Laws of Zipf and Benford, intermittency, and critical fluctuations

ROBLEDO Alberto¹,²

¹ Instituto de Física, Universidad Nacional Autónoma de México, México 01000 D.F., Mexico; ² Grupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas, Universidad Carlos III de Madrid, 28911 Leganés, Madrid, Spain

Received May 20, 2011; accepted August 2, 2011

We describe precise equivalences between theoretical descriptions of: (i) size-rank and first-digit laws for numerical data sets, (ii) intermittency at the transition to chaos in nonlinear maps, and (iii) cluster fluctuations at criticality. The equivalences stem from a common statistical-mechanical structure that departs from the usual via a one-parameter deformation of the exponential and logarithmic functions. The generalized structure arises when configurational phase space is incompletely visited such that the accessible fraction has fractal properties. Thermodynamically, the common focal expression is an (incomplete) Legendre transform between two entropy (or Massieu) potentials. The theory is in quantitative agreement with real size-rank data and it naturally includes the bends or tails observed for small and large rank.

**Zipf’s law, Benford’s law, tangent bifurcation, critical clusters, generalized statistical mechanics**

The empirical laws of Zipf and Benford enjoy a unique place in the field of study of complex systems due to their combined omnipresence and simplicity. Zipf’s law refers to the (approximate) power law that is displayed by sets of data when these are given a ranking in relation to magnitude or rate of recurrence [1]. Benford’s law is a simple logarithmic rule for the frequency of first digits found also in numerical data sets [2]. An explanation of their apparent universality has been offered recently [3,4] by exhibiting an underlying statistical-mechanical structure obeyed by the quantities employed in describing the empirical laws. Additionally, the same theoretical structure was found to hold in a very different problem, that of the simplest of so-called routes out of chaos in low-dimensional mappings, that via intermittency. At this transition from chaos to regular behavior, sequences of numerical values, called trajectories, obtained by consecutive iterations of a nonlinear map, transform from irregular to periodic in a particular manner that consists of approximately regular patterns, laminar episodes, separated by bursts of disorder [5]. Notably, there is a precise equivalence between this transition to chaos, known technically as the tangent bifurcation, and the rank-size and first-digit laws [4]. Furthermore, earlier work [6,7] has established the implications of yet another strict analogy between problems in different fields, the same tangent bifurcation in nonlinear dynamics and the nature of the transitory clusters of order parameter (e.g. magnetization) that occur at ordinary critical points in thermal systems. Here we put together these remarkable and exact analogies to emphasize the fact that the parallelisms amongst these apparently very dissimilar problems stem from a common kind of statistical mechanics that arises when configurational phase space is incompletely visited in a strict way. Specifically, the restriction is that the accessible fraction of this space has fractal properties.

Thus, in the next section 1 we reproduce the theoretical expressions [3,4] for the size-rank and first-digit laws relevant to our purposes and describe the generalized statistical-mechanical structure we observe in them. In the following section 2 we present the parallelism between the ranking of data and the dynamics at the tangent bifurcation in nonlinear maps and describe the finite-size effect of the former
in terms of the off-tangency feature of the latter [4]. In section 3 we recall the theoretical properties of large clusters of order parameter at critical points in thermal systems and highlight their association to the properties of both the trajectories at the tangent bifurcation and the size-rank functions [6,7]. We conclude in section 4 with a short summary and discussion.

1 The laws of Benford and Zipf from an statistical-mechanical viewpoint

The relationship between the laws of Benford and Zipf was first analyzed in [8]. Benford’s law, \( p(n) = \log(1+n^{-1}) \), where \( p(n) \) is the relative probability of occurrence of digit \( n \), is obtained by integration between consecutive digits \( n \) and \( n+1 \) of \( P(N) \sim N^{-1} \), the underlying probability distribution for the data \( N \) under consideration. The first step in [8] consisted of the consideration of the more general case \( P(N) \sim N^{-\alpha} \), \( \alpha > 1 \). The next step was to obtain the rank \( k \) from \( P(N) \), this time as an integration over \( P(N) \) from \( N(k) \), the number of data that define the rank \( k \), to a finite number \( N_{\text{max}} \) that corresponds to the first value of the rank \( k \). In the limit \( N_{\text{max}} \to \infty \) one obtains \( N(k) \sim k^{1/(1-\alpha)} \) that is Zipf’s law with exponent \( 1/(\alpha-1) \) when \( \alpha > 1 \). For many sets of real data \( \alpha \approx 2 \) and the standard Zipf law is \( \alpha = 2 \) [8].

In our work [3,4], the first, minor, step was to keep \( N_{\text{max}} \) finite, but this led us to a basic conclusion on a possible physical origin of the laws of Zipf and Benford. That is, these laws represent general thermodynamic relations that belong to a special type of thermodynamic structure obtained from the usual via a scalar deformation parameter represented by the power \( \alpha \). A further generalization of the already generalized form of Benford’s law in [8], not specialized to a first or other digit, but to the numbers already generalized form of Benford’s law in [8], not specifically a Zipf’s law [8].

As a convenient shorthand notation consider the \( q \)-deformed logarithmic function \( \log_q(x) = (1-q)^{-1}(x^{1-q} - 1) \) with \( q \neq 1 \) a real number, and its inverse, the \( q \)-deformed exponential function \( \exp_q(x) = 1 + (1-q)x^{1/(1-q)} \) that reduce, respectively, to the ordinary logarithmic and exponential functions when \( q = 1 \). In terms of these functions, eq. (2) and its inverse can be written more concisely as

\[
\log_q N(k) = \log_q N_{\text{max}} - N^{-1} k, \tag{3}
\]

and

\[
N(k) = N_{\text{max}} \exp_q \left[ -N_{\text{max}}^{-1} N^{-1} k \right]. \tag{4}
\]

As shown in [3,4], eq. (4) is a generalization of Zipf’s law that is capable of reproducing quantitatively the behavior for small rank \( k \) observed in real data where, as one would expect, \( N_{\text{max}} \) is finite.

In Figure 1 we compare the numbers of occurrences of English words in a corpus with \( N(k) \) as given by eq. (4),

\[
p(n) = \int N^{-\alpha} dN = \frac{1}{1-\alpha} \left[ (n+1)^{-\alpha} - n^{-\alpha} \right], \quad \alpha \neq 1. \tag{1}
\]

This is a generalization of Benford’s law [8]. The rank \( k \) for a set of \( N \) numbers extracted from the basic distribution \( P(N) \sim N^{-\alpha} \) is given by [8]

\[
k = N_{\text{max}}^{N_{\text{max}}^{-1}} N^{-\alpha} dN = \frac{N}{1-\alpha} \left[ N_{\text{max}}^{-1} - N(k)^{-1} \right], \quad \alpha \neq 1, \tag{2}
\]

where \( N_{\text{max}} \) and \( N(k) \) correspond, respectively, to rank \( k = 0 \) and nonspecific rank \( k > 0 \). Eq. (2) introduces a continuum space variable for the rank \( k \) and therefore the first value of the rank is \( k = 0 \). Inversion of the above in the limit \( N_{\text{max}} \to \infty \) yields Zipf’s law \( N(k) \sim k^{1/(1-\alpha)} \) [8].

Figure 1. Rank-order statistics for the occurrence of words (empty circles) in the British National Corpus [9]. Eq. (4) with \( \alpha = 0.09 \) (smooth curve) is fitted to the data. The straight line in the inset is drawn for visualization purposes.
where the reproduction of the small-rank bend displayed by the data before the power-law behavior sets in is evident. In the theoretical expression this regime persists up to infinite rank $k \to \infty$. Alternatively, we recover from eq. (4) the power-law $N(k) \sim k^{\lambda(1-\alpha)}$ in the limit $N_{\text{max}} > 1$, $\alpha > 1$.

A key physical interpretation of eq. (3) was put forward in [3,4]. This is based on the fact that both $\log N_{\text{max}}$ and $\log N(k)$ are given by the integrals

$$\log N_{\text{max}} = \sum_{i=1}^{N_{\text{max}}} N^{-\alpha} \, dN \quad \text{and} \quad \log N(k) = \int_{i=1}^{N(k)} N^{-\alpha} \, dN,$$

and these in turn can be seen, when $\alpha=1$, to conform to the evaluation of entropy $S = \log N_{\text{max}}$ or $S_k = \log N(k)$ where the probability of $N$ equally-probable configurations in phase space is $P(N) = N^{-1}$. The same interpretation was extended to the case $\alpha > 1$ by considering that $P(N) = N^{-\alpha}$ represents the probability of $N$ equally-probable configurations in a strongly restricted phase space. The entropies are written now as

$$S_\alpha = \log N_{\text{max}} \quad \text{and} \quad S_k = \log N(k),$$

where $N_{\text{max}}$ and $N(k)$ play the roles of total configurational numbers or partition functions. Eq. (3) is rewritten as

$$S_\alpha = S_k - N^{\alpha-1} k,$$

where $N_{\text{max}}$ and $N(k)$ play the roles of total configurational numbers or partition functions. Eq. (3) is rewritten as

$$S_\alpha = S_k - N^{\alpha-1} k,$$

and read as the expression of an incomplete Legendre transform from the Massieu potential $S_\alpha(N^{\alpha})$, a function of the inverse of the number $N^{\alpha}$ to the entropy $S_k(k)$, a function of the rank $k$. The variables $N^{\alpha-1}$ and $k$ are seen to be conjugate in the thermodynamic sense. Eq. (4), being the inverse of eq. (3), states the same relationship but in terms of the ‘partition functions’ $N(k)$ and $N_{\text{max}}(N^{\alpha})$. Formally, the Legendre transformation of $S_k(N^{\alpha})$ into $S_\alpha(k)$ is completed by the elimination of the variable $N^{\alpha-1}$ in favor of $k$. This would require optimization of $S_\alpha$, i.e. via the use of an ‘equation of state’

$$k = -\frac{d}{dN^{\alpha-1}} \log N_{\text{max}}(N^{\alpha-1}).$$

2 The laws of Benford and Zipf from the perspective of nonlinear dynamics

To make explicit the analogy between the generalized laws of Benford and Zipf and the nonlinear dynamics of intermittency [4], it is necessary to recall briefly the Renormalization Group (RG) treatment of the tangent bifurcation that mediates the transition between chaotic and periodic attractors [5]. The common procedure to study the transition to chaos from a trajectory of period $n$ starts with the $n$-th composition $f^n(x)$ of a one-dimensional map $f(x)$ at such bifurcation, followed by an expansion for the neighborhood of one of the $n$ points tangent to the line with unit slope [5]. With complete generality one obtains

$$x' = f(x) = x + ux', \quad x > 0, \ z > 1,$$

where $x' = \text{sign}(x)|x'|$. The RG fixed-point map is the solution $f(x)$ of

$$f^\mu \left( f(x) \right) = \lambda^{-1} f^\mu \left( \lambda x \right),$$

together with a specific value for $\lambda$ that upon expansion around $x=0$ reproduces eq. (9). An exact analytical expression for $f(x)$ was obtained in [10] with the use of the assumed translation property of an auxiliary variable, $y = x^{1-z}$. This property is written as

$$x^{1-z} = 1^{-1} z + (1-z) u,$$

or, equivalently, as

$$x' = x \exp \left( m x^{1-z} \right).$$

It is straightforward to corroborate that $x' = f(x)$ as given by eq. (12) satisfies eq. (10) with $\lambda = 2^{(1-z)}$. Repeated iteration of eq. (11) leads to

$$x_{\nu}^{1-z} = 1_{\nu}^{1-z} + (1-z) u,$$

or

$$\log x_{\nu} = \log x_0 + ut.$$

So that the iteration number or time $t$ dependence of all trajectories is given by

$$x_t = x_0 \exp \left[ \frac{t x_{\nu}^{1-z}}{1-u} \right],$$

where $x_0$ is the initial position. The $q$-deformed properties of the tangent bifurcation are discussed at greater length in [11]. The parallel between eqs. (14) and (15) with eqs. (3) and (4), respectively, is clear, and therefore we conclude that the dynamical system represented by the fixed-point map $f(x)$ operates in accordance to the same statistical-mechanical property described in the previous section for the generalized laws.

We notice that the absence of an upper bound for the rank $k$ in eqs. (3) and (4) is equivalent to the tangency condition in the map. Accordingly, we look at the changes in $N(k)$ brought about by shifting the corresponding map from tangency (Figure 2), i.e. we consider the trajectories $x_t$ with initial positions $x_0$ of the map

$$x' = x \exp \left( m x^{1-z} \right) + e, \quad 0 < e << 1,$$

with the identifications $k = t$, $N^{\alpha-1} = -\mu$, $N(k) = x_t + x'$, $N_{\text{max}} = x_0 + x'$ and $\alpha = z$, where the translation $x'$ ensures that all
In Figure 3 we illustrate the capability of this approach to reproduce quantitatively real data for ranking of eigenfactors (a measure of the overall value) of physics journals (http://www.eigenfactor.org/index.php).

In the intermittency route out of chaos it is relevant to determine the duration of the so-called laminar episodes [5], i.e. the average time spent by the trajectories going through the “bottleneck” formed in the region where the map is closest to the line of unit slope. Naturally, it is this property of the nonlinear dynamics that translates into the finite-size properties of the occurrence-rank function \( N(k) \). One more important result that follows from the analogy between nonlinear dynamics and the rank law is that the most common value for the degree of nonlinearity at tangency is \( z = 2 \), obtained when the map is analytic at \( x = 0 \) with nonzero second derivative, and this implies \( \alpha = 2 \), close to the values observed for most sets of real data.

### 3 The laws of Benford and Zipf in relation to critical phenomena

The local fluctuations of a thermal system undergoing a second-order phase transition are represented in a classical spin model by the deviations of the magnetization \( \phi \) from its equilibrium vanishing average. These fluctuations generate transient magnetic domains, or critical clusters, on all size scales whose combined behavior produce the familiar properties of critical phenomena. Critical clusters have been studied (see [6] and references therein) with the use of the statistical-mechanical method of a coarse-grained free energy, like the Landau-Ginzburg-Wilson (LGW) continuous spin model Hamiltonian, for the critical temperature and zero external field. The cluster of radius \( R \) has a fractal configuration whose amplitude \( \phi \) grows in time and eventually collapses when the instability is reached. This process is described by a nonlinear map with tangency and feedback features, such that the time evolution of the cluster is given in the nonlinear system as a laminar episode of intermittent dynamics (see [6] and references therein). The coarse-grained phase space for a critical cluster is the set of possible values \( 0 \leq \phi \leq 1 \) for the magnetization, and in a one-dimensional system (with no loss of generality as the range of interactions can be adjusted to ensure criticality) the cluster is described by the profile \( \phi(x) \) for each position \( x \). Also, it is sufficient to study a symmetrical cluster, \( \phi(-x) = \phi(x), 0 < x < R, \) of size \( 2R \).

The basic results can be more readily obtained via the following phenomenological arguments [7]. Firstly, assume that the fractal properties of the cluster imply an under-occupation of the coarse-grained phase space. This sub-occupation makes the ordinary entropy

\[
S_1(R) = \log \phi_R, \quad (17)
\]

where \( \phi_R = \phi(R) \), to be nonextensive. That is, \( S_1(R) \) does not grow linearly with \( R \) because the total magnetization

\[
\Phi(R) = 2 \int_0^R \phi(x), \quad (18)
\]

depends on \( R \) as

\[
\Phi(R) = R^{d_f}, \quad (19)
\]

where \( d_f < 1 \) is the fractal dimension of the cluster [6,7]. To restore the extensivity of the entropy we replace eq. (17) by

\[
S_q(R) = \log_q \phi_R, \quad (20)
\]

where \( \log_q \) is, as above, the \( q \)-logarithmic function, and tune \( q \) to that specific value \( q > 1 \) that makes \( S_q(R) \) grow linearly with \( R \). This is equivalent to the replacement \( \phi_R \Rightarrow \phi_R^q, \quad q > 1 \) in the derivative of \( S_q(R) \) as
Similarly, we write the entropy expression

$$S_q(0) = \log_q \phi_0,$$

(22)

associated with the value of the magnetization at the center of the cluster, \(\phi_0 = \phi(0)\), and use the same value of \(q\) as in eq. (20). Again, we have adopted the replacement \(\phi_0 \rightarrow \phi_0^q\), \(q > 1\).

Next we relate \(S_q(R)\) and \(\hat{S}_q(0)\) formally via the Legendre transform expression

$$S_q(R) = \hat{S}_q(0) + \beta R,$$

(23)

where the parameter \(\beta\) is given by

$$\beta = \frac{d\hat{S}_q}{dR}.$$

(24)

The entropy \(\hat{S}_q(0)\) can be thought of as a ‘free energy’ \(-\beta F(\beta)\) where \(\beta\) is the ‘inverse temperature’ conjugate to the ‘energy’ \(R\). Use of eqs. (20) and (22) into eq. (23) gives

$$\log_q \phi_R = \log_q \phi_0 + \beta R,$$

(25)

or

$$\phi_R = \phi_0 \exp_q [\phi_0^{q-1} \beta R].$$

(26)

Eq. (26) relates the values of the magnetization at the center and at the edge of the critical cluster. This is the main consequence we obtain from the above simple assumptions. The expression can be generalized to describe the magnetization profile \(\phi(x)\) of the entire cluster. Furthermore, eq. (26) can be derived from a LGW Hamiltonian within the saddle-point approximation. This more fundamental approach allows us to determine the parameters \(q\) and \(\beta\) in terms of Hamiltonian variables. In particular, the index \(q\) is given by \(q = (\delta + 1)/2\) where \(\delta\) is the isothermal critical exponent [6].

Comparison of eqs. (25) and (26) with eqs. (14) and (15), respectively, and with eqs. (3) and (4), respectively, demonstrates the equivalence between the properties of size-rank functions, dynamics at the intermittency transition out of chaos, and the properties of critical clusters.

Consequently, these three apparently different problems share the same statistical-mechanical interpretation. It should be noted however that the cluster magnetization \(\phi_R\) above is defined only for positive values of \(R\) and therefore this function coincides with the trajectory \(x_i\) at or close to the tangent bifurcation with its positive portion \(x_i > 0\), i.e. a trajectory out of tangency or past the midpoint of the “bottleneck”.

Similarly, \(\phi_R\) translates only to the portion of the size-rank function \(N(k)\) that corresponds to the finite-size induced down-bend that follows for large rank after the approximate power-law regime.

### 4 Summary and discussion

We have described the equivalence of the theoretical treatments for three topics or issues encountered in different fields: (i) the empirical laws of Zipf and Benford for diverse sets of numerical data; (ii) the dynamics associated to the intermittency route to chaos in nonlinear maps; and (iii) the nature of large critical clusters in many-body systems. These solid analogies are based on having in common the main expressions for their description and that these in turn share the same source for their derivation. We interpreted these expressions as the manifestation of a thermodynamic, or statistical-mechanical, structure that arises from a restriction in these systems to sample a large amount of configurations. More specifically, the theoretical expressions for the first-dig and size-rank laws, eqs. (1) and (2), were derived in [8] under the basic assumption that the data sets obeyed by them are statistically well reproduced when extracted from a power law distribution \(P(N) \sim N^{-\alpha}\). That is, the deviation from unity of the exponent \(\alpha\) implies a restricted access to the phase space for data configurations which when enumerated produce the numbers \(N\). The restriction involves an accessible subset of this space with a scale invariant property, i.e., a fractal set, as indicated by the power law \(N^{\alpha}\).

Our statistical-mechanical viewpoint becomes evident when \(P(N)\) is seen to represent the probability distribution of \(N\) equally-probable configurations in the phase space for the data, and, consequently, suggests the definition of the generalized entropies in eq. (6). The straightforward transformation of the pair of eqs. (3) and (4) into either eqs. (14) and (15) or eqs. (25) and (26) reveals the fact that the statistical-mechanical interpretation applies also to the critical nonlinear dynamics at the tangent bifurcation and to the growth and collapse of the critical clusters at second order transitions. The degree of deformation, \(\alpha, \gamma, \text{ or } q\), in the theory is an important index that characterizes the phase-space restriction. For instance, when \(\alpha = \gamma = q = 2\) one has: (i) the standard value for the power-law decay associated to Zipf’s law, minus one; (ii) the most general situation for a map at tangency, nonzero curvature; and (iii) the classical value for the critical isothermal exponent, \(\delta = 3\). More details and more ample discussions of the results commented here are found in [4,6,7,11].
1 Zipf G K. Human Behavior and the Principle of Least-Effort. Cambridge: Addison-Wesley, 1949
2 Benford F. The law of anomalous numbers. Proc Am Phil Soc, 1938, 78: 551–572
3 Altamirano C, Robledo A. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (LNICST). Berlin: Springer-Verlag, 2009. 2232–2237
4 Altamirano C, Robledo A. Possible thermodynamic structure underlying the laws of Zipf and Benford. Eur Phys J B, 2011, 81: 345–351
5 Schuster H G. Deterministic Chaos. An Introduction. 2nd ed. Weinheim: Wiley-VCH, 1988
6 Robledo A. Unorthodox properties of critical clusters. Mol Phys, 2005, 103: 3025–3030
7 Robledo A. q-statistical properties of large critical clusters. Int J Mod Phys B, 2007, 21: 3947–3953
8 Pietronero L, Tosatti E, Tosatti V, et al. The uneven distribution of numbers in nature. Physica A, 2001, 293: 297–304
9 Leech G, Rayson P, Wilson A. Word Frequencies in Written and Spoken English based on the British National Corpus. London: Longman, 2001
10 Hu B, Rudnick J. Exact solutions to the Feigenbaum renormalization-group equations for intermittency. Phys Rev Lett, 1982, 48: 1645–1648
11 Baldovin F, Robledo A. Sensitivity to initial conditions at bifurcations in one-dimensional nonlinear maps: Rigorous nonextensive solutions. Europhys Lett, 2002, 60: 518–524

Open Access This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.