Volume 387, issue 1 1 January 2008 ISSN 0378-4371

PHYSICA A

STATISTICAL MECHANICS AND ITS APPLICATIONS

Recognized by the European Physical Society

Available online at
ScienceDirect
www.sciencedirect.com

http://www.elsevier.com/locate/physa

Available online at www.sciencedirect.com
ScienceDirect
Physica A 387 (2008) 84–96
PHYSICA A
www.elsevier.com/locate/physa

Tail universalities in rank distributions as an algebraic problem: The beta-like function

G.G. Naumis[a,*], G. Cocho[b]

[a]*Departamento de Fisica-Quimica, Instituto de Fisica, Universidad Nacional Autónoma de México, Apdo.Postal 20-364, 01000 México, D.F., Mexico*

[b]*Departamento de Sistemas Complejos, Instituto de Fisica, Universidad Nacional Autónoma de México, Apdo.Postal 20-364, 01000 México, D.F., Mexico*

Received 4 May 2007; received in revised form 4 July 2007
Available online 8 August 2007

Abstract

Although power laws of the Zipf type have been used by many workers to fit rank distributions in different fields like in economy, geophysics, genetics, soft-matter, networks, etc. these fits usually fail at the tail. Some distributions have been proposed to solve the problem, but unfortunately they do not fit at the same time the body and the tail of the distribution. We show that many different data in rank laws, like in granular materials, codons, author impact in scientific journal, etc. can be very well fitted by the integrand of a beta function (that we call beta-like function). Then we propose that such universality can be due to the fact that systems made from many subsystems or choices, present stretched exponential frequency-rank functions which qualitatively and quantitatively are well fitted with the beta-like function distribution in the limit of many random variables. We give a plausibility argument for this observation by transforming the problem into an algebraic one: finding the rank of successive products of numbers, which is basically a multinomial process. From a physical point of view, the observed behavior at the tail seems to be related with the onset of different mechanisms that are dominant at different scales, providing crossovers and finite size effects.

PACS: 89.75.Fb; 87.10.+e; 89.75.Da; 89.65.Gh; 89.65.−s; 87.23.Cc

Keywords: Ranking distributions; Power law distribution; Zipf law; Multiplicative processes

1. Introduction

Both natural language texts and DNA sequences present power laws [1,2] in the observed frequency of a word as a function of its rank (r), where the rank is just the ordinal position of a word if all words are ordered according to their decreasing frequency. Usually, the most frequent word has rank 1, the next most frequent rank 2 and so on. This power law behavior of the ranking is known as the Zipf law, and it is very common in physics, biology, geography, economics, linguistics, etc. [3]. In physics one can cite the rank distribution of

*Corresponding author. Tel.: +555 5622 51 74; fax: +555 5622 50 08.
E-mail address: naumis@fisica.unam.mx (G.G. Naumis).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2007.08.002

stick–slip events in sheared granular media [4], earthquakes [4], radionuclides half-life time and nuclides mass number [5]. Many complex systems share as well the same phenomenology, as happens in networks [6], biological clocks [7] and metabolic networks [8]. Zipf discovered his rank law by analyzing manually the frequencies of 29,899 different words types in the novel "Ulysses" by James Joyce. When a larger set of words is considered, a deviation from a power law is observed for larger ranks [9]. A similar behavior is found in genetic sequences [1]. Deviations from the Zipf law are also found in the tails ranking of many physical systems [10]. In fact, is clear that one should expect a different behavior at the tails, since finite size effects are always present and the power law must be "stopped" at a certain region. In spite of this, many workers just ignore the tail effects by fitting the data in a restricted range, or they proceed in a very questionable way by fitting all the data with a power law. Others have fitted sets of data in nature and in economy with stretched exponentials [10] and log-normal distributions [11]. The problem with the previous expressions is that they do not fit the data at *both ending tails*, where different kinds of processes are set in once a crossover region is reached. Such crossovers are due to finite size effects, in which different mechanisms are set in when certain big and small scales are reached. This leads to the idea of using multiscaling physical modelling to understand such features. Maybe the best example of the previous situation occurs in turbulence, where Kolmogorov's power law is observed only in the inertial regimen [12]. In one tail (small length scales) energy dissipation plays the main role, while energy injection dominates at big scales [13]. For each of these limits, the scaling behavior is different [14,15]. One can conjecture that similar ideas are behind many other complex physical systems, since we report that many rank laws are well parametrized with the formula,

$$f(r) = K\frac{(R-r+1)^b}{r^a}, \tag{1}$$

where a and b are fitted from the data, r is the rank and R is the maximal r. If $f(r)$ is normalized to 1, then $K \equiv r^a/\sum_{r=1}^{R}(R-r+1)^b$. For $R \gg 1$, K can be transformed into an integral that yields $K \sim \Gamma(b-a+2)/\Gamma(1-a)\Gamma(1+b)$. We will propose that $f(r)$ is related with the ranking of multinomial events, in which a and b seem to be parameters that determine the onset of different mechanisms that operate at different scales. Notice that Eq. (1) is similar to the integrand of a beta function, and thus in what follows we will call it a modified beta-like function. Our work is in the same spirit of Moyano et al. [16] in the sense that we try to develop general properties of systems that are built from many subsystems or choices [17]. The outline of this paper is the following: in Section 2 we present some representative examples of the phenomenology that we have observed. In Section 3 we show how this phenomenology can be studied as a problem of hierarchies in the product of random variables, and then transformed into a related algebraic problem: the ranking of a set of numbers produced by the iterative product of an initial finite set of numbers. In Section 4, we analyze the proposed problem, and finally, in Section 5 we give the conclusions of this work.

2. Phenomenology of rank laws and the beta-like function

As starting point, we will provide some representative results of the wide phenomenology found in the tails of rank laws. We start with an example from geography. We took the population of municipalities and departments of different countries, and ranked them in order of decreasing population. Fig. 1 shows a semilog plot of the corresponding population rank of four representative municipalities in Mexico and Spain. For each state or department, a fit using Eq. (1) was made. Such fits are given by solid lines in the figure. The agreement is very good, with a correlation coefficient $\mathscr{R}^2$ bigger than 0.986 for all fits. The values of a and b for each fit are shown in the inset of the plot, and their numerical values are given in Table 1, with the corresponding correlation coefficients. We have verified that similar good results are obtained for the population of countries around the word.

As a second example, we ranked the impact factor of scientific journals from different fields. Fig. 2 shows the logarithm of the impact factor against the rank of scientific journals, taken from a recent study [18], compared with the fits given by Eq. (1). Again, all of the fits are excellent, with correlation coefficients above 0.998.

Similar good fitting results are obtained in genetics. Here we took the codon frequencies in genomes of different organisms and ranked each kind of codon according to its frequency. In Fig. 3, we plot the logarithm

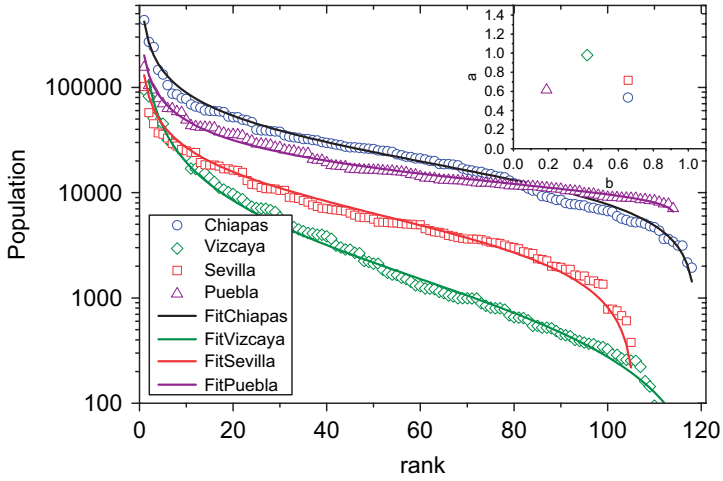

Fig. 1. Population ranking of four representative municipalities from Mexico and Spain. The solid lines are the fits obtained from Eq. (1). The inset presents the corresponding values of a and b used in the fits. The correlation coefficients $\mathscr{R}^2$ are bigger than 0.986 in all cases.

Table 1
Numerical values of a and b and the correlation coefficients $\mathscr{R}^2$ for all of the examples presented in this work

Example	a	b	R^2
Chiapas	0.534	0.654	0.999
Puebla	0.616	0.190	0.999
Sevilla	0.716	0.656	0.988
Vizcaya	0.979	0.421	0.986
Agroscience	0.221	0.959	0.999
Computer science	0.284	1.063	0.999
Physics	0.406	0.991	0.998
Ch. Tracho	0.220	0.501	0.991
E. Coli	0.247	0.503	0.998
Homo Sapiens	0.164	0.365	0.989
Jannasch	0.370	1.243	0.978
Slip–stick events	1.080	0.401	0.991

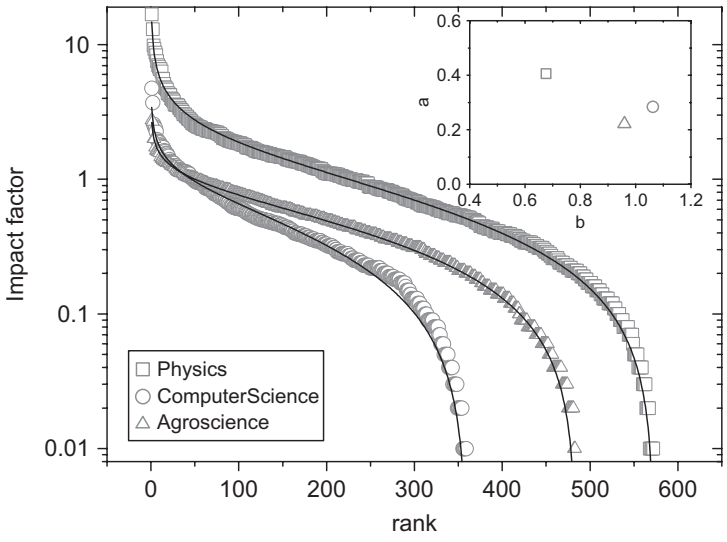

Fig. 2. Impact factor as a function of the rank for physics, computer science and agroscience. Fits using the beta-like function are shown as solid lines. Inset: values of a and b. The correlation coefficients are bigger than $\mathscr{R}^2 = 0.998$.

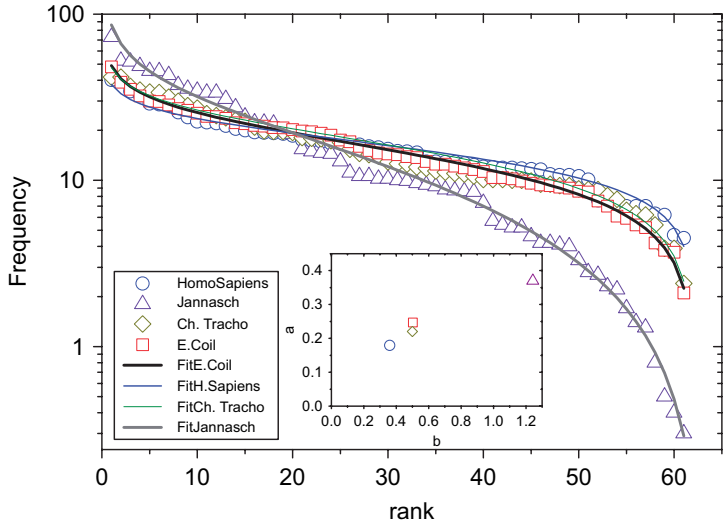

Fig. 3. Frequency of codons (normalized to 1000) as a function of the rank for the genome of four different species, with their corresponding fits shown as solid lines. Inset: values of a and b used for the fits in the beta-like distribution. The correlation coefficients are bigger than $\mathscr{R}^2 = 0.978$.

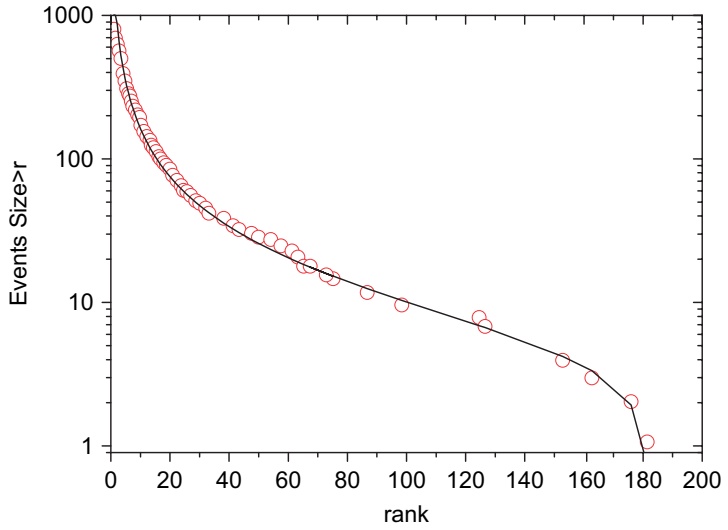

Fig. 4. Rank-ordered distribution of stick–slip events in a slowly sheared granular media. Circles are data taken from Ref. [4], and the solid line is a fit using Eq. (1), with $a = 1.08$ and $b = 0.40$. The correlation coefficient is $\mathscr{R}^2 = 0.991$.

of codon frequencies (normalized to 1000) as a function of the rank for different representative organisms, taken from a genome database [19]. For all the organisms, the resulting correlation parameters are bigger than 0.978.

Now we turn our attention to physics. In Fig. 4 we plot the rank-ordered distribution of stick–slip events in a slowly sheared granular media taken from Ref. [4], fitted using Eq. (1). Although a modified power law was proposed in Ref. [4] to explain the results, the present fit gives a better correlation coefficient, $\mathscr{R}^2 = 0.991$.

All of these results are summarized in Table 1, where we present the numerical values of a and b and the correlation coefficients for all of the previous examples.

Here we presented examples of four different fields in which the beta-like function appears, but Eq. (1) can be used with excellent results in order to correct the Gutenberg–Ritcher law in earthquakes ranking, Bénard convection cells and in other fields, like architecture, music or road networks [20]. For example, in music we have found [21] that Eq. (1) produces a good fit for around 2000 different musical works (including classical,

jazz and rock) if the notes are ranked according to their statistical frequency. The actual values of a and b depend on the composer and on the type of composition. In the case of Bach, a clear pattern is observed for minor and major modes. For road networks, the ranking of distances between main cities in Mexico also follows the beta-like function.

As was explained in the introduction, there are many fitting functions that have been proposed by others [10,11]. Some of these functions use only one fit parameter, like the Lavalette's law, and some others use two, like the Yule–Simon distribution [22]. Since our function uses two fit parameters, plus normalization, is clear that in general, it will provide a better fit than one-parameter functions but at the expense of an increased number of parameters. However, as was explained in the introduction, it seems that in general we can expect that different physical mechanisms are set in once different scales are reached. Thus, in order to give the scaling for each different physical mechanism, one needs at least two-parameters. For example, in turbulence the cross-overs are two, the one that determines energy injection versus inertial regimen, and the other scale at which the inertial behavior becomes dissipative. Plus, one needs a critical exponent for the inertial regimen.

The comparison of Eq. (1) with others two-parameter fits is qualitative. For example, the Yule–Simon distribution can be used to reproduce a Zipf law, but it introduces an exponential cutoff in the upper tail [22]. The stretched exponentials [10] and log-normal distributions [11] usually reproduce one of the tails but not the other. Usually, such deviations do not change in a dramatic way the correlation coefficient with respect to Eq. (1) since the tails do not have a great impact upon this coefficient. Compared with other two-parameter distributions, the main difference is that Eq. (1) captures the qualitative behavior of the system, which seems to be related with a hierarchy in multinomial events.

3. Hierarchy in a multiplicative stochastic process

The previous section leads to the conclusion that the tails of the ranking present some degree of universality, and Eq. (1) seems to be an excellent fitting function due to the fact that it gives the right shape of the curve. Also, it is simple and can be reduced to a pure power law by using an appropriate choice of a and b. As the $\{a, b\}$ distribution is indeed ubiquitous, one can try to associate it to some generic mechanism.

In the dynamics of population, scientific journal impact factor, codon usage and stick–slip events, there are many important issues that determine the behavior. In the case of the impact factor, we can cite for example the ability to select a good problem for investigation, the gift for writing clear papers, etc. Similar comments would be valid for the dynamics of granular media, as well as in economy, linguistics, genetics, etc. All of the previous systems share a common feature: their complex nature, i.e., they are build from many subsystems or path choices that produce a final result. One can try to model such complexity as follows. Consider a system made from N identical subsystems, where each can have s different states or choices with probability p_j, and $j = 1, \ldots, s$. When N such subsystems are put together, the state space consists of all s^N possible sequences of length N. If we do not care about the order of the choices or states in the string, there are just $(N + s - 1)!/(s - 1)!N!$ *different combinations*. For example, if a system is made from $N = 2$ subsystems, where each has two states or choices, say 1 or 0, the possible global states are $(0, 0)$, $(1, 0)$, $(0, 1)$ and $(1, 1)$, while there are only three combinations: $(0, 0)$, $(1, 1)$ and $(1, 0)$, the last one has multiplicity 2. Each combination has a certain probability that we call *reduced probability* $x_N(n_1, n_2, \ldots, n_s)$, where n_j is the number of subsystems in the j-esim state. The multiplicity of each different state is given by the multinomial coefficient $N!/(n_1!n_2!n_3! \ldots n_s!)$. The probability of a global state of the whole system is,

$$P_N(n_1, n_2, \ldots, n_s) = \frac{N!}{n_1!n_2!n_3! \ldots n_s!} x_N(n_1, n_2, \ldots, n_s) \tag{2}$$

with $n_1 + n_2 + n_3 + \cdots + n_s = N$. Notice that $P_N(n_1, n_2, \ldots, n_s)$ is a multinomial distribution function, which has well known properties. However, we are interested in the rank of the observed different values of the macrostates, not in the distribution of probability. To tackle this problem, we notice that each value $x_N(n_1, n_2, \ldots, n_s)$ corresponds to a *different macrostate of the system*. In our example, the states $(0, 1)$ and $(1, 0)$ produce the same global macrostate. These two internal states lead to one global state that has the same characteristics. If one assume that a certain characteristic (X) of a process or object is a function of $n_1, n_2, \ldots, n_s$, then each value of $X(n_1, n_2, \ldots, n_s)$ can be mapped to $x_N(n_1, n_2, \ldots, n_s)$ and $X(n_1, n_2, \ldots, n_s) =$

$X(x_N(n_1, n_2, \ldots, n_s))$. From the previous considerations, is clear that any rank hierarchy of $x_N(n_1, n_2, \ldots, n_s)$ will be inherited to $X(n_1, n_2, \ldots, n_s)$, but the actual hierarchy of X will also depend in the functional form of $X(x_N(n_1, n_2, \ldots, n_s))$. Is clear that the general problem cannot be solved without a reasonable assumption for this functional form. Here we will suppose that $X(x_N(n_1, n_2, \ldots, n_s))$ can be expressed as a power series in $x_N(n_1, n_2, \ldots, n_s)$,

$$X(x_N(n_1, n_2, \ldots, n_s)) = X_0 + X_1 x_N(n_1, n_2, \ldots, n_s) + \ldots, \tag{3}$$

where X_0 and X_1 are constants. Up to first order, this assumption means that X is proportional to $x_N(n_1, n_2, \ldots, n_s)$. Here X_1 plays the role of a susceptibility, as in any linear response theory. Under this approximation, the rank features of a system are reduced to study the hierarchy present in $x_N(n_1, n_2, \ldots, n_s)$. Observe that a priory, it is difficult to know if a system follows such approximation, but the assumption can be tested a posteriori once the ranking of X is obtained and compared with the experimental results.

To study the hierarchy of $x_N(n_1, n_2, \ldots, n_s)$, there are two cases. In the first, the subsystems are independent, as in a Bernoulli process,

$$x_N(n_1, n_2, \ldots, n_s) = p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_s^{n_s} \tag{4}$$

and the other is the general case of interacting subsystems, in which the addition of a new subsystem leads to a functional relationship of the type,

$$x_{N+1}(n_1, n_2, \ldots, n_s) = f(x_N(n_1, n_2, \ldots, n_s)). \tag{5}$$

In the following section, we will only consider the case of independent subsystems, in which no extra information is needed in order to model the system. This allows to produce the beta-like function in a simple form.

4. The rank hierarchy as an algebraic problem

For independent subsystems, an inspection of Eq. (4) shows that the rank structure can be reduced to the following algebraic problem. Take s numbers $p_1, p_2, \ldots, p_s$ at random (normalization can be imposed at the end of the process), labeled in such a way that $p_1 > p_2 > \cdots > p_s$, and multiply once each number by all the numbers in the set. With these resulting numbers, repeat the process N times to obtain a set of numbers that have the form $p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_s^{n_s}$, where $n_1 + n_2 + \cdots + n_s = N$. If the resulting numbers are arranged in decreasing magnitude, we can assign a rank (r) to each one according to its order in the hierarchy. The rank $r = 1$ is assigned to p_1^N, while the lowest rank $r = R$ corresponds to p_s^N. For example, chose at random three numbers p_1, p_2 and p_3 and form all the possible products: $p_1^2, p_1p_2, p_1p_3, p_2^2, p_2p_3, p_3^2$. We remark again that the events of the type p_1p_2 and p_2p_1 are considered as the same, since as was explained in the previous section, we are only interested in the value of the observed number. The corresponding multiplicity can be obtained from the multinomial distribution. In Fig. 5, we present a plot of $\log\, x_N(n_1, n_2, n_3)$ as a function of r for $N = 77$ and $p_1 = 0.5202$, $p_2 = 0.3125$ and $p_3 = 0.1673$. For obtaining this graph, we used arbitrary precision for the products. Then, the logarithm of each number was calculated. A second version of the program was made using an algebraic procedure that we will discuss later, and the results were completely equivalent. Fig. 5 shows the resulting ranking, and it is interesting to notice that they already display the shape presented in the phenomenology of real systems. In fact the results can be fitted by the same two-parameter beta-like function, with $a = 9.36 \pm 0.2$ and $b = 10.52 \pm 0.2$, with a correlation coefficient of 0.972. The quality of the fitting is improved as we consider more than three numbers, i.e., when $s \to \infty$. The message from this numerical experiment is simple: if this product is seen as a multiplicative process where each number is the probability of making a certain choice or state in a process, then each possible result has a well determined hierarchy.

The task that remains is how to calculate $x_N(n_1, n_2, \ldots, n_s)$ in terms of the rank. The problem is more easily solved using the logarithm of $x_N(n_1, n_2, \ldots, n_s)$,

$$\log\, x_N(n_1, n_2, \ldots, n_s) = n_1 \log\, p_1 + n_2 \log\, p_2 + \cdots + n_s \log\, p_s. \tag{6}$$

Each set of values $(n_1, n_2, \ldots, n_s)$ is a point with integer coordinates in a s-dimensional space. Since $n_1 + n_2 + \cdots + n_s = N$, all the points are in a subspace of dimension $s - 1$. The problem of ranking is reduced

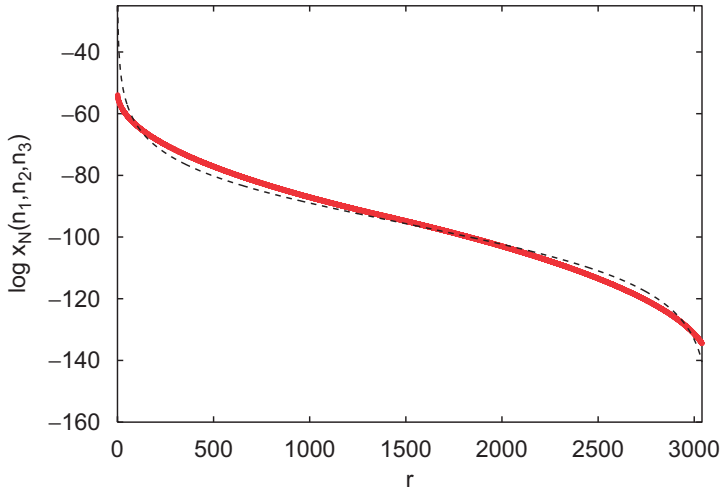

Fig. 5. Successive products of three numbers $p_1 = 0.5202$, $p_2 = 0.3125$, $p_3 = 0.1673$ as a function of the rank (bold solid line) for $N = 77$, and a fitting using Eq. (1), with $a = 9.36$, $b = 14.53$.

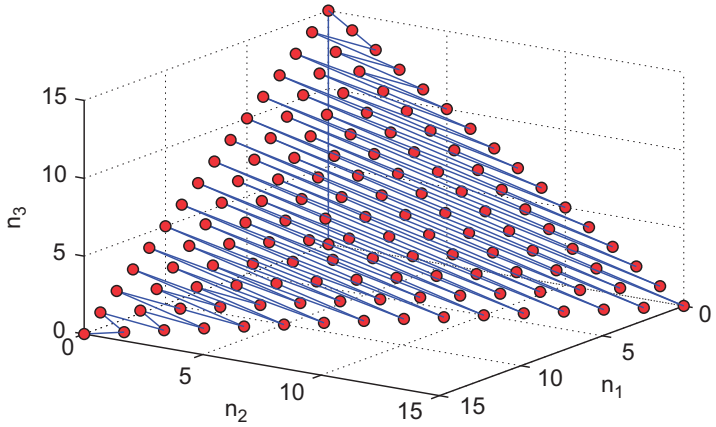

Fig. 6. Path of decreasing rank in the n_1, n_2 and n_3 space, for $N = 15$ and three random numbers $p_1 = 0.5202$, $p_2 = 0.3125$, $p_3 = 0.1673$.

to find a path between the maximal rank point (with coordinates $(N, 0, 0, \ldots, 0)$) to the minimum $(0, 0, 0, \ldots, N)$ in such a way that $\log\, x_N(n_1, n_2, \ldots, n_s)$ decreases in each step. For $s = 2$, the solution is easy to find. Using that $n_1 + n_2 = N$,

$$x_N(n_1, n_2) = x_N(n_2) = p_1^{N-n_2} p_2^{n_2}, \tag{7}$$

it follows that the range is given by $r = n_2 + 1$. Then,

$$x_N(r) = p_1^N \left(\frac{p_2}{p_1}\right)^{r-1} = p_1^N\, \mathrm{e}^{-A(r-1)} \tag{8}$$

with $A = |\ln(p_2/p_1)|$. Eq. (8) shows that the numbers decrease in an exponential way as a function of the rank.

The case $s = 3$ can be easily visualized in Fig. 6, where the points in the integer lattice defined by Eq. (6) are shown as circles.

A path between points of decreasing $\log\, x_N(n_1, n_2, \ldots, n_s)$ is indicated as a line that joins the lattice points in Fig. 6, for a given set of numbers p_1, p_2 and p_3. Fig. 7 shows how the values of n_1, n_2 and n_3 vary as a function of the range. A very complicated oscillatory pattern is seen, although a well defined envelope is also observed.

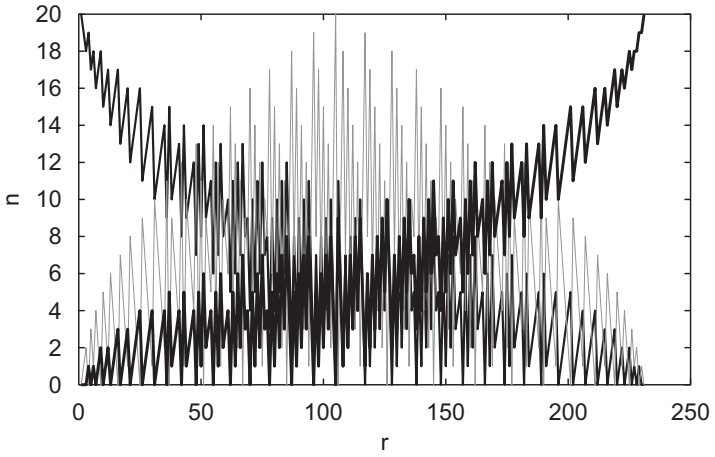

Fig. 7. Values of n_1 (thin solid line), n_2 (gray line) and n_3 (solid bold line) as a function of the rank, for $N = 20$ and $p_1 = 0.5202$, $p_2 = 0.3125$, $p_3 = 0.1673$.

This envelope is in fact the key to solve the problem, since it is the responsible of the ranking behavior. Notice also that all paths always start at $(N, 0, 0)$ and finish at $(0, 0, N)$, since $\log p_1 > \log p_2 > \log p_3$.

In general, since the index n_j is a function of the rank r, we can write that $n_j = n_j(r)$ where r is just the number of steps used to go from the point $(N, 0, \ldots, 0)$ to a certain $(n_1, n_2, n_3, \ldots, n_s)$. It follows that,

$$\log x_N(r) = n_1(r) \log p_1 + n_2(r) \log p_2 + \cdots + n_s(r) \log p_s. \tag{9}$$

The task is reduced to find the functions $n_j(r)$ for a given set $\{p_j\}$. Consider again the case of an initial set of three numbers, $s = 3$. Using that $n_1 + n_2 + n_3 = N$, $\log x_N(r)$ can be written as

$$\log x_N(r) = N \log p_1 + n_2(r) \log \delta_{21} + n_3(r) \log \delta_{31} \tag{10}$$

with $\delta_{21} = p_2/p_1$ and $\delta_{31} = p_3/p_1$. The solution for any set p_1, p_2, p_3 is complicated, because some paths are not periodic. However, one can work out first the cases $p_1 \sim p_2 \gg p_3$ and $p_1 \gg p_2 \sim p_3$ that give insights about how to treat others.

Let us first consider the limit $p_1 \sim p_2 \gg p_3$, and $\delta_{21}^2 \gg \delta_{31}$. The corresponding path is easy to find because it is similar to an odometer with an increased range after each turn, as seen in Fig. 7, due to the hierarchy $1 > \delta_{21} > \delta_{21}^2 > \delta_{31} > \delta_{21}\delta_{31} > \delta_{31}^2 > \cdots > \delta_{31}^N$. For example, when $N = 2$ this leads to the following table that contains the number $x_N(r)$ as a function of the rank, and the corresponding path given by n_2 and n_3.

$x_N(r)$	n_2	n_3	r	$n_{2MAX}(r)$
p_1^2	0	0	1	–
$p_1^2\delta_{21}$	1	0	2	–
$p_1^2\delta_{21}^2$	2	0	3	2
$p_1^2\delta_{31}$	0	1	4	–
$p_1^2\delta_{21}\delta_{31}$	1	1	5	1
$p_1^2\delta_{31}^2$	0	2	6	0

The sequence of the path goes as follows, first $n_2(r)$ is increased one by one as n_3 remains constant, until it reaches a maximal value called $n_{2MAX}(r)$ which in fact determines the envelope of the ranking sequence and thus the basic shape of the curve $x_N(r)$ (the envelope that contains $n_{2MAX}(r)$ is shown in Fig. 8 as a dotted line). Once $n_2(r)$ increases from zero to $n_{2MAX}(r)$, a new cycle begins with $n_2(r+1) = 0$ and $n_3(r+1) = n_3(r) + 1$. As a result, the number of steps to reach $n_{2MAX}(r)$ from the maximal rank (R) is given by summing over all the number of steps made for each constant value of $n_3(r)$ (corresponding basically in Fig. (8) to a sum of all

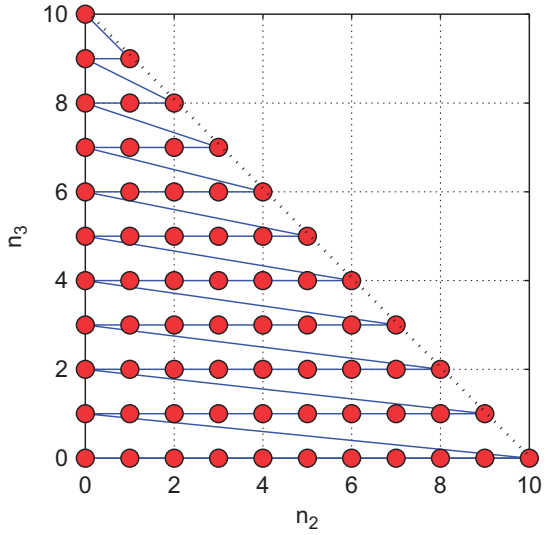

Fig. 8. Path of decreasing ranks in the n_2 and n_3 plane for $p_1 \sim p_2 \gg p_3$, where the n_1 coordinate was eliminated using that $n_1 + n_2 + n_3 = N$. The dotted line corresponds to all the $n_{2MAX}(r)$, which defines the envelope of the ranking sequence.

"row lengths" from the top of the triangle to $n_{2MAX}(r)$). This sum can be written as

$$R - r = \sum_{j=1}^{n_{2MAX}(r)} j = \frac{n_{2MAX}(r)(n_{2MAX}(r)+1)}{2}. \tag{11}$$

The previous equation is quadratic in $n_{2MAX}(r)$, and can be solved in terms of r and R, to give,

$$n_{2MAX}(r) = \frac{-1 \pm \sqrt{1 + 8(R-r)}}{2}. \tag{12}$$

One can verify that Eq. (12) agrees with the table when the positive sign is used, since $n_{2MAX}(3) = 2$, $n_{2MAX}(5) = 1$ and $n_{2MAX}(6) = 0$. However, R and N are not independent. If for example we analyze the first three rows of the table, is clear that in general, for $r = N + 1$ the value of $n_{2MAX}(r)$ is N. On the other hand, Eq. (12) must also be satisfied, from where it follows that,

$$n_{2MAX}(N+1) = N = \frac{-1 + \sqrt{1 + 8(R-N-1)}}{2}, \tag{13}$$

thus,

$$8 = \frac{(2N+1)^2 - 1}{(R-N-1)}. \tag{14}$$

The previous expression can be inserted into Eq. (12), and since $1 \ll N \ll R$, we obtain that the leading term of $n_{2MAX}(r)$ is

$$n_{2MAX}(r) \approx N\left(1 - \frac{(r-1)}{R}\right)^{1/2}. \tag{15}$$

We have verified that the previous equation is in excellent agreement with the numerical results. The corresponding value of $n_3(r)$ can be obtained from the condition $n_2 + n_3 \leqslant N$. Finally, the number as a function of the rank is given by,

$$x_N(r) \approx \left[p_1 \left(\frac{p_2}{p_1}\right)^{\left(1-(r-1)/R\right)^{1/2}} \left(\frac{p_3}{p_1}\right)^{1-\left(1-(r-1)/R\right)^{1/2}} \right]^N. \tag{16}$$

Fig. 9 shows the excellent agreement between Eq. (16) and the curve obtained for $p_1 = 0.5250$, $p_2 = 0.4250$, $p_3 = 0.000047$. Furthermore, Eq. (16) can be written as a stretched exponential as follows:

$$x_N(r) \approx p_3^N \exp\left[D\left(1 - \frac{(r-1)}{R}\right)^{1/2}\right] \tag{17}$$

with $D = N|\log(p_2/p_3)|$ and R is the maximal value of r.

The case $p_1 \gg p_2 \sim p_3$ can be tackled in a similar way. The result is,

$$x_N(r) \approx p_1^N \exp\left[-E\left(\frac{r}{R}\right)^{1/2}\right] \tag{18}$$

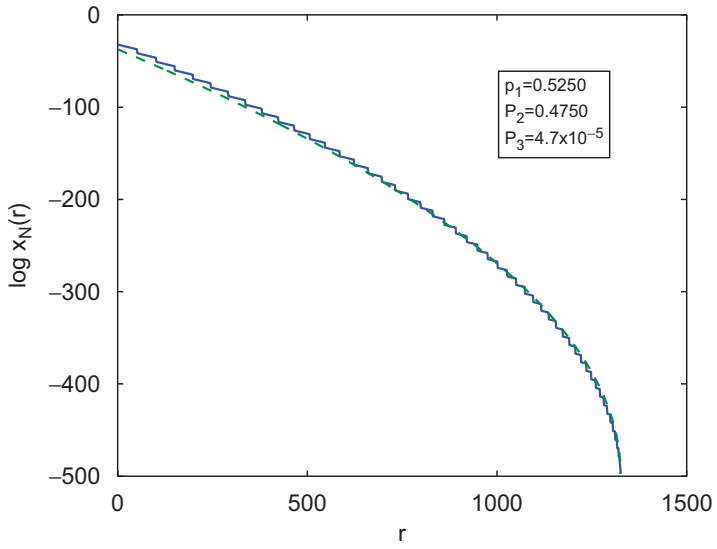

Fig. 9. Numerical results for the ranking of the successive product of three numbers such that $p_1 \sim p_2 \gg p_3$. The dotted line is the prediction using Eq. (16).

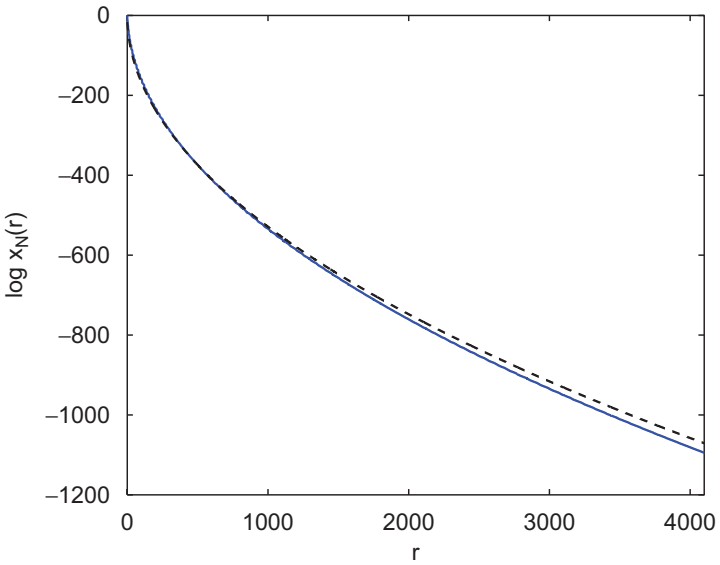

Fig. 10. Ranking of the successive product of three numbers such that $p_1 \gg p_2 \sim p_3$, for $p_1 = 0.99999$, $p_2 = 6.2 \times 10^{-6}$, $p_3 = 3.8 \times 10^{-6}$. The dashed line is the prediction made from Eq. (18), compared with the numerical result for $N = 100$ iterations (solid line).

with $E = N(\log(p_1/p_3) - \log(p_2/p_3))$, and as shown in Fig. 10, the agreement is also good, specially for low values of r.

Now consider the general case in which p_1, p_2 and p_3 have the same order of magnitude, as in Fig. 5, where two tails appears, one for small r and the other at r near R. The tail at low r is produced basically by the hierarchy in the biggest probabilities, i.e., by numbers where $n_1 \sim N$. In a similar way, the tail for r near R is produced by the lowest probability hierarchy, $n_3 \sim N$. The main effect in these tails when $p_1 \approx p_2 \approx p_3$ is that the sequence of ordering is not uniform as can be observed in Fig. 6, for which a very complicate path appears. As a result, Eq. (11) changes with the appearance of new subcycles in the rank path. These changes are the result of the increasing number of cycles in the odometer that we have discussed, as is also clear in the exponents that are transformed from 1 to $\frac{1}{2}$ as s goes from $s = 2$ to 3. Then we propose that Eq. (17) can be transformed into a generalized expression,

$$x_N(r) \approx p_3^N \exp\left[D\left(1 - \frac{(r-1)}{R}\right)^{\beta}\right] \tag{19}$$

in which β is a yet unknown exponent, always less than 1. Although we do not have a general proof in order to get analytically the value of β, is clear that the functional form is due to the way in which the area of the triangular surface defined by $n_1 + n_2 + n_3 = N$ is filled. In a similar way, Eq. (18) should be replaced by,

$$x_N(r) \approx p_1^N \exp\left[-E\left(\frac{r}{R}\right)^{\alpha}\right] \tag{20}$$

with $\alpha < 1$. These generic exponents for the tails also appear for $s > 3$ since from the degree of the polynomial equivalent to Eq. (11), one can prove that $\alpha \leqslant 1/(s-1)$ and $\beta \leqslant 1/(s-1)$. A simple procedure to combine the tails represented by Eqs. (19) and (20) is obtained by making the observation that for a given tail, only one stretched exponential produces curved tails in a semi-log plot, while the other tends toward a constant, i.e., if we consider the derivative of Eq. (19):

$$\left(\frac{\mathrm{d}\,\ln\, x_N(r)}{\mathrm{d}r}\right) = -\frac{\beta D}{R}\left(1 - \frac{(r-1)}{R}\right)^{\beta-1} \tag{21}$$

is clear that $x'_N(r)$ is nearly a constant if $r \ll 1$, corresponding to the limit in which Eq. (20) has greater curvature. Analyzing the limit $r \to R$ gives a similar result,

$$\left(\frac{\mathrm{d}\,\ln\, x_N(r)}{\mathrm{d}r}\right) = -\frac{\alpha E}{R}\left(\frac{r}{R}\right)^{\alpha-1}. \tag{22}$$

From these considerations, a simple way to produce a function with the required dependences when $r \to R$ and $r \to 1$ is by proposing the following ansatz which reproduces both tails:

$$x_N(r) \approx C_1 \exp\left[D\left(1 - \frac{(r-1)}{R}\right)^{\beta}\right] \exp\left[-E\left(\frac{r}{R}\right)^{\alpha}\right], \tag{23}$$

where C_1 is a constant. A plot of the previous expression is presented in Fig. 11, showing the basic shape of the studied beta function.

Finally, Eq. (23) can be simplified when many states are present since $\alpha \leqslant 1/(s-1)$, $\beta \leqslant 1/(s-1)$ and for $s \gg 1$, $\alpha \to 0$ and $\beta \to 0$. Then, by using the observation about the derivatives that appears in Eqs. (21) and (22), one can approximate the derivatives like in Eq. (21) as follows:

$$\left(\frac{\mathrm{d}\,\ln\, x_N(r)}{\mathrm{d}r}\right) = -\frac{\beta D}{R}\left(1 - \frac{(r-1)}{R}\right)^{\beta-1} \approx -\frac{\beta D}{R}\left(1 - \frac{(r-1)}{R}\right)^{-1}. \tag{24}$$

A similar thing can be done in the tail $r \to 1$, for which α can be neglected with respect to one in Eq. (22). Combining both tails in a sole expression we get:

$$\left(\frac{\mathrm{d}\,\ln\, x_N(r)}{\mathrm{d}r}\right) \approx -\frac{\beta D}{R}\left(1 - \frac{(r-1)}{R}\right)^{-1} - \frac{\alpha E}{R}\left(\frac{r}{R}\right)^{-1}.$$

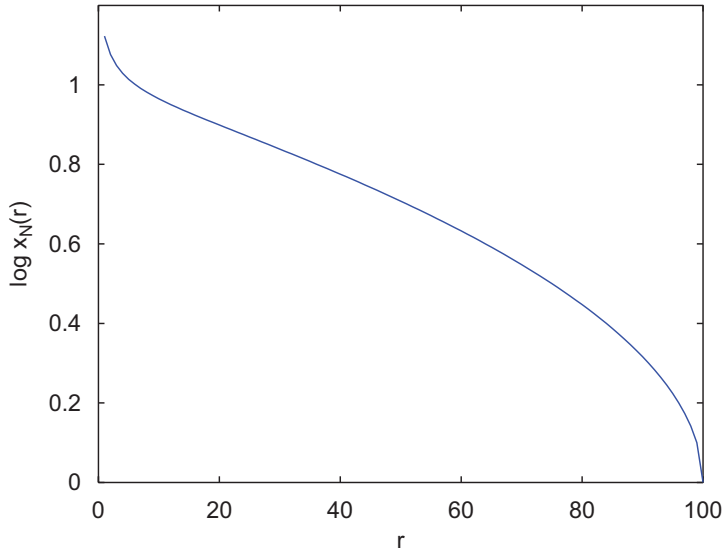

Fig. 11. A plot of Eq. (23) using $C_1 = 2$, $D = 1$, $E = 1$, $\beta = 1$ and $\alpha = 1$.

By integrating the previous equation, we finally obtain the beta-like function given by Eq. (1), where the exponents a and b are given by

$$a = \alpha E \quad \text{and} \quad b = \beta D. \tag{25}$$

Thus, the beta-like function is obtained when we have a large number of states in the system. Notice how the parameters a and b are determined mainly by the behavior in the tails.

5. Conclusions

In conclusion, we have proposed a simple formula that allows to fit many different rank phenomena. Although there are many formulas with one and two parameters that fit the observed ranking, the present one seems to give a better qualitative agreement. Furthermore, we have proposed that this formula arises as the result of ranking multinomial events. To do so, we considered an equivalent algebraic problem: finding the rank of a successive product of numbers. The case of an initial set with two numbers has been solved in a complete form, while for three numbers we have solved some particular cases. Using such solutions, we constructed an ansatz that agrees with extensive computer simulations. This construction is based on a detailed study of the rank at the tails, which is given by stretched exponentials. In the limit of an infinite set of numbers, the ansatz leads to the proposed beta-like function.

A task that remains is how to get the coefficients a and b from physical principles, using for example master equations and the concept of multiscaling modelling. To do so, simple models are required to shine some light on the problem. As an example, we can cite an ad-hoc expansion-modification algorithm in DNA models [23]. In such model, the beta-like function is also obtained for the ranking of codons, but $a > b$ if the expansion probability of the genetic code is bigger than the mutation rate. This particular model seems to confirm the ideas presented in this article, i.e., that a and b represent the relative influence of two general mechanisms, where each of them dominate at a given tail. According to some preliminary results, a seems to be related with a certain funnel type of energy landscape, as in protein folding, which leads to a deterministic sequence, while b is associated with a many valley landscape, as seen in spin glasses. This last opposite effect provides much more variability in the sequence of results. Such correlation is consistent with associating b to the stochastic component of the dynamics and a with the most deterministic features [23]. In future works, we will elucidate with more detail such mechanisms.

Acknowledgments

This work was supported by DGAPA-UNAM project IN-117806, CONACyT 48783-F and 50368. We thank G. Martínez-Meckler for a critical revision of the manuscript.

References

[1] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.K. Peng, M. Simons, H.E. Stanley, Phys. Rev. Lett. 73 (1994) 3169.
[2] S.C. Manrubia, D.H. Zanette, Phys. Rev. E 59 (1999) 4945.
[3] W. Li, Phys. Rev. E 43 (1991) 5240;
W. Li, ⟨http://www.nslij-genetics.org/wli/zipf/⟩, 2003.
[4] M. Bretz, R. Zaretzki, S.B. Field, N. Mitarai, F. Nori, Europhys. Lett. 74 (2006) 1116.
[5] G. Audi, O. Bersillon, J. Blachot, A.H. Wapstra, Nucl. Phys. A 624 (1997) 124.
[6] S. Fortunato, A. Flammini, F. Menczer, Phys. Rev. Lett. 96 (2006) 218701.
[7] A.C.C. Yang, S.S. Hseu, H.W. Yien, A.L. Goldberger, C.K. Peng, Phys. Rev. Lett. 90 (2003) 108103.
[8] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabasi, Nature 407 (2000) 651.
[9] H. Le Quan, E.I. Sicilia-García, J. Minj, F.J. Smith, Proceedings of the 17th International Conference on Computer Lingustics, Montreal, 2002.
[10] J. Laherrere, D. Sornette, Eur. Phys. J. B 2 (1998) 525.
[11] E.W. Montroll, M.F. Shlesinger, J. Stat. Phys. 32 (1983) 209.
[12] A.N. Kolmogorov, Dokl. Akad. Nauk SSSR 30 (1941) 299 (reprinted in Proc. R. Soc. London A 434 (1991) 9).
[13] A.N. Kolmogorov, Dokl. Akad. Nauk SSSR 32 (1941) 16 (reprinted in Proc. R. Soc. London A 434 (1991) 15).
[14] A. Kolmogorov, J. Fluid. Mech. 13 (1962) 82.
[15] Z. Warhaft, Annu. Rev. Fluid Mech. 32 (2000) 203.
[16] L.G. Moyano, C. Tsallis, M. Gell-Mann, Europhys. Lett. 72 (2006) 355.
[17] J.A. Marsh, M.A. Fuentes, L.G. Moyano, C. Tsallis, Physica A 372 (2006) 183.
[18] I. Popescu, Glottometrics 6 (2003) 83.
[19] Codon Usage Database, NCBI-GenBank ⟨http://www.kazuza.or.jp/codon⟩.
[20] G. Cocho, G. Martínez-Mekler, submitted for publication.
[21] M. Beltran, G. Cocho, G. Naumis, submitted for publication.
[22] C. Rose, D. Murray, D. Smith, Mathematical Statistics with Mathematica, Springer, New York, 2002, p. 107.
[23] R. Mansilla, G. Cocho, Complex Syst. 12 (2000) 207.