Incidence of *q* statistics in rank distributions

G. Cigdem Yalcin[a], Alberto Robledo[b], and Murray Gell-Mann[c,1]

[a]Department of Physics, Istanbul University, Vezneciler, Istanbul 34134, Turkey; [b]Instituto de Fisica y Centro de Ciencias de la Complejidad, Universidad Nacional Autonoma de Mexico, 01000 Mexico D.F., Mexico; and [c]Santa Fe Institute, Santa Fe, NM 87501

Contributed by Murray Gell-Mann, June 27, 2014 (sent for review May 29, 2014)

We show that size-rank distributions with power-law decay (often only over a limited extent) observed in a vast number of instances in a widespread family of systems obey Tsallis statistics. The theoretical framework for these distributions is analogous to that of a nonlinear iterated map near a tangent bifurcation for which the Lyapunov exponent is negligible or vanishes. The relevant statistical–mechanical expressions associated with these distributions are derived from a maximum entropy principle with the use of two different constraints, and the resulting duality of entropy indexes is seen to portray physically relevant information. Whereas the value of the index α fixes the distribution's power-law exponent, that for the dual index 2 − α ensures the extensivity of the deformed entropy.

rank-ordered data | generalized entropies

Zipf's law refers to the (approximate) power law obeyed by sets of data when these are sorted out and displayed by rank in relation to magnitude or rate of recurrence (1). The sets of data originate from many different fields: astrophysical, geophysical, ecological, biological, technological, financial, urban, social, etc., suggesting some kind of universality. Over the years this circumstance has attracted much attention and the rationalization of this empirical law has become a common endeavor in the study of complex systems (2, 3). Here we pursue further the view (4, 5) that an understanding of the omnipresence of this type of rank distribution hints at an underlying structure similar to that which confers systems with many degrees of freedom the familiar macroscopic properties described by thermodynamics. That is, the quantities used in describing this empirical law obey expressions derived from principles akin to a statistical–mechanical formalism (4, 5). The most salient result presented here is that the reproduction of the data via a maximum entropy principle indicates that access to its configurational space is severely hindered to a point that the allowed configurational space has a vanishing measure. This feature appears to be responsible for the entropy expression not to be of the Boltzmann–Gibbs or Shannon type but instead to take that of the Tsallis form (6), while the extensivity of entropy is preserved. It is perhaps worth clarifying that our study is set in discrete space and it does not consider any formal Hamiltonian system.

In Fig. 1 we show three examples of ranked data that appear to display power-law behavior along a considerably large interval of rank values. Fig. 1 (*Top*) shows data for the wealth of billionaires in the United States (7), Fig. 1 (*Middle*) shows data for the energy released by earthquakes in California (8), and Fig. 1 (*Bottom*) shows data for the intensity of solar flares (9). In Fig. 1 (*Left*), logarithmic scales are used for both size and rank, whereas Fig. 1 (*Right*) shows the same data in log-linear scales. Fig. 1 (*Left*) indicates approximate power-law decay for large rank and a clear deviation from this for small to moderate rank. As we shall show below, the theoretical description reproduces the data in Fig. 1 for the entire rank interval.

In *Distribution Functions that Generate Zipf's Law* we recall (4, 5, 10) the concise stochastic approach for raw data generated by a power-law distribution P for the size random variable N that yields an analytical expression for the size-rank distribution $N(k)$. This analytical expression involves a deformed exponential that has been shown to reproduce quantitatively real data and has as a limiting form the classical Zipf law (4, 5, 10). We also recall (4, 5) the analogy that exists between the stochastic approach and the deterministic nonlinear dynamics at and close to the tangent bifurcation. This analogy allows for a convenient description of finite-sized data that deviates from power-law behavior for both small and large rank. In *Rank Distributions from Maximum Entropy Principle* we derive the rank distribution $N(k)$ from a maximum entropy principle (MEP) and this allows us, via a well-known deformation index duality, to discuss two different entropy expressions of the Tsallis type obtained from two different sets of constraints (11–13). The values of the two entropy expressions coincide but they yield different information for the set of data under consideration. In *Statistical Mechanics of Contracted Configuration Space* we use this duality to discuss entropy extensivity of the ranked data and the presence of a strong phase-space contraction. This is shown to be the source of a generalized entropy that departs from the usual Shannon expression. This departure is extreme for the classical Zipf case, implying that the data can sample only a set of zero measure. Finally, in *Discussion* we discuss and summarize our results.

Distribution Functions That Generate Zipf's Law

A basic approach for the study of ranked data consists of three simply related distribution functions (4, 5, 10). The input is the distribution $P(N)$ of the data N under consideration, that is, it is assumed that the data are generated by a source described by $P(N)$ such that N can be thought of as a random variable. With no loss of generality we restrict N to take positive values within an interval $N_{\min} \le N \le N_{\max}$, where we allow for the limiting possibilities $N_{\min} = 0$ and/or $N_{\max} \to \infty$. The total number of data extracted from $P(N)$ is denoted by $\mathcal{N}$. Next, the (complementary) cumulative distribution $\Pi(N, N_{\max})$ is determined from $P(N)$,

$$\Pi(N, N_{\max}) = \int_{N}^{N_{\max}} P(N')dN', \qquad [1]$$

where the normalization of $P(N)$ implies $\Pi(N_{\min}, N_{\max}) = 1$. We can recover $P(N)$ from $\Pi(N, N_{\max})$,

Significance

The contents presented are of prime importance to the field of generalized statistical mechanics. We fulfill a longstanding need of exhibiting the kind of abundant real-world data that match the formal developments in this subject. These are size-rank distributions for which we provide a solid bridge between experimental data and theory. Also, this work delivers a working explanation for the existing duality between the two Tsallis-type entropy expressions that generalize the canonical expression. One relates to the distribution's power-law exponent whereas the other ensures entropy extensivity. The generalized entropies arise from a drastic reduction of configurations available to the system. We argue that this phase-space contraction is farthest for ranked data of the Zipf type.

Author contributions: G.C.Y., A.R., and M.G.-M. designed research; performed research; and wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: mgm@santafe.edu.

$$P(N) = -\frac{\partial}{\partial N}\Pi(N, N_{\max}). \quad \textbf{[2]}$$

By construction, the distribution $\Pi(N, N_{\max})$ sorts out data according to its magnitude: As N is decreased from $N_{\max}$ the distribution Π increases monotonically, taking values from $\Pi(N_{\max}, N_{\max}) = 0$ to $\Pi(N_{\min}, N_{\max}) = 1$, so it can be identified with $k/\mathcal{N}$, where k is the rank and $\mathcal{N}$ is the total number of data extracted from $P(N)$, and $k_{\max} = \mathcal{N}$. The last and third distribution is the size-rank function $N(k)$ and can be obtained by solving

$$\frac{k}{\mathcal{N}} = \int_{N(k)}^{N_{\max}} P(N')dN', \quad \textbf{[3]}$$

for $N(k)$. If k is to be an integer the possible lower limits in the integral in Eq. **3**, $N(1), N(2),\ldots,N(k_{\max})$ are such that the integral takes values $1/\mathcal{N}, 2/\mathcal{N},\ldots,k_{\max}/\mathcal{N}$.

If we make use of a power-law form for $P(N)$,

$$P(N) \sim N^{-\alpha}, 1 \le \alpha < \infty, \quad \textbf{[4]}$$

we have (4, 5, 10)

$$\Pi(N(k), N_{\max}) = \int_{N(k)}^{N_{\max}} N^{-\alpha}dN \quad \textbf{[5]}$$

$$= \frac{1}{1-\alpha}\left[N_{\max}^{1-\alpha} - N(k)^{1-\alpha}\right], \quad \textbf{[6]}$$

or, in terms of the q-deformed logarithmic function $\ln_q(x) \equiv (1-q)^{-1}[x^{1-q} - 1]$ with q a real number,

$$\ln_\alpha N(k) = \ln_\alpha N_{\max} - \mathcal{N}^{-1}k. \quad \textbf{[7]}$$

The size-rank distribution $N(k)$ is explicitly obtained from the above with use of the inverse of $\ln_q(x)$, the q-deformed exponential function $\exp_q(x) \equiv [1 + (1-q)x]^{1/(1-q)}$; this is

$$N(k) = N_{\max}\exp_\alpha\left(-N_{\max}^{\alpha-1}\mathcal{N}^{-1}k\right). \quad \textbf{[8]}$$

When $\alpha = 1$, Eq. **8** acquires the ordinary exponential form

$$N(k) = N_{\max}\exp\left(-\mathcal{N}^{-1}k\right), \quad \textbf{[9]}$$

whereas in the limit $N_{\max} \to \infty$ Eq. **8** becomes the power law $N(k) \sim k^{1/(1-\alpha)}$, which when $\alpha = 2$ gives the simple classical Zipf's law form $N(k) \sim k^{-1}$.

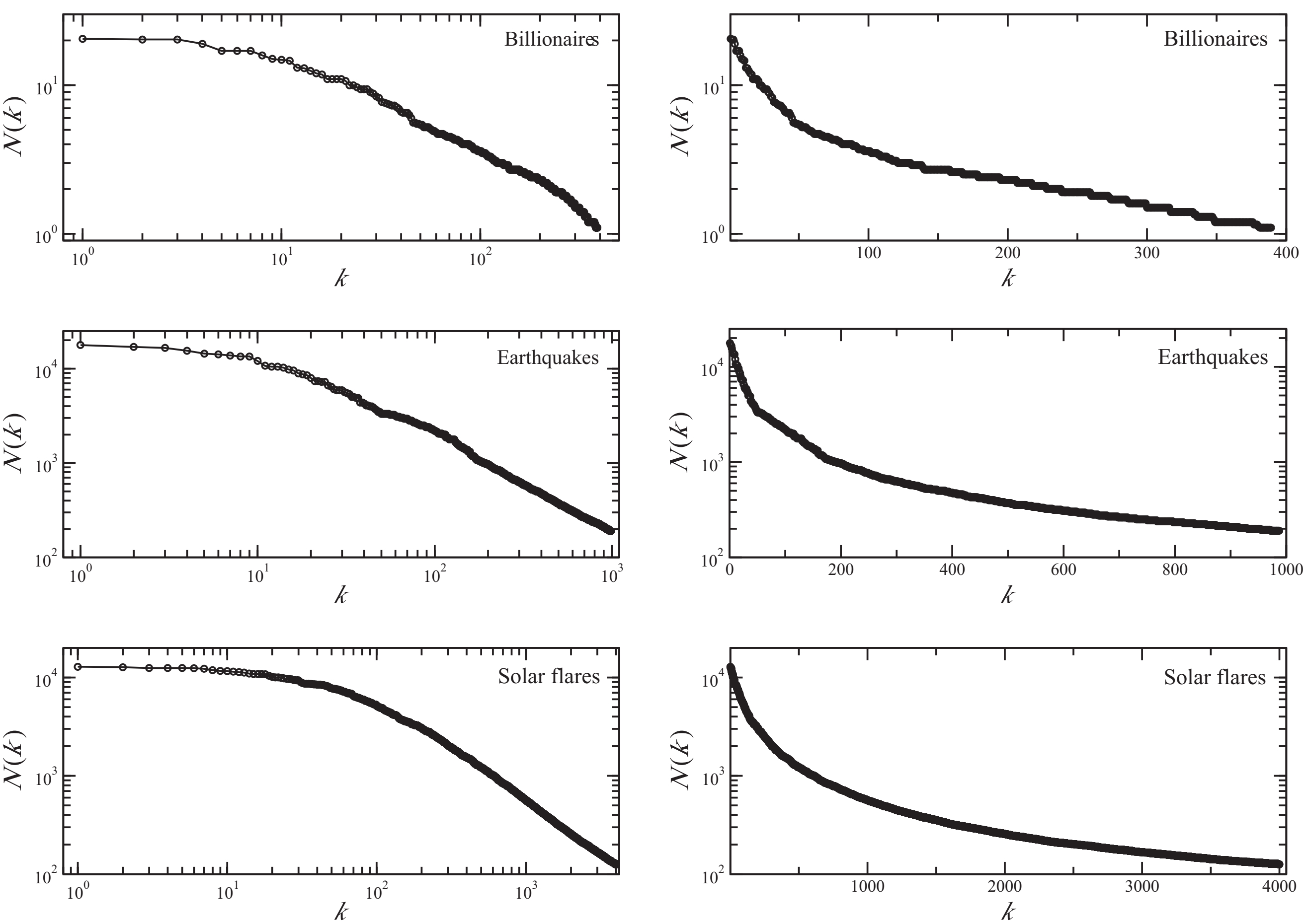

Fig. 1. Three examples of ranked data that appear to display power-law behavior along an interval of rank values. (*Top*) Data for the wealth of billionaires in the United States (7). (*Middle*) Data for the energy released by earthquakes in California (8). (*Bottom*) Data for the intensity of solar flares (9). (*Left*) Data are shown in logarithmic scales. (*Right*) Same data shown in log-linear scales. See text for description.

An explicit analogy between the generalized law of Zipf and the nonlinear dynamics of intermittency has been studied (4, 5). We recall the renormalization group fixed-point map for the tangent bifurcation. The trajectories x_t, $t = 1, 2, 3,\ldots$, produced by this map, comply (analytically) with

$$\ln_z x_t = \ln_z x_0 + ut \qquad [10]$$

or

$$x_t = x_0 \exp_z\left[x_0^{z-1} ut\right], \qquad [11]$$

where the x_0 are the initial positions. The parallels between Eqs. **10** and **11** with Eqs. **7** and **8**, respectively, are clear, and therefore we conclude that the dynamical system represented by the fixed-point map operates in accordance with the same q-generalized statistical–mechanical properties discussed below. We notice that the absence of an upper bound for the rank k in Eqs. **7** and **8** is equivalent to the tangency condition in the map. Accordingly, to describe data with finite maximum rank, we look at the changes in $N(k)$ brought about by shifting the corresponding map from tangency, i.e., we consider the trajectories x_t with initial positions x_0 of the map:

$$x' = x \exp_z\left(ux^{z-1}\right) + \varepsilon, 0 < \varepsilon \ll 1, \qquad [12]$$

with the identifications $k = t$, $\mathcal{N}^{-1} = -u$, $N(k) = x_t + x^*$, $N_{\max} = x_0 + x^*$, and $\alpha = z$, where the translation x^* ensures that all $N(k) \geq 0$. The capability of this approach to reproduce quantitatively real data for ranked data with deviations from power law for large rank has been discussed (4, 5).

Rank Distributions from Maximum Entropy Principle

The rank distribution $N(k)$ described in the previous section can be obtained from an MEP, and, as we shall see, this allows one to put forward important interpretations regarding the nature of the systems that give rise to it. However, first we adjust our interpretation of $N(k)$. This quantity is actually the size or magnitude of the data under consideration, the number of units that, in a microcanonical ensemble description, is the number of configurations that take place for a fixed value of k. Therefore, its inverse, $p_k = 1/N(k)$, is the (uniform) probability for the occurrence of each unit that constitutes $N(k)$. The probability p_k is normalized for fixed k, and we denote its limiting values by $p_{\min} = 1/N_{\max}$ and $p_{\max} = 1/N_{\min}$, $N_{\min} \leq N(k) \leq N_{\max}$.

A formal investigation of the possible entropy expressions that generalize the Boltzmann–Gibbs or Shannon canonical form has been systematically carried out with the use of the MEP under the assumption that only three of the Shannon–Kinchin axioms hold (11–13). (Inclusion of the fourth, composability, uniquely defines the canonical form.) Here we focus only on the Tsallis expressions (14).

Consider the entropy functional $\Phi_1[p_k]$ with Lagrange multipliers a and b,

$$\Phi_1[p_k] = S_1[p_k] + a\left[\sum_{k=0}^{k_{\max}} p_k - \mathcal{P}\right] + b\left[\sum_{k=0}^{k_{\max}} kp_k - \mathcal{K}\right], \qquad [13]$$

where the entropy expression $S_1[p_k]$ has the trace form (11)

$$S_1[p_k] = \sum_{k=0}^{k_{\max}} s_1(p_k). \qquad [14]$$

Optimization via $\partial\Phi_1[p_k]/\partial p_k = 0$, $k = 0, 1, 2,\ldots,k_{\max}$, gives

$$s_1'(p_k) = -a - bk. \qquad [15]$$

Now, the choices

$$s_1'(p_k) = \alpha \ln_\alpha p_k^{-1} - 1, \quad a = -\alpha \ln_\alpha p_{\min}^{-1} + 1, \quad b = \alpha\mathcal{N}^{-1}, \qquad [16]$$

lead to

$$\ln_\alpha p_k^{-1} = \ln_\alpha p_{\min}^{-1} - \mathcal{N}^{-1}k \qquad [17]$$

or

$$p_k^{-1} = p_{\min}^{-1} \exp_\alpha\left(-p_{\min}^{1-\alpha}\mathcal{N}^{-1}k\right), \qquad [18]$$

from which we immediately recover Eqs. **7** and **8**.

We repeat the same optimization procedure but with a constraint change (11). Consider the functional $\Phi_2[p_k]$ with Lagrange multipliers c and d,

$$\Phi_2[p_k] = S_2[p_k] + c\left[\sum_{k=0}^{k_{\max}} p_k - \mathcal{P}\right] + d\left[\sum_{k=0}^{k_{\max}} kp_k^{\alpha'} - \mathcal{K}_{\alpha'}\right], \qquad [19]$$

and where the entropy expression $S_2[p_k]$ has also a trace form

$$S_2[p_k] = \sum_{k=0}^{k_{\max}} s_2(p_k). \qquad [20]$$

Optimization via $\partial\Phi_2[p_k]/\partial p_k = 0$, $k = 0, 1, 2,\ldots,k_{\max}$ gives

$$s_2'(p_k) = -c - dk. \qquad [21]$$

This time, the choices

$$s_2'(p_k) = -(2-\alpha')\ln_{\alpha'} p_k - 1, c = (2-\alpha')\ln_{\alpha'} p_{\min} + 1, \qquad [22]$$

$$d = (2-\alpha')\mathcal{N}^{-1}, \qquad [23]$$

give the expressions

$$\ln_{\alpha'} p_k = \ln_{\alpha'} p_{\min} + \mathcal{N}^{-1}k \qquad [24]$$

or

$$p_k = p_{\min} \exp_{\alpha'}\left(p_{\min}^{\alpha'-1}\mathcal{N}^{-1}k\right). \qquad [25]$$

A comparison of Eqs. **7** and **8** with Eqs. **24** and **25**, respectively, indicates that they become equivalent with the identifications $p_k = 1/N(k)$, $p_{\min} = 1/N_{\max}$, $\alpha' = 2 - \alpha$. Furthermore, $s_2'(p_k) = s_2'(p_k)$ (as given by Eqs. **15**, **16**, **21**, and **23**), and therefore

$$S_2[p_k] = S_1[p_k], \qquad [26]$$

where their optimized expressions are

$$S_1[p_k] = \sum_{k=0}^{k_{\max}} p_k \ln_\alpha p_k^{-1}, \qquad [27]$$

and

$$S_2[p_k] = -\sum_{k=0}^{k_{\max}} p_k \ln_{\alpha'} p_k. \qquad [28]$$

Under the assumption of validity of only the first three Shannon–Kinchin axioms, it has been shown (11, 12) that there are only two ways to construct entropy expressions via the MEP procedure. These correspond to the constraints used in Eqs. **13** and **19** and the resulting entropy expressions are those in Eqs. **27** and **28**. The two approaches are related via the deformation index duality $\alpha' = 2 - \alpha$, and, for the same distribution p_k, their values are equal as in Eq. **26**. For an earlier account of this duality property, see ref. 15. See also ref. 16. From our earlier discussion we know that the index α fixes the shape of the rank distribution $N(k)$; its departure from unity generates its power-law feature and the value $\alpha = 2$ reproduces the classic Zipf law. To complete the picture we need to clarify the role of the dual index α' and the distribution p_k, and from this obtain an understanding of the dual entropy expressions in Eqs. **27** and **28**. Interestingly, when $\alpha = \alpha' = 1$ the duality collapses into the Boltzmann–Gibbs or Shannon entropy expressions and the exponential form for $N(k)$, but for $\alpha = 2$ we have $\alpha' = 0$ and p_k grows linearly with k.

In Fig. 2 we show the same three sets of data in Fig. 1 in log-linear scales. This time we fit them with Eqs. **8** and **25** and observe that the data are well described with values of the deformations $\alpha = 2$ and $\alpha' = 0$.

Statistical Mechanics of Contracted Configuration Space

The function $N(k)$ has the properties of a microcanonical partition function (4, 5). That is, the size $N(k)$ is the result of $N(k)$ equally probable configurations, and the probabilities p_k are correspondingly normalized for fixed k. However, these probabilities are not normalized if the rank k runs across its values $k = 0, \ldots, k_{\max}$, and we do not make an attempt here to do so. Instead, we look at the rank dependence in Eq. **25**, that we identify as the system's size dependence. As it can be observed in Fig. 2 (*Right*), the probabilities p_k rise sharply and then saturate as k increases. The pure deformed exponential

$$\frac{p_k}{p_{\min}} = \frac{N_{\max}}{N(k)} = \exp_{\alpha'}\left(p_{\min}^{\alpha'-1}\mathcal{N}^{-1}k\right) \qquad [29]$$

measures the change in the number of microcanonical configurations with the size of the system k. We define the size-dependent entropy

$$S(k) \equiv \ln_{\alpha'}\left(\frac{N_{\max}}{N_k}\right),\ k \text{ fixed}, \qquad [30]$$

and from Eqs. **29** and **30** we observe that $S(k)$ is extensive; doubling the numbers of billionaires, earthquakes, or solar flares in the data sets doubles the value of $S(k)$, and it can be seen to be so because the deformation index α' has the precise value to ensure this property. The constraint

$$\sum_{k=0}^{k_{\max}} k\, p_k^{\alpha'} = \mathcal{K}_{\alpha'} \qquad [31]$$

in Eq. **19** for entropy maximization indicates that the phase space $N_{\min} \le N \le N_{\max}$ is highly constrained because the probabilities

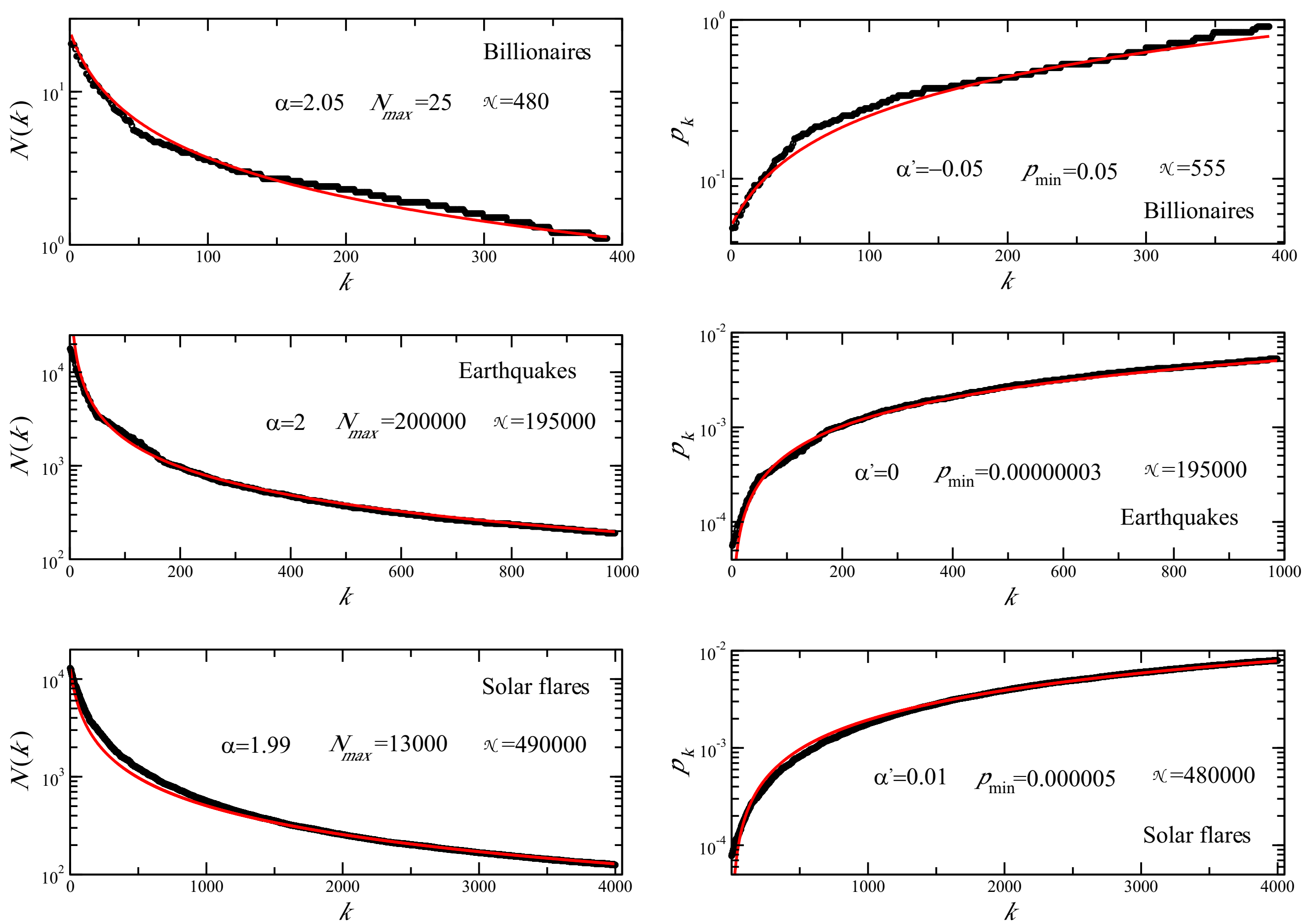

Fig. 2. Same three examples in Fig. 1 are fitted with the expressions in Eqs. **8** and **25**. As can be seen, the values of α needed for fitting are close to $\alpha \simeq 2$ and $\alpha' = 2 - \alpha \simeq 0$. The value $\alpha = 2$ gives the classical Zipf law exponent, whereas the value $\alpha' = 0$ indicates extreme configuration-space contraction. See text for description.

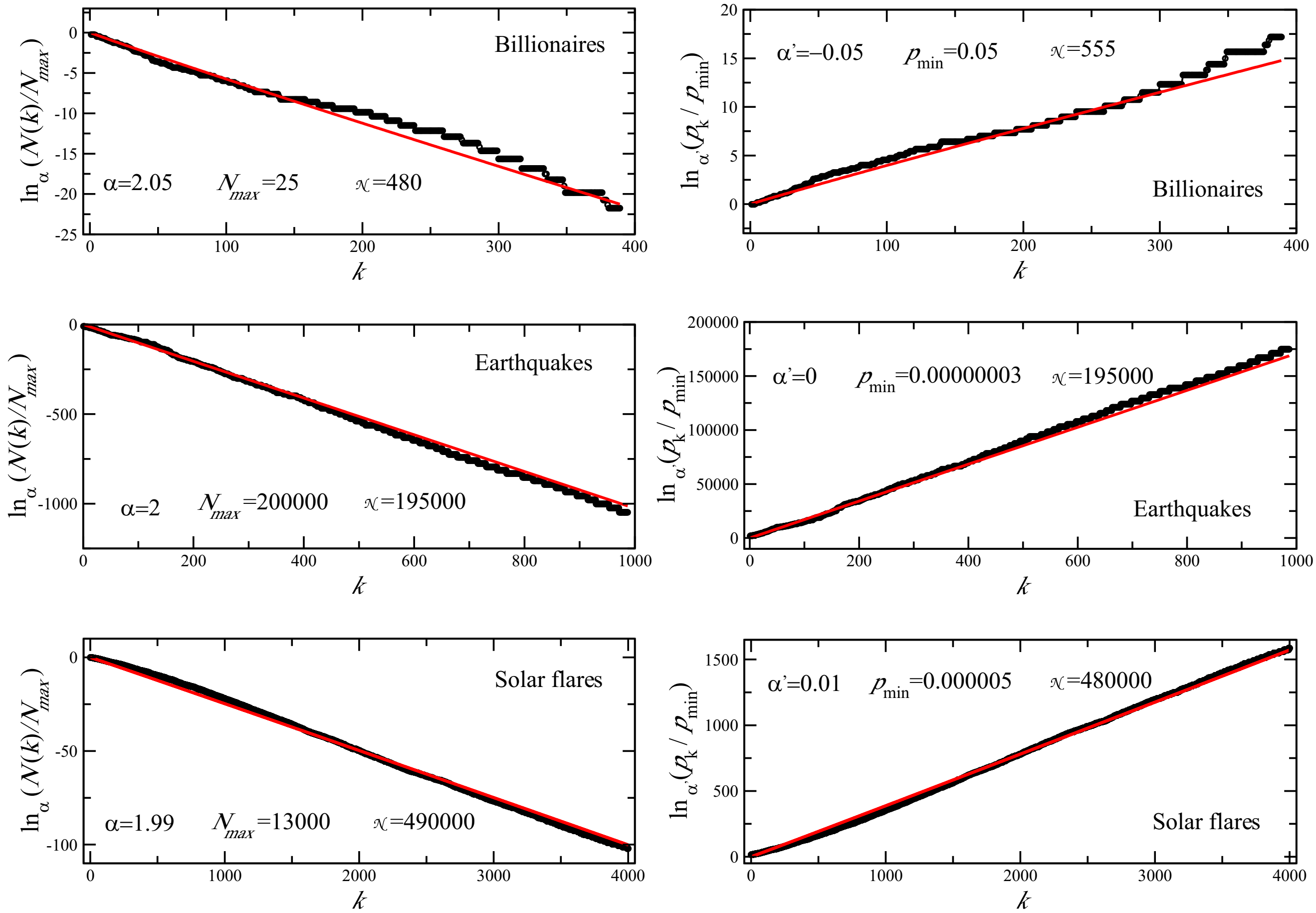

Fig. 3. Same three examples in Figs. 1 and 2 plotted in $\ln_\alpha(N(k)/N_{max})$ (*Left*) and $\ln_{\alpha'}(p_k/p_{min})$ (*Right*) scales. Data plotted in these scales are designed to display linear behavior if the theoretical expressions in Eqs. **8** and **25** are fulfilled by the data.

$p_k < 1$ need to be enhanced up to $p_k^{\alpha'}$, $\alpha' < 1$, to obtain a meaningful average of k. In relation to this, notice that $N(k)$ is a monotonously decreasing function with a power-law feature. This phase-space contraction is extreme for the case of Zipf law because α' reaches its minimum value of zero. For a system with normal occupation of phase space, the number of configurations grows exponentially and $S(k)$ above becomes extensive in k for index value $\alpha' = 1$, whereas for the phase space in the most contracted stage the number of configurations grows only linearly; this linearity is preserved in $S(k)$ when $\alpha' = 0$.

In Fig. 3 we show the same data as in Figs. 1 and 2 but this time plotted in deformed logarithmic scales with deformation indexes $\alpha \simeq 2$ and $\alpha' \simeq 0$. Data in these scales are displayed linearly and should be fitted by the theoretical expressions Eqs. **8** and **25** if these equations represent the behavior of the data.

Discussion

We have shown that size-rank distributions with power-law decay for moderate and large values of rank obey Tsallis statistics. The small-rank behavior that departs from the power law is also well reproduced by the deformed exponential expression in Eq. **8** for $N(k)$. For the specific data we presented (US billionaires, California earthquakes, and solar-flare intensities) the values of the exponential deformations were found to be $\alpha \simeq 2$, the value needed to obtain the classical Zipf law. To advance further in the characterization of the apparent relationship between rank distributions and generalized statistical mechanics, such as that of Tsallis, we rederived Eq. **8** for $N(k)$ from a maximum entropy procedure. This was done in accordance with the consideration of validity of only the first three Shannon–Kinchin axioms (11, 12). Under these conditions duality of entropy expressions appears according to the use of two different constraints. In doing this we introduce the (unnormalized) distribution $p_k = 1/N(k)$; actually $N(k)$ is the number of data for the same rank k (playing the role of a partition function) (4, 5). We obtain equality of the entropy expressions $S_1[p_k] = S_2[p_k]$ in Eqs. **27** and **28** and a companion rank distribution expression for p_k, Eq. **25**. As it is known (11, 12) the two entropies $S_1[p_k]$ and $S_2[p_k]$ correspond to the dual deformation indexes α and $\alpha' = 2 - \alpha$. We have inquired as to the different roles of the two entropy expressions and identify the physically relevant information carried by each one. We found that the value of the index α fixes the distribution's power-law exponent for $N(k)$ and that the dual index $\alpha' = 2 - \alpha$ ensures the extensivity of the deformed entropy. Finally, we argued that the value $\alpha = 2$, which corresponds to the classical Zipf law, manifests as $\alpha' = 0$, which we interpret as an extreme contraction of the phase space from which the data originate.

ACKNOWLEDGMENTS. G.C.Y. and A.R. gratefully acknowledge the hospitality of the Santa Fe Institute. Support by Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México-IN100311 and Consejo Nacional de Ciencia y Technología CB-2011-167978 is acknowledged. G.C.Y. was supported by the Scientific Research Projects Coordination Unit of Istanbul University, Project 36529. M.G.-M. acknowledges the generous support of Insight Venture Partners and the Bryan J. and June B. Zwan Foundation.

1. van der Galien SJG (2003) http://en.wikipedia.org/wiki/Zipfs_law.
2. Schroeder M (1991) *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* (Freeman, New York).
3. Gell-Mann M (1994) *The Quark and the Jaguar: Adventures in the Simple and the Complex* (Freeman, New York).
4. Altamirano C, Robledo A (2011) Possible thermodynamic structure underlying the laws of Zipf and Benford. *Eur Phys. J B* 81(3):345–351.
5. Robledo A (2011) Laws of Zipf and Benford, intermittency, and critical fluctuations. *Chin Sci Bull* 56(34):3645–3648.
6. Tsallis C (2009) *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World* (Springer, New York).
7. www.forbes.com/billionaires. Accessed May 29, 2014.
8. Southern California Earthquake Data Center. Available at www.data.scec.org. Accessed May 29, 2014.
9. http://tuvalu.santafe.edu/~aaronc/powerlaws/data.htm. Accessed May 29, 2014.
10. Pietronero L, Tosatti E, Tosatti V, Vespignani A (2001) Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. *Physica A* 293(1-2):297–304.
11. Hanel R, Thurner S, Gell-Mann M (2011) Generalized entropies and the transformation group of superstatistics. *Proc Natl Acad Sci USA* 108(16):6390–6394.
12. Hanel R, Thurner S, Gell-Mann M (2012) Generalized entropies and logarithms and their duality relations. *Proc Natl Acad Sci USA* 109(47):19151–19154.
13. Hanel R, Thurner S, Gell-Mann M (2014) How multiplicity determines entropy and the derivation of the maximum entropy principle for complex systems. *Proc Natl Acad Sci USA* 111(19):6905–6910.
14. Abe S (2000) Axioms and uniqueness theorem for Tsallis entropy. *Phys Lett A* 271(1-2):74–79.
15. Baldovin F, Robledo A (2004) Nonextensive Pesin identity: Exact renormalization group analytical results for the dynamics at the edge of chaos of the logistic map. *Phys Rev E Stat Nonlin Soft Matter Phys* 69(4 Pt 2):045202.
16. Robledo A (2013) Generalized statistical mechanics at the onset of chaos. *Entropy* 15(12):5178–5222.

PHYSICS