gamma(n-\beta)}{\Gamma(n+1)}, \tag{8}$$

and so we find $V_{tot} = \Gamma(1-\beta)A^\beta$, that is essentially Heaps' law [26, 50], also called the type-token relationship [32] (recall that A has to scale linearly with system size, i.e., $A = L_{tot}/\zeta(\alpha)$ when $V \to \infty$, with $\zeta(\alpha)$ the Riemann zeta function). In terms of $S(n)$, one gets

$$S(n) = \frac{\Gamma(n-\beta)}{\Gamma(1-\beta)\Gamma(n)},$$

which, again, is only a power law asymptotically. Note that this distribution is different from the so-called Yule (or Yule-Simon) distribution [12, 53].

Table II clarifies the different options for the different versions of Zipf's law and their different representations.

B. Zipf's law for sizes

Nevertheless, strictly speaking, for a random variable n (size), a discrete power-law distribution is defined in terms of $f(n)$ and neither in terms of $S(n)$ (although there seems to be some confusion, as in Ref. [54]) nor in terms of its underlying rank-size relation. Thus, when considering the size of types as a discrete random variable, a power-law distribution would mean that the size probability mass function is given by

$$f(n) = \frac{1}{\zeta(\gamma, n_a) n^\gamma}, \tag{9}$$

TABLE II: Different versions of Zipf's law in different representations. Each column is a different model and each row is a different representation (equivalent for each model); the horizontal line separates the rank-size representation (given by $n(r)/L$) from the distribution-of-sizes representation (given by $S(n)$ and $f(n)$). n is size (or frequency), L is total size (text length), r is rank, $S(n)$ is complementary cumulative distribution of n, $f(n)$ is probability mass function of n, ζ is Hurwitz zeta function, ζ_2^{-1} is the inverse of ζ with respect its second argument, and Γ is the gamma function. Both $S(n)$ and $f(n)$ are properly normalized, which yields an $n(r)/L$ that is only an approximation of the empirical value, except for the second column, where $n(r)/L$ is a proper normalized distribution. In the first column, n goes to ∞ in the limit $L \to \infty$ (i.e., $A \to \infty$) and $V \to \infty$. These distributions were also used in Ref. [35].

	Zipf's law for types	Mandelbrot's version [32]	Zipf's law for sizes
$n(r)/L$	$\frac{A}{L}\left(\frac{1}{r}\right)^{1/\beta}$ $1 \le r \le V$	$\frac{1}{\zeta(\beta^{-1},1)}\left(\frac{1}{r}\right)^{1/\beta}$ $r = 1, 2 \dots \infty$	$\frac{1}{L}\zeta_2^{-1}\left(\beta+1, \frac{\zeta(\beta+1,n_a)}{V}r\right)$ $r \le V$
$S(n)$	$\frac{1}{V}\left(\frac{A}{n}\right)^{\beta}$ $n = \frac{A}{V^{1/\beta}} \dots \infty$	$\frac{\Gamma(n-\beta)}{\Gamma(1-\beta)\Gamma(n)}$ $n = 1, 2 \dots \infty$	$\frac{\zeta(\beta+1,n)}{\zeta(\beta+1,n_a)}$ $n = n_a, n_a + 1 \dots \infty$
$f(n)$	$\frac{1}{V}\left[\left(\frac{A}{n}\right)^{\beta} - \left(\frac{A}{n+1}\right)^{\beta}\right]$	$\frac{\beta}{\Gamma(1-\beta)}\frac{\Gamma(n-\beta)}{\Gamma(n+1)}$	$\frac{1}{\zeta(\beta+1,n_a)}\left(\frac{1}{n}\right)^{\beta+1}$

for $n = n_a, n_a+1, \dots$ with n_a the lower cut-off, $\gamma > 1$ and $\zeta(\gamma, n_a)$ the Hurwitz zeta function (a generalization of the Riemann zeta function, $\zeta(\gamma) = \zeta(\gamma, 1)$), defined as

$$\zeta(\gamma, n_a) = \sum_{k=0}^{\infty} \frac{1}{(n_a + k)^{\gamma}},$$

and providing the normalization of the distribution. When $n_a = 1$ the distribution given by Eq. (9) is sometimes called the zeta distribution.

The corresponding cumulative distribution function (for $n_a \ge 1$) is obtained from $S(n) = \sum_{n'=n}^{\infty} f(n')$, yielding

$$S(n) = \frac{\zeta(\gamma, n)}{\zeta(\gamma, n_a)}, \tag{10}$$

which does not have a strict power-law shape, but an asymptotic power-law shape. The empirical rank-size relation corresponding to this distribution of sizes is given by

$$n(r) = \zeta_2^{-1}\left(\gamma, \zeta(\gamma, n_a)\frac{r}{V}\right), \tag{11}$$

for $r \leq V$, where ζ_2^{-1} is the inverse of the Hurwitz zeta function with respect to its second argument. These formulas are also included in Table II. Equation (2.3) of Ref. [55] states that

$$\zeta_2^{-1}\left(\beta+1, \frac{1}{\beta\nu^\beta}\right) = \nu + \Delta(\beta,\nu),$$

with the function $\Delta(\beta,\nu)$ fulfilling $0 < \Delta(\beta,\nu) < 1$. This means that we can write $n(r)$ as a power law in r plus an extra term

$$n(r) = \left(\frac{V}{\beta\zeta(\beta+1,n_a)r}\right)^{1/\beta} + \Delta(\beta,\nu),$$

with $\nu = [\beta\zeta(\beta+1,n_a)r/V]^{-1/\beta}$.

Note that although the empirical size usually takes values in the range $n \geq 1$, the theoretical $f(n)$ only considers $n \geq n_a$, with perhaps $n_a \gg 1$. The reason is that the theoretical power law only pretends to fit the tail of the empirical distribution, with the tail defined by the power-law range $n \geq n_a$. Thus, all empirical values of n below n_a are disregarded for the power-law fit (in line with Clauset et al.'s framework [12]).

As outlined in the introduction, the previous equation for $f(n)$, Eq. (9), can be understood as a second, different definition of Zipf's law (alternative to Eqs. (4)), which we call Zipf's law for sizes (as, by counting repetitions of the size variable, it leads to a power law). Both definitions of Zipf's law [Eqs. (4) and (9)] are not equivalent, only asymptotically equivalent in the limit of large n, i.e., small r. Such a distinction between the definitions would disappear completely if n were a continuous variable; thus, it is an effect of the discreteness of the tokens. It is important to clarify that, although the descriptions of systems candidate to fulfill Zipf's law in terms of the rank-size (or rank-frequency) plot and in terms of the distribution of sizes are fully equivalent (in the sense that one can recover any of the two with the knowledge of the other [26, 27]), a power-law relationship in one case does not imply a power law in the other, and reciprocally. This leads to the two distinct definitions (or "flavors") of Zipf's law just explained.

In summary, we have two representations of Zipfian systems, in terms of the rank-size relation or in terms of the distribution of sizes, and both approaches are equivalent to describe such systems. However, a pure power law in one of the representations does not imply a pure power law in the other, and *vice versa* (Table II), and therefore we have two alternative, different definitions of Zipf's law.

III. RECOVERY OF POWER-LAW RELATIONSHIPS FROM SIMULATIONS OF ZIPFIAN SYSTEMS

In this section, we deal with simulated systems built using any of the two different versions of Zipf's law discussed above, Eq. (1) and Eq. (2), respectively. The empirical values of the power-law exponents will be obtained by means of maximum likelihood estimation applied in any case to both the rank-size relation $n(r)$ and the distribution of sizes $f(n)$. Further, for illustrative and simplifying purposes, we will compare the results of applying the ML-estimation formulas for continuous power laws (which is an approximate method for the discrete systems we are interested in) with the results of ML estimation when the discreteness of the power-law distributions is taken into account. For continuous power laws, we apply the method developed in Refs. [13, 56] (see also Ref. [17]). The procedure for discrete power-law distributions is fully explained in Appendix A. An overview of both procedures follows here.

A. Maximum likelihood estimation and goodness-of-fit tests

Given a continuous, non-truncated power-law distribution, $g(x) = (\tau-1)a^{\tau-1}/x^{\tau}$, defined for $x \geq a > 0$ with $g(x)$ the probability density of x and a the lower cut-off, the ML estimation of the exponent τ is straightforwardly obtained as

$$\hat{\tau} = 1 + \frac{1}{\ln G_a - \ln a}, \tag{12}$$

where G_a is the geometric mean of the values of x in the sample fulfilling $x \geq a$ [12, 13].

For a discrete, non-truncated power-law distribution, $g(x) = 1/[\zeta(\tau, a)x^{\tau}]$, defined for $x = a, a+1, \dots$ (with $g(x)$ the probability mass function and $\zeta(\tau, a)$ the Hurwitz zeta function defined in the previous section), the ML estimation of τ comes from the value that maximizes the (per-datum) log-likelihood

$$\ell(\tau) = -\ln \zeta(\tau, a) - \tau \ln G_a. \tag{13}$$

In this case, a closed solution for $\hat{\tau}$ is not possible, due to the presence of the function $\zeta(\tau, a)$ in the expression, and one has to perform the maximization numerically [12, 57].

It is clear that ML estimation for power-law distributions is much simpler for continuous random variables, and for this reason it will be used here, but together with the more complicated discrete case (which is the natural procedure). The random variable x and the exponent τ will represent either the rank r and the exponent α appearing in the version of Zipf's law for types, Eqs. (1) or (4), or the size n and the exponent γ of Zipf's law for sizes, Eqs. (2) or (9).

It is interesting to observe that ML estimation explicitly assumes independence in the sample [13, 37]. However, this assumption is implicit in any case in which one calculates a histogram from a sample and considers it as representative of the distribution of the whole population. Thus, the usual power-law (or Zipf's law) paradigm has to be understood as trying to answer the following question: if the data were independent, which would be the univariate probability distribution describing it? In case of dependences in the data (as in time series of meteorological events), one can filter the data eliminating the dependent events [58].

For the sake of generality, the power-law fit is not performed on the full range $x \geq 1$ but on the upper tail of the distribution, whose starting point is given by the parameter a (which corresponds to r_a and n_a in each of the two representations and could take the value $a = 1$ as a particular case). This allows one to deal with empirical distributions that are not pure power laws but only asymptotic power laws. When the number of data is not infinite (always, in practice), there will exist a value of a such that an asymptotic power law will be confused with a pure power-law; the fact of fitting power-law tails takes advantage of this fact [16].

As, *a priori*, a is undetermined, one needs to do the fits for different values of a, and compare the goodness of each fit (we sweep 20 values of a per order of magnitude, equispaced in log-scale). We take the smallest value of a such that the fit is clearly non-rejectable ($p-$value greater than 0.2), using the Kolmogorov-Smirnov goodness-of-fit test, with the distribution of the Kolmogorov-Smirnov statistic calculated from 100 Monte-Carlo simulations [13, 57]. The simulations will allow us also to estimate the standard deviation of the estimated value of the exponent. This approach to power-law fitting and testing is inspired by Clauset et al.'s method [12], but correcting some of its important shortcomings [13, 14].

B. Simulation of Zipf's law with rank as the random variable

We start by generating a synthetic sample following the Zipf's law for types. This means that we take (individual) tokens from the types contained in a directory (e.g., word types in a special dictionary), which, in addition to the list of all possible types also contains some information on their global size or global frequency in the population. We assume that information comes from the discrete power-law relation

$$n(z) \propto \frac{1}{z^{\alpha}}, \tag{14}$$

between the global size n and the random variable z associated to each type, with $z = 1, 2, \ldots,$ up to infinite (non-truncated power law), and with $\alpha > 1$.

Notice that z, which is the variable which is power-law distributed, just represents an arbitrary labeling of the types and has no physical meaning, except for its monotonic relation with n (n gives the size or absolute frequency in the population and is directly proportional to the probability mass function of z). This is the same model considered by Mandelbrot and mentioned above [32], where the number of possible types (number of possible values of z in the population) is infinite.

If we draw a sample of L_{tot} independent random numbers z following the previous distribution, we obtain a sequence of tokens, which constitutes our system of size L_{tot}. By construction, this is a *Zipfian system* obeying, in principle, Zipf's law for types. Notice that the empirical value of n, for each type, will be a binomial random variable with expected value $n(z)$ given by Eq. (14). The algorithm to simulate the discrete power-law distribution is explained in Appendix B and is a generalization of the one presented in Ref. [59].

The outcome of the process for a particular realization with $\alpha = 1.2$ and $L_{tot} = 10^6$ is shown in Fig. 1. The observed sizes n of each type (or absolute frequencies, just counting tokens with the same z) are plotted versus z; in addition, a less naive estimation of the distribution of z is obtained by adapting the logarithmic-binning procedure explained in Refs. [60, 61] to discrete distributions (Appendix C) [84].

However, in practice, one has only access to the resulting sizes n of the different types and not to the random variable z, which we may consider then a hidden variable (i.e., a hidden rank with no physical representation that one can measure). The substitution of z by the rank r is a useful trick, performed ordering the resulting types (only those contained in the sequence) by decreasing size n, and assigning rank values from 1 to V_{tot} (remember that V_{tot} is the total number of resulting types, and equal to 133,146 in the realization of our example). Figure 1 shows how the rank is in good correspondence with the hidden variable z for small values of z (up to about 1000 for this concrete example), but not for intermediate and large values, totally missing the long tail of z.

This failure of correspondence is due to the unavoidable statistical fluctuations in finite samples, which make that although the probability of z_1 is higher than that of z_2 if $z_1 < z_2$, the value of z_1 does not appear necessarily more frequently than that of z_2 if z_1 is large enough; even more, z_1 may not appear at all in the sample. It is clear then that the rank is not the same that the true random variable z on which the distribution is defined. Note also that, by the definition of the concept of random-variable, one has to associate events to numbers for the whole sample space [62], but in the case of ranks the association is done *a posteriori*, after the random sample is drawn; thus, different samples lead to different assignations. Put differently, the rank is not univocally associated to elements of the sample space when two or more random samples are drawn.

Let us apply the maximum-likelihood-estimation method, together with the goodness-of-fit testing procedure (outlined in the previous subsection and fully explained in Appendix A), to the rank data obtained from the simulation. To be precise, the quantities defined in the previous subsection, x, $g(x)$, τ, and a correspond to r, $n(r)/L$, α, and r_a, respectively. The outcome of the procedure leads to the rejection of the power-law distribution for r, no matter the minimum value r_a used to truncate the distribution from below, and no matter also if a discrete or a continuous power law is fitted; that is, we do not find $p-$values larger than the 0.20 threshold for any value of r_a. Thus, no power-law distribution can be fitted to the rank-size relation by ML estimation, although the relation is constructed from a true power law. For the particular example used for illustration, the case $r_a = 1$ leads to $\hat{\alpha} = 1.2214 \pm 0.0002$ for the discrete fit and $\hat{\alpha} = 1.2516 \pm 0.0003$ for the continuous fit, close to the original value (with some positive bias) but rejected for the reasons described below (as a test of consistence, for the fit with the hidden rank z taking minimum value $z_a = 1$ we get $\hat{\alpha} = 1.1994 \pm 0.0002$).

Figure 2 shows the positive bias of the ML estimation $\hat{\alpha}$ as a function of the lower cut-off r_a and for different system sizes L_{tot}. Indeed, a true power law would generate much higher values of the variable (well beyond the 133,146 of maximum rank in our example, see

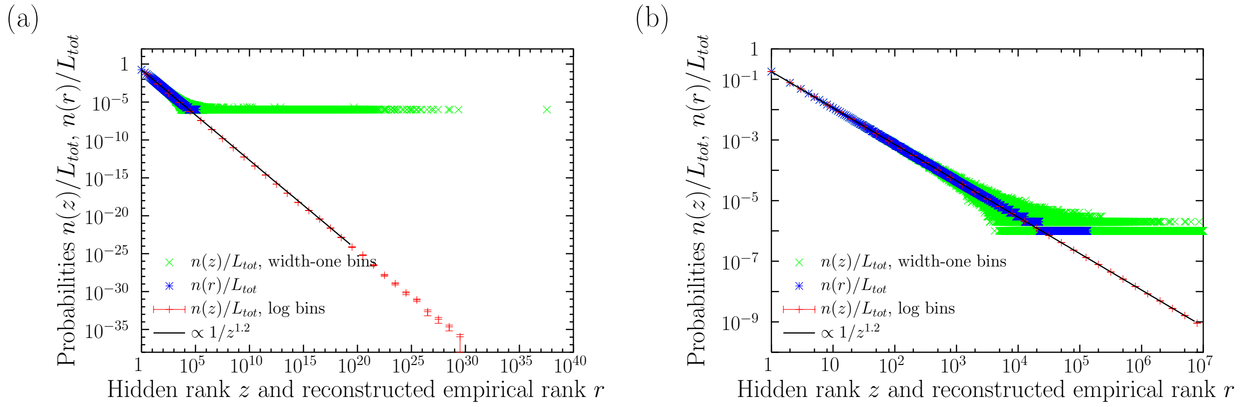

FIG. 1: Estimation of the probability mass function $n(z)/L_{tot}$ of the hidden variable z (hidden rank) associated to type size n for a synthetic sequence with discrete power-law distribution of z, Eq. (14), with exponent $\alpha = 1.2$ and length $L_{tot} = 10^6$, together with the corresponding empirical rank-size relation $n(r)/L_{tot}$. Both size-one bins and logarithmic bins are shown for $n(z)/L_{tot}$. The solid line is the original discrete power law from which the tokens are drawn. Notice how the only available quantity in practice, the rank-size relation, deviates from a power law for large ranks (corresponding to intermediate and large values of z). This deviation (apart from the bias in the ML-estimated exponent) is responsible for the rejection of the power-law hypothesis. (a) Full plot. (b) Restricted plot.

Fig. 1), which would lead to a larger geometric mean and to smaller exponents, through Eqs. (12) or (13). Thus, the fact that $r \ll z$ for large z leads to an underestimation of the geometric mean of the variable and to an overestimation of the exponent. The bias is specially pronounced for large r_a, corresponding to a decreasing number of data in the power law. For small r_a, specially when L_{tot} is large, the bias of the exponent practically disappears, as seen in Fig. 2.

Nevertheless, the rejection of the power law is achieved through the Kolmogorov-Smirnov test, due to the lack of resemblance of the rank data with true power-law distributed data. Visual inspection of the rightmost part of Figure 1 seems to indicate that a "flat" power law, i.e., one with an exponent equal to zero, could fit the largest ranks (those with $n = 1$); nevertheless, such distribution is not normalizable when defined over an infinite support and

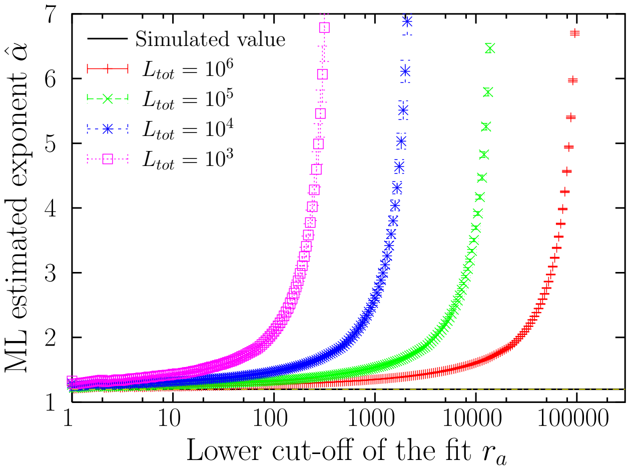

FIG. 2: ML exponent $\hat{\alpha}$ from the rank-size representation in the continuous approximation, as a function of the cut-off in rank, r_a, for synthetic systems fulfilling Zipf's law for types, Eq. (14), with original value of α equal to 1.2 and different values of L_{tot}. Observe the increasing positive bias from the true value with increasing r_a. In all cases, the Kolmogorov-Smirnov test rejects the power-law hypothesis with the ML-estimated exponent. Different values of the simulated $\alpha-$value lead to analogous results. The fitting of a discrete power law yields essentially the same results (not shown).

cannot arise therefore from the ML formalism (except if one introduces an upper truncation in the power law). On the other side, one could envisage a goodness-of-fit test in which empirical rank-size data are compared with simulated rank-size data. This is not contemplated in the standard algorithms provided in Refs. [12, 13] and it should be the subject of future research.

On the contrary, for the same data generated from Zipf's law for types, the application of ML estimation to the distribution of type sizes $f(n)$ (counting how many types have a given size) leads to the acceptance (that is, non-rejection) of the power-law hypothesis for precise values of the lower cut-off n_a. In this case, the correspondence with the quantities of Section III A is $x = n$, $g(x) = f(n)$, $\tau = \gamma$, and $a = n_a$. In the concrete example mentioned above with $\alpha = 1.2$, we get, for discrete ML estimation, an estimated power-law exponent $\hat{\gamma} = 1.86 \pm 0.01$ starting to hold for $n \geq 7$, with a $p-$value 0.85. In the simpler continuous approximation, the results are $\hat{\gamma} = 1.84 \pm 0.02$ for $n \geq 32$, with a $p-$value 0.29. The corresponding values of α are, using Eq. (6), $\hat{\alpha} = 1.16$ and $\hat{\alpha} = 1.19$, respectively; thus, the results are close to the true value but with some bias (the true exponent of the asymptotic power law is $\gamma = 1 + 1/1.2 = 1.833$, after Eq. (6)). Both fits are shown in Fig. 3(a) in terms of $f(n)$ and $S(n)$. The "translation" of the fits into the rank-size format appears in Fig. 3(b). Table III summarizes more results from simulations of this kind. A total of 20 systems are simulated from Zipf's law for types for each value of α, being these $\alpha = 1.2, 1.3$, and 1.4. Under no circumstance, there is a contradiction with the conclusions obtained from the example shown in Fig. 3.

TABLE III: Maximum likelihood fitting of $f(n)$ for synthetic systems fulfilling (by construction) Zipf's law for types, Eq. (14). For each value of α we generate 20 systems with $L_{tot} = 10^6$ tokens each, and both discrete ML estimation (upper row for each α) and continuous ML estimation (lower row) are performed. The averages (represented by the bar) of the estimated exponent $\hat{\gamma}$, of the cut-off n_a, and of the $p-$value, are over the 20 samples, and the error is one standard deviation of the variable (not of its mean, which is a factor $\sqrt{20}$ smaller). Also, the average and the estimation of the standard deviation of the resulting number of types V_{tot} are included. Note the higher (positive) bias of the estimated exponent in the discrete case, due to the smaller value of n_a.

α	γ	$\bar{V}_{tot}$	$\bar{\hat{\gamma}}$	$\bar{n}_a$	$\bar{p}$	ML estimation
1.20	1.833	132,934±258	1.861±0.010	7.1±3.0	0.51±0.28	discrete n
			1.842±0.007	32.5±3.2	0.47±0.19	continuous n
1.30	1.769	56,771±168	1.794±0.009	6.0±1.6	0.58±0.25	discrete n
			1.774±0.006	25.3±4.1	0.38±0.14	continuous n
1.40	1.714	27,098± 88	1.739±0.007	5.0±1.2	0.57±0.25	discrete n
			1.722±0.008	22.8±3.4	0.42±0.17	continuous n

These results lead to the remarkable situation that, although the underlying pure Zipf's law may be valid for types, we find the distribution of sizes $f(n)$ more reliable than the statistics of types $n(r)$ in order to test the power-law hypothesis when the best method of parameter estimation (ML, as explained in Appendix A) is used [13]. In fact, the approximate continuous procedure seems to yield better results than the exact discrete case. The reason is that the true empirical distribution of sizes is only a power-law asymptotically, and presents a strong excess of probability for the smallest values of n (in comparison to a pure power law, see Fig. 3(a)). As the continuous ML method works worse for discrete data, it rejects the power-law hypothesis for small values of n_a, avoiding the deviations from the power law and yielding an exponent closer to the original one. In contrast, the discrete ML method is not able to reject the hypothesis for smaller n_a's, at the price that there is a certain bias in the value of the exponent. In practice, both fits look very satisfactory, as shown in Fig. 3, although, due to its smaller bias one may prefer the continuous case. Figure 4 shows the biased result for the exponent $\hat{\gamma}$ for small values of the lower cut-off n_a. Note that it is not the ML method that leads to biased results, but that the distribution deviates from a power law when small values of n are taken into account, as seen in Fig. 3(a).

C. Simulation of Zipf's law with size as the random variable

Now we generate a synthetic Zipf's law not for types (as in the previous subsection) but for sizes, i.e., the discrete power-law distribution for $f(n)$, Eq. (9), holds exactly in the population. Thus (in order to compare with that subsection), we generate $V_{tot} = 133,000$ independent random numbers from a discrete power law, $f(n) = 1/[\zeta(\gamma)n^{\gamma}]$, defined for $n = 1, 2, \ldots$ (i.e., $n_a = 1$), with $\zeta(\gamma) = \zeta(\gamma, 1)$ the Riemann zeta function and exponent $\gamma = 1.833$ (corresponding to $\alpha = 1.2$), using the algorithm explained in Appendix B. Each of these random numbers represents the size n of a type, and thus, $f(n)$ is obtained directly from the statistics of n. To plot the rank-size relation, we order the list of types by decreasing value of n, and assign ranks $r = 1, 2, \ldots V_{tot}$.

So as to apply ML estimation to the rank variable, we need to generate a synthetic system of tokens (the key step is the calculation of the empirical value of the geometric mean of the random variable r, Eqs. (12) and (13)), this is done by creating $n(r)$ copies of each of the V_{tot} types, each labeled by its corresponding rank r. This list of L_{tot} values is the data entering as the input of the ML routine in the rank-size approach. This procedure leads to the same results as in the previous subsection, that is, the power-law hypothesis for the rank-size relation is rejected, no matter how large the value of the lower cut-off r_a is.

In contrast, ML for the random variable n does not allow one to reject a description in terms of a power-law distribution. This is obvious, as n comes indeed from a power-law distribution, such that, at variance with Zipf's law for types, the value of its random

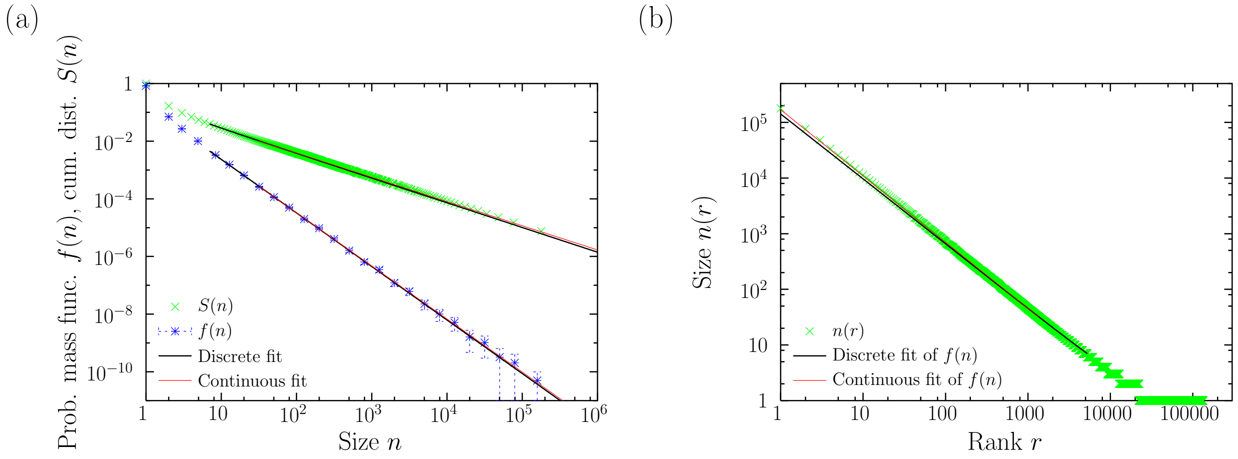

FIG. 3: (a) Estimated probability mass function $f(n)$ and survivor function $S(n)$ for simulated Zipf's law for types, Eq. (14), taking $\alpha = 1.2$ and $L_{tot} = 10^6$ (same data as in Fig. 1), together with the corresponding power-law ML fit for the random variable n, Eq. (9), in the discrete case and also in the continuous approximation (lines). The fit in terms of $S(n)$, given by Eq. (10), is also shown. We remark that ML estimation does not rely on the estimations of $f(n)$ of $S(n)$, these are shown here as a visual verification of the goodness of the fits. Expression (8) provides a good fit of $f(n)$ for all n (not shown). (b) Translation of the previous ML fit of the n variable into the rank-size representation, given by the inverted Hurwitz zeta function of Eq. (11). The corresponding values of the upper cut-off r_b turn out to be $r_b \simeq 5250$ (discrete fit) and $r_b \simeq 1350$ (continuous fit).

variable is not hidden. The results for the selected example (equivalent to Fig. 1) are $\hat{\gamma} = 1.835 \pm 0.004$ for $n \geq 2$, with a $p-$value 0.31 for ML estimation of a discrete power law and $\hat{\gamma} = 1.838 \pm 0.015$ for $n \geq 56$, with a $p-$value 0.23 for the continuous-version approximation. Figure 5(a) shows the direct outcome of the fit, both in terms of $f(n)$ and $S(n)$, whereas Fig. 5(b) shows how the fit of the distribution of n translates into the rank-size representation.

Note that the resulting system size, $L_{tot} = \sum_{r=1}^{V_{tot}} n(r)$, turns out to be, in the particular realization chosen as an example, $L_{tot} = 3,417,385$. The large difference with the value $L_{tot} = 10^6$ used in the previous subsection for about the same V_{tot} is due to the distinct

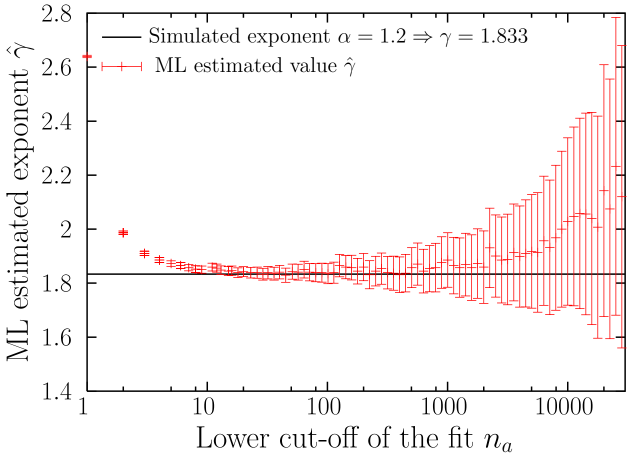

FIG. 4: ML exponent from the distribution-of-sizes representation, $\hat{\gamma}$, as a function of the cut-off in size, n_a, for a synthetic system fulfilling Zipf's law for types, Eq. (14), using the (exact) discrete fitting procedure. The original value of α is equal to 1.2 (represented as a straight line for $\gamma = 1.833$) and $L_{tot} = 10^6$ (same data as in Fig. 3). The continuous fitting leads to similar results.

balance between types with $n = 1$ and the rest. In the simulation based on Zipf's law for types, there was an excess of very small $n-$values; thus, for the same V_{tot} the size of the system L_{tot} becomes smaller there. To ease comparison, Figure 6 shows together the $f(n)$ resulting from simulating Zipf's law for types (previous subsection) with Zipf's law for sizes (this subsection).

Another source of variation in the value of the resulting system size L_{tot} is that this arises as a sum of independent power-law distributed sizes n. As the exponent of the power law γ is smaller than 2, the law of large numbers does not apply and the sum is not scaling linearly with the number of terms (types) V_{tot}. Instead, the sum is broadly distributed, as expected from the generalized central limit theorem [63–65]. Table IV provides the results obtained from other examples simulating Zipf's law for sizes [Eq. (9)]; these results are in total agreement with the example chosen for illustration and show the wild dispersion in the resulting values of L_{tot}. In this case, it is clear that the discrete fit is preferred, as it leads to a much smaller value of n_a, widening the power-law regime and reducing the uncertainty in the exponent.

TABLE IV: Equivalent to Table III, for systems fulfilling Zipf's law for sizes. For each value of γ we generate 20 systems with $133,000$ types each. In this case it is clear that discrete ML estimation provides better results than the continuous approximation. The bias of the estimated exponent becomes negligible. Notice the enormous variation on L_{tot}, described by its minimum and maximum values.

γ	α	min L_{tot}	max L_{tot}	$\bar{\hat{\gamma}}$	$\bar{n}_a$	$\bar{p}$	ML estimation
1.833	1.20	$2.9 \cdot 10^6$	$2.4 \cdot 10^8$	1.833±0.003	1.2± 0.5	0.61±0.24	discrete n
				1.834±0.018	56.2±23.9	0.32±0.13	continuous n
1.769	1.30	$7.2 \cdot 10^6$	$1.2 \cdot 10^9$	1.769±0.003	1.4± 0.9	0.61±0.23	discrete n
				1.773±0.012	55.4±17.4	0.38±0.14	continuous n
1.714	1.40	$1.6 \cdot 10^7$	$5.9 \cdot 10^9$	1.713±0.002	1.3± 0.4	0.66±0.26	discrete n
				1.716±0.012	63.0±23.8	0.35±0.17	continuous n

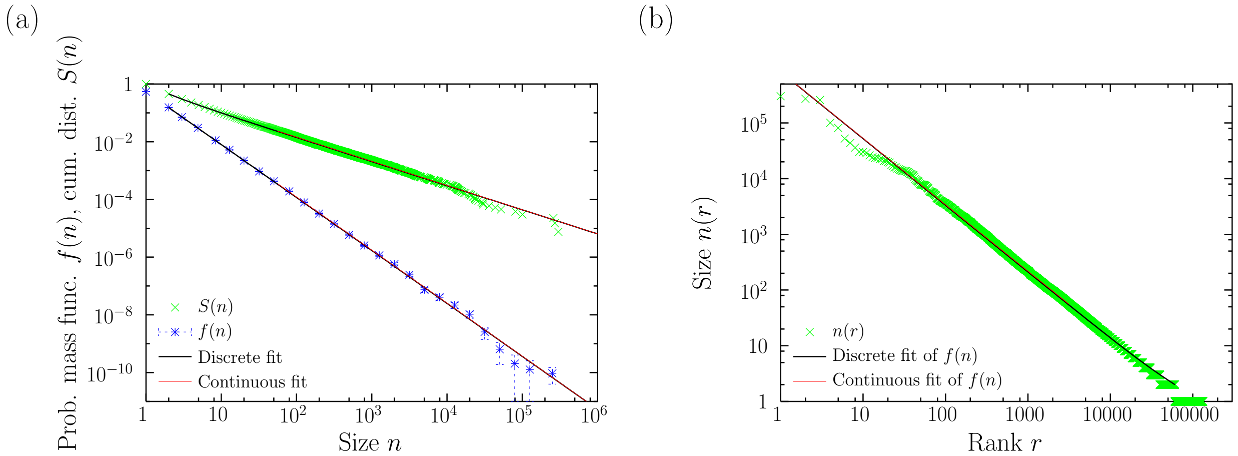

FIG. 5: Same as Fig. 3 replacing the simulation of Zipf's law for types by the simulation of Zipf's law for sizes, Eq. (9), with $\gamma = 1 + 1/1.2 \simeq 1.83$, $n_a = 1$, and $V_{tot} = 133,000$. (a) Empirical cumulative distribution and probability mass function of sizes, together with discrete and continuous fits. (b) Empirical rank-size representation and fits.

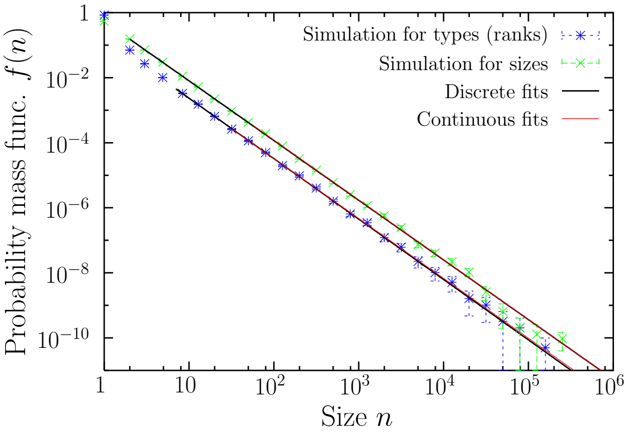

FIG. 6: Direct comparison of the estimated probability mass functions $f(n)$ arising from the simulation of both versions of Zipf's law, Eq. (14) [Fig. 3(a)] and Eq. (9) with $n_a = 1$ [Fig. 5(b)]. Notice the bending of $f(n)$ for small n when Zipf's law is simulated for types.

IV. DISCUSSION

We have shown, empirically, the clear advantages of testing Zipf's law by using maximum likelihood estimation applied to the distribution of sizes, instead of to the rank-size relation. We have dealt with two different generative processes. The first one consists of generating tokens from Zipf's law for types (a power law for $n(z)$, with z a hidden rank, and n the size, or frequency). The second one consists of generating the size of types from Zipf's law for sizes (a power law for $f(n)$). In each of them, we apply ML estimation to both possible representations of the system (in terms of $n(r)$ and in terms $f(n)$, with r the observed, empirical rank). Applying ML to $n(r)$ always leads to bad results ($p = 0$, due to the fact that $n(r)$ does not resemble a power law for large r; in addition, the power-law exponent may show considerable bias, see Fig. 2). Applying ML to $f(n)$ leads to good results (with some small bias) when the power law comes from $n(z)$ and to perfect results (no bias) when the power law comes from $f(n)$. That is the reason why we recommend applying ML to the representation in terms of $f(n)$, no matter the generation process. Table V summarizes these results.

In our experimental setup, the generation process is known (it is determined by us) but in a practical situation the generation process is unknown. In such a case, for the same reasons, we recommend applying ML to the representation in terms of $f(n)$. Nevertheless, knowing the generation process could allow one to fit models that are more refined than a power law, for instance Eq. (8) when tokens are drawn from a pure power law.

More general arguments in favor of the distribution of sizes are the following:

- *Parsimony.*

  According to standard model selection, the evaluation of the goodness of a model (e.g., a power law) must be based not only on the quality of the fit of the model but also on a penalty for the number of parameters employed [66] (although the very meaning of what a single parameter can be has been recently questioned [67], those criticisms are not relevant for our study). The fact that $\alpha = 1/(\gamma - 1)$ [recall Eq. (6)] means that if $\gamma \geq 2$ then $\alpha \leq 1$. These values of the exponent have been claimed for human language under certain conditions [68].

  The problem is that a power-law distribution needs $\alpha > 1$ for normalization (otherwise, the summation of the probabilities does not converge). Therefore, when sizes are taken for statistical modeling, just two parameters suffice for a power-law description: γ and a minimum cut-off n_a; in contrast, the corresponding power-law for the rank-size representation needs three parameters: α, a minimum rank r_a, and a maximum rank r_b. In fact, this maximum rank is unavoidable in this representation, even for $\alpha > 1$, due to the distortions introduced in the rank-size relation by the artefactual nature of the rank (as we have seen, because of the plateaus corresponding to the smallest sizes, which lead to the rejection of the non-truncated power-law hypothesis for ranks). Note that it has been claimed [15] that Clauset et al.'s method [12] is not able to fit $\alpha-$exponents close to one.

  We have not considered truncated power laws in this article [41] because they are not "genuine" power laws (as mentioned in the introduction) and because they are not possible to fit using common fitting software [12]. In any event, although we consider a non-truncated power law a simpler model than a truncated power law, even using model selection criteria such as BIC, AIC, etc. [66] it is not clear to us at this point how to compare non-truncated power laws in the distribution-of-sizes representation with truncated power laws in the rank-size representation, due to the fact that the number of data that are compared are different (in fact, the data sets can be considered different, one with V elements and the other with L). This is left for future research.

Note also that as Zipf's law is expected to hold for the most common words (less common words have a different power-law behavior [33, 69]) its practical description in terms of non-truncated power laws is more naturally addressed in the representation in terms of the distribution of sizes (nevertheless, in contrast to real systems, in our simulations Zipf's law holds for the whole range of the random variable, so this effect is not present).

- *More flexibility for the largest sizes.*

  Taking the rank-size representation as the true form of Zipf's law, $n(r) \propto 1/r^{\alpha}$, leads to a very narrow range of variation of the largest sizes. For example, for the population of USA cities, as New York, with rank $= 1$ has about 8,600,000 inhabitants, this form of Zipf's law enforces that the second largest city (L.A.) has to have a number of inhabitants very close to 4,300,000 (assuming $\alpha = 1$), and the third largest city has to be close to 2,900,000, with little margin of variation. This is a very strict form of Zipf's law, almost impossible to fulfill in practice (except when one has very little data, e.g., for a very-low-populated country).

  In contrast, a power-law relation for the distribution of sizes, $f(n) \propto 1/n^{\gamma}$, allows a large range of variability between the largest types, as there is no strong link between the highest values that the random variable can attain in practice (the largest n divided by the second largest n is not enforced to yield a value very close to two). This means that under this form, for example, the tail of the size of U.S. cities is well described by a power-law distribution [29]. This is illustrated in the large sizes (low ranks) displayed in Figs. 5(a) and 5(b). Whereas the fitting based on size as a random variable does not reject a power-law distribution in that range (despite the erratic proportions between the values of n for $r = 1, 2, 3$, etc.), a fitting based on rank leads to a clear rejection of the power law in any range comprising those values of r. A consequence of this is that the fitting of the size variable n leads to a large tolerance of dragon-kings [70], see Ref. [34].

- *Bias for a negative correlation between rank and size.*

  The definition of rank forces the size (i.e., the number of tokens) of a rank to not increase as rank increases. This implies a negative correlation between a rank and its size. In contrast, the number of types with a certain size is free with regard to size. In principle, it can increase, decrease or remain constant as size increases. This difference is vital for model testing as the null hypothesis of a power law might be harder to reject in terms of the rank-size relation because of this correlation. The situation is analogous to the problems arising when fitting a probability distribution from its cumulative distribution function [71].

In our case, although we get $p-$values equal to zero when a non-truncated power law is fitted to the rank-size relation, the opposite effect, leading to inflated $p-$values, may arise if we truncate the rank-size power law at a maximum value r_b. Indeed, we have some preliminary results indicating that between $r = 100$ and 1000 the $p-$values of a rank-size relation generated from a discrete power law with exponent $\alpha = 1.2$ are not uniformly distributed between 0 and 1 but biased to the high values.

TABLE V: Summary of results found in this article. PL stands for the null hypothesis of a power-law distribution.

Simulating	ML applied to rank r	ML applied to size n
Zipf's law for types, Eq. (4)	PL rejected	PL not rejected (slight positive bias of $\hat{\gamma}$ for the discrete fit)
Zipf's law for sizes, Eq. (9)	PL rejected	PL not rejected (too large value of n_a in the continuous fit)

On the opposite side, an argument against the distribution of sizes and in favor of the rank-size relation is that the distribution-of-sizes approach entails a substantial reduction of the sample size when the texts are large (the size of the sample is the number of tokens in the rank representation and the number of types in the size representation). Indeed, Heaps' law [26, 50] approximately relates vocabulary (empirical number of types) with text length (empirical number of tokens) as $V_{tot} \propto L_{tot}^{1/\alpha}$, if $\alpha > 1$, which implies $V_{tot} \ll L_{tot}$ if L_{tot} is large (we have clearly seen this in the example chosen for several figures, with $L_{tot} = 10^6$ and $V_{tot} \simeq 133,000$). This can make the power-law hypothesis more difficult to reject in the distribution-of-sizes representation, because $p-$values are known to be data-size dependent, when the null-hypothesis does not hold [58, 72]. Nevertheless, the advantages shown in this article for the distribution-of-sizes representation regarding bias and goodness-of-fit testing (in particular that the $p-$value turns out to be zero when n is a power law by construction) overcome this effective data-size disadvantage.

Another issue related with the number of data is the possible existence of correlations or dependence between the data points. In our case, we have generated them independently, but in a real situation one does not have control over that. In that case, our approach implies that we are trying if a description of the system (sizes or ranks) is possible in terms of power-law distributed independent events (independence is the maximum-entropy outcome when no information about dependence is available [73]). However, in a real scenario it may be possible that one cannot disregard dependences, for instance if the data comes ordered in time (in a time series or in a point process, as it happens with natural disasters for instance). In such a case, a useful procedure is to eliminate part of data, i.e., the dependent data, and perform the fitting with the remainder of the data [58].

In summary, we have presented wide evidence that the description of Zipf's power law is a different matter in terms of the rank-size relation and in terms of the distribution of sizes. Put differently, both descriptions lead to different distributions of tokens into types. Whatever version of Zipf's law might hold in real systems, or even, if neither of the two versions is expected to hold, the application of maximum likelihood estimation should be improved, in general, by taking size as the random variable. Therefore, we recommend investigating Zipf's law from the size-distribution's viewpoint. Although we have worked in the context of Zipf's law, our results are general for discrete power-law distributions. Our analysis can be applied to other distributions with more parameters than simple power laws.

Acknowledgments

We thank P. Puig for discussions about statistical testing, G. Altmann and G. Boleda for his advice from a quantitative linguistics perspective, E. Galbraith and I. A. Hatton for comments, L. Devroye for providing a copy of his valuable book for free, and A. Deluca and F. Font-Clos for feedback on the maximization of $g(m)/q(m)$, which should complement Devroye's book. Support from projects FIS2012-31324, FIS2015-71851-P, TIN2017-89244-R, PGC-FIS2018-099629-B-I00, María de Maeztu Program MDM-2014-0445, from Spanish

MINECO; 2014SGR-1307, 2017SGR-856 (MACDA) from AGAUR; and the Collaborative Mathematics Project from La Caixa Foundation (I.S.) is acknowledged.

APPENDIX A: FITTING AND TESTING DISCRETE POWER LAWS

Here we explain the procedure of finding accurate values of the parameters of the discrete power-law distribution. The method presented is an extension to the discrete case of the method introduced and developed for continuous power laws in Refs. [13, 56], that was in turn inspired by that of Clauset *et al.* [12], but introducing important modifications that yielded a better performance [14, 16]. As the continuous case has been treated in the previous literature, we explain here the peculiarities of the discrete fitting. In the exposition, we use a generic representation that is valid both for the rank-size representation and for the distribution of sizes. Table VI provides an equivalence between the notation used in each representation.

TABLE VI: Correspondence of notation between the generic representation used in this Appendix (and in the Maximum likelihood estimation and goodness-of-fit tests subsection), the rank-size representation, and the representation in terms of the distribution of sizes. Being strict, we should have defined a function $h(r)$ as $h(r) = n(r)/L$ but we have preferred to avoid a growth in the notation.

	Variable	Mass func.	Cumul. distrib.	Exponent	Lower cut-off	Number of data
Generic representation	x	$g(x)$	$G(x)$	τ	a	N_a
Rank-size representation	r	$n(r)/L$	–	α	r_a	L
Distribution of sizes	n	$f(n)$	$S(n)$	γ	n_a	V

1. The discrete power-law distribution

Let us consider a discrete power-law distribution, defined for $x \geq a$, with a a natural number ($a \geq 1$). The corresponding probability mass function is

$$g(x) = \frac{1}{\zeta(\tau, a) x^{\tau}}, \quad \text{for } x = a, a+1, a+2, \dots \tag{A1}$$

(and zero otherwise), where normalization is ensured by the Hurwitz zeta function, defined as

$$\zeta(\tau, a) = \sum_{k=0}^{\infty} \frac{1}{(a+k)^{\tau}};$$

for $a = 1$, this function becomes the standard Riemann zeta function, and the distribution is called the zeta distribution (or rather confusingly, the discrete Pareto distribution [54]). The corresponding (complementary) cumulative distribution function is

$$G(x) = \frac{\zeta(\tau, x)}{\zeta(\tau, a)},$$

for $x \geq a$, giving, by definition, Prob[variable $\geq x$]. Our approach fits the value of τ corresponding to different values of a and selects the one that yields the largest power-law range provided that the quality of the fit is acceptable (i.e., the power-law model in Eq. (A1) is not rejected), as explained next.

2. Advantages of ML estimation

The superiority of ML estimation in front of other methods of fitting has already been pointed out by several authors. In particular, for many years the most common used method, at least for fitting Zipf's law (and in complex-systems research, in general), has been least-squares fitting. Important problems arise in this case when the empirical probability density or the empirical mass function (for which the minimization with respect to the fitting curve is performed) are obtained using naive linear binning [10–12, 74, 75]. Logarithmic binning of this function corrects some of these flaws (when empty bins are not present), as it does also the least-squares fitting of the cumulative distribution (showing however other problems [71, 76]), but still the least-squares method shows a considerable bias, high variance, and bin-size dependence, and yields distributions that are not normalized (as it does not use the fact that the curves to be fitted are probability distributions).

In contrast, the ML estimator (for distributions in the exponential family, where the power law belongs) is the one with minimum variance among all asymptotically unbiased estimators, a property that is called asymptotic efficiency [10, 11, 37]. Also, the ML estimator is invariant under reparametrizations [13, 77], which means that ML fits the distribution, not the parameters. An additional advantage is that the ML result is also invariant under change of variables (under very general conditions). For all these reasons, ML estimation is employed in this article for the study of Zipf's law.

3. ML estimation and computation of the Hurwitz zeta function

The first step (step 1) then is the fitting of τ by maximum likelihood estimation. Considering a as a fixed parameter, the (per-datum) log-likelihood function ℓ for a discrete power-law distribution is defined as the logarithm of the likelihood function, divided by the total number of data N_a in the power-law range (i.e., those such that $x \geq a$); this is,

$$\ell(\tau) = \frac{1}{N_a} \ln \prod_{i=1}^{N_a} g(x_i) = \frac{1}{N_a} \sum_{i=1}^{N_a} \ln g(x_i),$$

with x_i the recorded values of x, numbered from $i = 1$ to N_a. Values below a must be disregarded. This yields

$$\ell(\tau) = -\ln \zeta(\tau, a) - \tau \ln G_a,$$

where G_a is the geometric mean of the data in the range, that is,

$$\ln G_a = \frac{1}{N_a} \sum_{i=1}^{N_a} \ln x_i$$

for $x_i \geq a$.

As a and the data are constants, ℓ is only a function of τ, and the maximum of this function yields the estimation of τ, which we call τ_{emp}, that is,

$$\tau_{emp} = \arg\max_{\forall \tau} \ell(\tau),$$

where arg max denotes the argument that makes the function maximum. This maximization is performed in our algorithm through the downhill simplex method as implemented in Ref. [78], restricted here to one dimension. The computation of the Hurwitz zeta function uses an algorithm based upon the Euler-Maclaurin series [79],

$$\sum_{k=0}^{\infty} \tilde{g}(k) \simeq \sum_{k=0}^{M-1} \tilde{g}(k) + \int_M^{\infty} \tilde{g}(k) dk + \frac{\tilde{g}(M)}{2} - \sum_{k=1}^{P} \frac{B_{2k}}{(2k)!} \tilde{g}^{(2k-1)}(M),$$

where B_{2k} are the Bernoulli numbers ($B_2 = 1/6, B_4 = -1/30, B_6 = 1/42, B_8 = -1/30, \dots$) [52]. The desired approximation is obtained by applying the formula to $\tilde{g}(k) = (a+k)^{-\tau}$, whose derivatives of order $2k-1$ are

$$\tilde{g}^{(2k-1)}(M) = \tilde{g}^{(2k-3)}(M) \frac{(\tau + 2k - 3)(\tau + 2k - 2)}{(a+M)^2},$$

with $\tilde{g}^{(1)}(M) = \tilde{g}'(M) = -\tau/(a+M)^{\tau+1}$ and the integral yields $(a+M)^{1-\tau}/(\tau-1)$. Thus,

$$\zeta(\tau, a) \simeq \sum_{k=0}^{M-1} \frac{1}{(a+k)^\tau} + \frac{1}{(\tau-1)(a+M)^{\tau-1}} + \frac{1}{2(a+M)^\tau} + \sum_{k=1}^{P} B_{2k} C_{2k-1},$$

with

$$C_{2k-1} = \frac{(\tau+2k-2)(\tau+2k-3)}{2k(2k-1)(a+M)^2} C_{2k-3} \text{ and } C_1 = \frac{\tau}{2(a+M)^{\tau+1}}.$$

The sum from $k=1$ to P is stopped when a minimum value term is reached [79], otherwise $P = 18$. We also take $M = 14$.

As a check, the reader can verify that this method allows to calculate $\zeta(2,1) = \pi^2/6$ with more than 16 correct significant figures. This constitutes an improvement with respect to other approximations for the Riemann and Hurwitz zeta functions [80].

4. Kolmogorov-Smirnov goodness-of-fit test

As, given a, the fit only depends on the geometric mean of the data, maximum likelihood can yield very bad fits if the data are not power-law distributed (because the estimation assumes *a priori* that the power-law hypothesis holds). The second step (step 2) of the procedure is to measure the deviation between the data and the fit. For that purpose, we use the Kolmogorov-Smirnov statistic d_{emp}, defined as the maximum absolute difference between the (complementary) cumulative distributions corresponding to the empirical data and to the fit (parameterized by $\tau = \tau_{emp}$) [78], i.e.,

$$d_{emp} = \max_{\forall x \geq a} \left| \frac{N_x}{N_a} - G(x; \tau_{emp}, a) \right|,$$

where the maximization is performed for all values of $x \geq a$, integer and not integer, and N_x counts the number of data with values equal to x or larger (defined only for $x \geq a$). Therefore, large and small values of d_{emp} denote, respectively, bad and good fits. We recall that although $g(x)$ is a pure power law above a, $G(x)$ is only a power law asymptotically.

5. Simulation procedure

The next step (step 3) consists in the evaluation of which is bad and which is good; this is done with simulated data following the distribution obtained by maximum likelihood estimation, that is, a discrete power law defined for $x \geq a$ with exponent τ_{emp}, Eq. (A1).

Once N_a simulated values of x have been generated with the procedure described in Appendix B, they are treated in exactly the same way as the empirical data, following steps 1 and 2 above [12] (see also the Supporting Information of Ref. [81]): first (step 4), maximum likelihood estimation leads to a value of the exponent, this time denoted as τ_{sim} (notice that this value will be close but distinct to τ_{emp}, due to statistical fluctuations); second (step 5), the Kolmogorov-Smirnov statistic, now called d_{sim}, leads to a quantification of the distance between the simulated data (with τ_{emp}) and their fit $G(x; \tau_{sim}, a)$ (notice that parameterized by τ_{sim}).

However, comparison of this single value d_{sim} with the empirical one, d_{emp}, does not allow to draw any conclusion. Naturally, one needs an ensemble of values of d_{sim}, which are obtained by repeating the simulation procedure (steps 3, 4, and 5) many times. The position of d_{emp} in relation to the distribution defined by the obtained values of d_{sim} allows us to calculate the $p-$value of the fit. This is just defined as the probability that true power-law data (as the simulated data is), with exponent τ_{emp}, yield a Kolmogorov-Smirnov statistic equal or larger than d_{emp} (that is, the probability that for real power-law data the fit is worse than for the empirical data). The estimation of the $p-$value from the simulations is just

$$p = \frac{\text{number of simulations with } d_{sim} \geq d_{emp}}{\text{number of simulations}}.$$

The uncertainty in p can be obtained from the fact that the number of simulations with $d_{sim} \geq d_{emp}$ will be binomially distributed; therefore, its standard deviation can be estimated as

$$\sigma_p = \sqrt{\frac{p(1-p)}{\text{number of simulations}}}.$$

For 1000 simulations and p around 0.5 this is 0.016, but if p or $1-p$ are about 0.01 we find 0.003. The simulation procedure also allows one to obtain the uncertainty of τ_{emp} as the standard deviation of the values of τ_{sim}.

6. Selection of the value of the lower cut-off

Thus, for a fixed value of a we end with a value of p that tells us the goodness of the fit. Usually, values of p below 0.05 are considered bad, and therefore the hypothesis under testing (the goodness of the fit) is rejected, although this value is rather arbitrary. Repeating the whole procedure for different values of a, we will obtain (or not) a set of acceptable $a-$values (i.e., a set of $a-$values such that Eq. (A1) is not rejected), together with their corresponding exponents τ_{emp} (which will depend on the value of a). In order to select one of this set of values, we just choose the smallest $a-$value (which yields the largest range) provided that $p > 0.20$. This concludes the fitting and testing procedure, leading to final values a^* and τ^*_{emp} (denoted in the main text simply as a and τ, r_a and α, or n_a and γ). Formally,

$$a^* = \min\{a \text{ such that } p > 0.20\},$$

which has associated the resulting exponent τ^*_{emp}. It is worth mentioning that the final $p-$value of the fit for varying a is not the one corresponding to fixed $a = a^*$; nevertheless, for our purposes its computation is not necessary.

APPENDIX B: GENERATION OF DEVIATES FOLLOWING THE DISCRETE POWER-LAW DISTRIBUTION

1. The general procedure

We present a method to generate random deviates following Eq. (A1; the method generalizes to the case $a > 1$ the rejection method explained by Devroye [59] for the case $a = 1$. Although more efficient procedures have been proposed [82], we were not aware of them at the time of writing and running our code.

The generalization of the method of Ref. [59] proceeds as follows: First, a uniform random number u is generated between 0 and u_{max}, where u_{max} fulfills $a = 1/u_{max}^{1/(\tau_{emp}-1)}$. Then, a new random number m is obtained as the integer part of $y = 1/u^{1/(\tau_{emp}-1)}$, i.e.,

$$m = \text{int}(1/u^{1/(\tau_{emp}-1)}),$$

which will verify $m \geq a$ if u_{max} fulfills $a = 1/u_{max}^{1/(\tau_{emp}-1)}$. Notice that y is continuous and has probability density $(\tau_{emp} - 1)a^{\tau_{emp}-1}/y^{\tau_{emp}}$, which is the continuous approximation of our $g(x)$, and its cumulative distribution is $(a/y)^{\tau_{emp}-1}$. The same cumulative distribution holds for m (but only for its integer values). From here, the probability mass function of m turns out to be $q(m) = (a/m)^{\tau_{emp}-1} - (a/(m+1))^{\tau_{emp}-1}$ (which, for large m is $(\tau_{emp} - 1)a^{\tau_{emp}-1}/m^{\tau_{emp}}$).

Then, the rejection method gives a random number $x = m$ distributed following $g(x)$ if

m is kept when a new uniform random number v (between 0 and 1) fulfills

$$v \leq \frac{g(m)}{cq(m)}$$

(and rejected otherwise), with c the rejection constant

$$c = \max_{\forall m \geq a} \frac{g(m)}{q(m)} = \frac{g(a)}{q(a)} = \frac{1}{\zeta(\tau_{emp}, a) a^{\tau_{emp}}} \left(\frac{(a+1)^{\tau_{emp}-1}}{(a+1)^{\tau_{emp}-1} - a^{\tau_{emp}-1}}. \right)$$

The resulting random variable $x = m$ corresponds to the hidden rank z in Section III B and to the size n in Section III C. The maximum value of $g(m)/q(m)$ takes place at $m = a$ because this is a decreasing function; this is shown in the next subsection. Defining the auxiliary variable $T = (1 + m^{-1})^{\tau_{emp}-1}$ and the constant $b = (a+1)^{\tau_{emp}-1}$ the acceptation condition is equivalent to

$$vm \frac{T-1}{b - a^{\tau_{emp}-1}} \leq \frac{aT}{b},$$

which shows clearly that the simulation procedure does not require the computation of the Hurwitz zeta function.

2. Calculation of the rejection constant

The efficiency of the simulations of discrete power-law distributed numbers depends on finding the optimum rejection constant, which is given by the maximum of $g(m)/q(m)$, where the functions are defined in the previous subsection. We will show that the maximum is reached for the smallest value of m, as $g(m)/q(m)$ is a monotonically decreasing function. Let us calculate (removing irrelevant multiplicative constants),

$$\frac{q(m)}{g(m)} \propto m - \frac{m}{(1+m^{-1})^{\tau_{emp}-1}},$$

whose derivative is

$$\left(\frac{q(m)}{g(m)} \right)' \propto 1 - \left(1 + \frac{1}{m} \right)^{-(\tau_{emp}-1)} \left(1 + \frac{\tau_{emp} - 1}{m+1} \right).$$

Then, it is enough to show that $1 + (\tau_{emp} - 1)/(m+1)$ is smaller than $(1+m^{-1})^{\tau_{emp}-1}$, as this implies that the previous derivative is positive and $q(m)/g(m)$ is monotonically increasing.

For $\tau_{emp} > 2$, we can write $1 + (\tau_{emp} - 1)/(m+1) < 1 + (\tau_{emp} - 1)m^{-1}$, which in turn is smaller than $(1 + m^{-1})^{\tau_{emp}-1}$, as the Bernoulli's inequality states for $\tau_{emp} > 2$ and $m^{-1} > 0$.

Indeed, a version of Bernoulli's inequality states that $1+sz < (1+z)^s$ with s and z any real numbers fulfilling $s > 1$ and $z > 0$.

For $1 < \tau_{emp} \leq 2$, we write

$$1+\frac{\tau_{emp}-1}{m+1} = \frac{1+(\tau_{emp})m^{-1}}{1+m^{-1}},$$

which is again smaller than $(1+m^{-1})^{\tau_{emp}-1}$, using the Bernoulli's inequality for $\tau_{emp} > 1$ (this demonstration also holds for $\tau_{emp} > 2$, but the previous one is simpler).

APPENDIX C: LOGARITHMIC BINNING IN THE DISCRETE CASE

In the plots, the fits are compared against the empirical probability mass functions. These are estimated adapting logarithmic binning [60, 71], which uses a constant number of bins per order of magnitude (5 in our case), to discrete distributions [60]. Let us consider the intervals $[x_{(k)}, x_{(k+1)})$, labeled by $k = 0, 1\dots$ with $x_{(k+1)} = Bx_{(k)}$ and $B = \sqrt[5]{10}$ (in our case); the starting value $x_{(0)}$ is irrelevant, but the values of $x_{(k)}$ should not be integer, for numerical convenience. Then, each occurrence of x in the data set is associated to a value of k using the formula

$$k = \log_B(x/x_{(0)}).$$

Next, the number of occurrences of x in the interval k (i.e., the number of types with size in the interval range, see the denominator of the formula below) is divided by the total number of occurrences of any value of x (which is $N(a) = N_a$, changing notation for convenience) and by $\text{int}(x_{(k+1)}) - \text{int}(x_{(k)})$ (which counts the number of possible values of x in the interval, i.e., the number of integers). This yields the estimated value of the probability mass function $g_{emp}(x_k^*)$ in the $k-$th interval,

$$g_{emp}(x_k^*) = \frac{N(\text{int}(x_{(k)})+1) - N(\text{int}(x_{(k+1)})+1)}{N(a)[\text{int}(x_{(k+1)}) - \text{int}(x_{(k)})]},$$

where the value of the probability mass function is associated to a point x_k^* in the interval given by the geometric mean of the smallest and largest integer in the interval,

$$x_k^* = \sqrt{\text{int}(x_{(k)}+1)\text{int}(x_{(k+1)})},$$

see Ref. [71]. Compare the last two formulas with Eq. 1.12 in Ref. [26]. Notice that our procedure estimates directly the probability mass function for small values of x (as the number of integers in each bin is one, or zero), but tends to its continuous version (the probability density) for large x (as the number of integers approaches the width of the bin). Estimation of probability distributions for discrete but non-integer variables was described in Ref. [61].

The error bars associated to $g_{emp}(x)$ can be estimated from the fact that the number of counts in a given bin can be considered binomially distributed (assuming that the data are not correlated, but this assumption is also made in order to apply maximum likelihood estimation; in practice it is enough that the number of data is much larger than the range of correlations). For a binomial variable, the ratio between the standard deviation and the mean (the relative error) is given, approximately, by the inverse of the square root of the mean of the variable (i.e., mean number of counts in the bin; if there is no bin that accumulates most of the counts). The same relation holds for $g_{emp}(x)$, because it is proportional to the number of counts and the proportionality constants vanish when the ratio standard deviation / mean is taken. Approximating the mean number of counts to the actual number of counts, then, the standard deviation of $g_{emp}(x)$ in bin k is obtained as

$$\sigma_k \simeq \frac{g_{emp}(x_k^*)}{\sqrt{\text{counts in } k}} = \frac{g_{emp}(x_k^*)}{\sqrt{N(\text{int}(x_{(k)}) + 1) - N(\text{int}(x_{(k+1)}) + 1)}}.$$

For an alternative approach, see Sec. 7.3 of Ref. [83].

Finally, notice that the estimation of the empirical mass function does not play any role in the fitting and testing procedures, and it is shown in the plots just for illustrative purposes. There, we compare $g_{emp}(x)$, defined for $x \geq 1$, with the fit $g(x)$ defined for $x \geq a$; then, a correction constant is applied to the latter so that the fit is properly displayed. Thus $g_{emp}(x)$ is plotted together with $g(x)N_a/N$. In the case of the rank-size representation this means that we plot the empirical $n_{emp}(r)/L_{tot}$ together with the theoretical $(n(r)/L)(L/L_{tot}) = n(r)/L_{tot}$.

[1] P. Bak. *How Nature Works: The Science of Self-Organized Criticality.* Copernicus, New York, 1996.

[2] J. P. Sethna, K. A. Dahmen, and C. R. Myers. Crackling noise. *Nature*, 410:242–250, 2001.

[3] D. Sornette. *Critical Phenomena in Natural Sciences.* Springer, Berlin, 2nd edition, 2004.

[4] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Math.*, 1 (2):226–251, 2004.

[5] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Cont. Phys.*, 46:323–351, 2005.

[6] M.V. Simkin and V.P. Roychowdhury. Re-inventing Willis. *Physics Reports*, 502(1):1–35, 2011.

[7] M. P. H. Stumpf and M. A. Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[8] A.-L. Barabási. Love is all you need. Clauset's fruitless search for scale-free networks. *https://www.barabasilab.com/post/love-is-all-you-need*, 2018.

[9] P. Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature Comm.*, 10:1016, 2019.

[10] H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. *Eur. Phys. J. B*, 58:167–173, 2007.

[11] E. P. White, B. J. Enquist, and J. L. Green. On estimating the exponent of power-law frequency distributions. *Ecol.*, 89:905–912, 2008.

[12] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661–703, 2009.

[13] A. Deluca and A. Corral. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophys.*, 61:1351–1394, 2013.

[14] A. Corral, F. Font, and J. Camacho. Non-characteristic half-lives in radioactive decay. *Phys. Rev. E*, 83:066103, 2011.

[15] Hanel R., Corominas-Murtra B., Liu B., and Thurner S. Fitting power-laws in empirical data with estimators that work for all exponents. *PLoS ONE*, 12(2):e0170920, 2017.

[16] I. Voitalov, P. van der Hoorn, R. van der Hofstad, and D. Krioukov. Scale-free networks well done. *Phys. Rev. Research*, 1:033034, 2019.

[17] A. Corral and A. González. Power law distributions in geoscience revisited. *Earth Space Sci.*, 6(5):673–697, 2019.

[18] T. Hernández and R. Ferrer i Cancho. *Lingüística Cuantitativa.* El País Ediciones, Madrid, 2019.

[19] E. U. Condon. Statistics of vocabulary. *Science*, 67(1733):300–300, 1928.

[20] G. K. Zipf. *The psycho-biology of language.* Houghton Mifflin, Boston, 1935.

[21] G. K. Zipf. *Human behaviour and the principle of least effort.* Addison-Wesley, Cambridge (MA), USA, 1949.

[22] L. A. Adamic and B. A. Huberman. Zipf's law and the Internet. *Glottom.*, 3:143–150, 2002.

[23] C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Phys. Rev. Lett.*, 90:088102, 2003.

[24] R. L. Axtell. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, 2001.

[25] J. Serrà, A. Corral, M. Boguñá, M. Haro, and J. Ll. Arcos. Measuring the evolution of contemporary western popular music. *Sci. Rep.*, 2:521, 2012.

[26] H. Baayen. *Word Frequency Distributions.* Kluwer, Dordrecht, 2001.

[27] M. Baroni. Distributions in text. In A. Lüdeling and M. Kytö, editors, *Corpus linguistics: An international handbook, Volume 2*, pages 803–821. Mouton de Gruyter, Berlin, 2009.

[28] S. T. Piantadosi. Zipf's law in natural language: a critical review and future directions. *Psychon. Bull. Rev.*, 21:1112–1130, 2014.

[29] Y. Malevergne, V. Pisarenko, and D. Sornette. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Phys. Rev. E*, 83:036111, 2011.

[30] D. Zanette. *Statistical patterns in written human language.* 2012.

[31] L. Lü, Z.-K. Zhang, and T. Zhou. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE*, 5(12):e14139, 12 2010.

[32] B. Mandelbrot. On the theory of word frequencies and on related Markovian models of discourse. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 190–219. American Mathematical Society, Providence, RI, 1961.

[33] R. Ferrer i Cancho and R. V. Solé. Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited. *J. Quant. Linguist.*, 8(3):165–173, 2001.

[34] A. Corral, F. Udina, and E. Arcaute. Truncated lognormal distributions and scaling in the size of naturally defined population clusters. *Phys. Rev. E*, 101:042312, 2020.

[35] I. Moreno-Sánchez, F. Font-Clos, and A. Corral. Large-scale analysis of Zipf's law in English texts. *PLoS ONE*, 11(1):e0147073, 2016.

[36] A. Corral, R. Garcia-Millan, N. R. Moloney, and F. Font-Clos. Phase transition, scaling of moments, and order-parameter distributions in Brownian particles and branching processes with finite-size effects. *Phys. Rev. E*, 97:062156, 2018.

[37] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford UP, Oxford, 2001.

[38] D. Zanette and M. Montemurro. Dynamics of text generation with realistic Zipf's distribution. *J. Quant. Linguist.*, 12(1):29–40, 2005.

[39] W. Li, P. Miramontes, and G. Cocho. Fitting ranked linguistic data with two-parameter functions. *Entropy*, 12(7):1743–1764, 2010.

[40] G. J. Stephens and W. Bialek. Statistical mechanics of letters in words. *Phys. Rev. E*, 81:066119, 2010.

[41] J. Baixeries, B. Elvevåg, and R. Ferrer-i-Cancho. The evolution of the exponent of Zipf's law in language ontogeny. *PLoS ONE*, 8(3):e53227, 2013.

[42] J. Kwapień and S. Drozdz. Physical approach to complex systems. *Phys. Rep.*, 515:115–226, 2012.

[43] M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, 2013.

[44] Christian Bentz, D. Kiela, F. Hill, and P. Buttery. Zipf's law and the grammar of languages: A quantitative study of old and modern English parallel texts. *Corpus Linguistics and Ling. Theory*, 10:175–211, 2014.

[45] J. Tuldava. The frequency spectrum of text and vocabulary. *J. Quantitative Linguistics*, 3(1):38–50, 1996.

[46] V. K. Balasubrahmanyan and S. Naranan. Quantitative linguistics and complex system studies. *J. Quantitative Linguistics*, 3(3):177–228, 1996.

[47] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, 2:943, 2012.

[48] R. Ferrer-i-Cancho and R. Gavaldà. The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Association for Information Science and Technology*, 60(4):837–843, 2009.

[49] E. G. Altmann and M. Gerlach. Statistical laws in linguistics. In M. D. Esposti, E. G. Altmann, and F. Pachet, editors, *Creativity and Universality in Language. Lecture Notes in Morphogenesis.* Springer, 2016.

[50] F. Font-Clos, G. Boleda, and A. Corral. A scaling law beyond Zipf's law and its relation with

Heaps' law. *New J. Phys.*, 15:093033, 2013.

[51] H. S. Heaps. *Information retrieval: computational and theoretical aspects.* Academic Press, 1978.

[52] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions.* Dover, New York, 1965.

[53] H. A. Simon. On a class of skew distribution functions. *Biomet.*, 42:425–440, 1955.

[54] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions.* Wiley-Interscience, New Jersey, 3rd edition, 2005.

[55] N. Batir. New inequalities for the Hurwitz zeta function. *Proc. Indian Acad. Sci. (Math. Sci.)*, 118(4):495–503, 2008.

[56] O. Peters, A. Deluca, A. Corral, J. D. Neelin, and C. E. Holloway. Universality of rain event size distributions. *J. Stat. Mech.*, P11030, 2010.

[57] A. Corral, A. Deluca, and R. Ferrer-i-Cancho. A practical recipe to fit discrete power-law distributions. *ArXiv*, 1209:1270, 2012.

[58] M. Gerlach and E. G. Altmann. Testing statistical laws in complex systems. *Phys. Rev. Lett.*, 122:168301, 2019.

[59] L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.

[60] K. Christensen and N. R. Moloney. *Complexity and Criticality.* Imperial College Press, London, 2005.

[61] A. Deluca and A. Corral. Scale invariant events and dry spells for medium-resolution local rain data. *Nonlinear Proc. Geophys.*, 21:555–567, 2014.

[62] A. N. Kolmogorov. *Foundations of the Theory of Probability.* Chelsea Publising Company, New York, 2nd edition, 1956.

[63] J.-P. Bouchaud and A. Georges. Anomalous diffusion in disordered media: statistical mechanisms, models and physical applications. *Phys. Rep.*, 195:127–293, 1990.

[64] A. Corral. Scaling in the timing of extreme events. *Chaos. Solit. Fract.*, 74:99–112, 2015.

[65] A. Corral and F. Font-Clos. Dependence of exponents on text length versus finite-size scaling for word-frequency distributions. *Phys. Rev. E*, 96:022318, 2017.

[66] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference. A practical information-theoretic approach.* Springer, New York, 2nd edition, 2002.

[67] S. T. Piantadosi. One parameter is always enough. *AIP Advances*, 8(9):095118, 2018.

[68] R. Ferrer i Cancho. The variation of Zipf's law in human language. *Eur. Phys. J. B*, 44:249–257, 2005.

[69] A. Corral and I. Serra. The brevity law as a scaling law, and a possible origin of Zipf's law for word frequencies. *Entropy*, page 224, 2019.

[70] D. Sornette. Dragon-kings, black swans and the prediction of crises. *Int. J. Terraspace Sci. Eng.*, 2(1):1–18, 2009.

[71] S. Hergarten. *Self-Organized Criticality in Earth Systems.* Springer, Berlin, 2002.

[72] S. Kunte and A. P. Gore. The paradox of large samples. *Current Sci.*, 62(5):393–395, 1992.

[73] T. Broderick, M. Dudík, G. Tkacik, R. E. Schapireb, and W. Bialek. Faster solutions of the inverse pairwise Ising problem. *arXiv*, 0712.2437, 2007.

[74] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *Eur. Phys. J. B*, 41:255–258, 2004.

[75] S. Pueyo and R. Jovani. Comment on "A keystone mutualism drives pattern in a power function". *Science*, 313:1739c–1740c, 2006.

[76] S. M. Burroughs and S. F. Tebbens. Power-law scaling and probabilistic forecasting of tsunami runup heights. *Pure Appl. Geophys.*, 162:331–342, 2005.

[77] G. Casella and R. L. Berger. *Statistical Inference.* Duxbury, Pacific Grove CA, 2nd edition, 2002.

[78] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in FORTRAN.* Cambridge University Press, Cambridge, 2nd edition, 1992.

[79] L. Vepstas. An efficient algorithm for accelerating the convergence of oscillatory series, useful for computing the polylogarithm and Hurwitz zeta functions. *Numer. Algor.*, 47:211–252, 2008.

[80] S.Naranan. "Power law" version of Bradford's law: Statistical tests and methods of estimation. *Scientomet.*, 17(3–4):211–226, 1989.

[81] R. D. Malmgren, D. B. Stouffer, A. E. Motter, and L. A. N. Amaral. A Poissonian explanation for heavy tails in e-mail communication. *Proc. Natl. Acad. Sci. USA*, 105:18153–18158, 2008.

[82] W. Hörmann and G. Derflinger. Rejection-inversion to generate variates from monotone discrete distributions. *ACM Trans. Model. Comput. Simul.*, 6(3):169–184, 1996.

[83] A. Turiel, C. J. Pérez-Vicente, and J. Grazzini. Numerical methods for the estimation of multifractal singularity spectra on sampled data: A comparative study. *J. Comp. Phys.*,

216(1):362–390, 2006.

[84] Note that, for a power law without upper truncation, the value of z can become colossal if the exponent α is close to one, as shown in Fig. 1, but this weird fact is not relevant for our purposes.