Eur. Phys. J. B **13**, 777–780 (2000)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences


# Are citations of scientific papers a case of nonextensivity?

C. Tsallis[a] and M.P. de Albuquerque

Centro Brasileiro de Pesquisas Fisicas, Rua Xavier Sigaud 150, 22290-180 Rio de Janeiro-RJ, Brazil

Received 13 April 1999

**Abstract.** The distribution $N(x)$ of citations of scientific papers has recently been illustrated (on ISI and PRE data sets) and analyzed by Redner (Eur. Phys. J. B **4**, 131 (1998)). To fit the data, a stretched exponential ($N(x) \propto \exp-(x/x_0)^\beta$) has been used with only partial success. The success is not complete because the data exhibit, for large citation count $x$, a power law (roughly $N(x) \propto x^{-3}$ for the ISI data), which, clearly, the stretched exponential does not reproduce. This fact is then attributed to a possibly different nature of rarely cited and largely cited papers. We show here that, within a nonextensive thermostatistical formalism, the same data can be quite satisfactorily fitted with a *single* curve (namely, $N(x) \propto 1/[1+(q-1)\,\lambda\,x]^{\frac{q}{q-1}}$ for the available values of $x$. This is consistent with the connection recently established by Denisov (Phys. Lett. A **235**, 447 (1997)) between this nonextensive formalism and the Zipf-Mandelbrot law. What the present analysis ultimately suggests is that, in contrast to Redner's conclusion, the phenomenon might essentially be *one and the same* along the *entire* range of the citation number $x$.

**PACS.** 02.50.-r Probability theory, stochastic processes, and statistics – 01.75.+m Science and society – 89.90.+n Other topics of general interest to physicists (restricted to new topics in section 89)

Half a century ago, Zipf [1] made his remarkable observations about some basic linguistic laws. More precisely, if we order the words appearing in a text (*e.g.*, Homer's Iliad) from the most to the less frequent ones, thus obtaining a ranking (low rank for the most used, and high rank for the less used), we can plot, as a function of the rank, the number of times those words appear. Zipf showed that, excepting the words with extremely low rank, an inverse power law emerges (so called *Zipf's law*). The exponent exhibits interesting universal aspects. For instance, for the spoken language, it appears to be very sensitive to the degree of instruction (primary, intermediate, highly academic) of the speaker, but very little to the particular culture (French, German, Anglo-saxon). Later on, Mandelbrot pointed out connections of this phenomenon with fractals [2], and also suggested a further correction, namely that substantially better fittings can be obtained by using an inverse power law of the sum of the rank with a constant (so called *Zipf-Mandelbrot law*). A further step along this line was provided recently by Denisov [3]. Indeed, using within the Sinai-Bowen-Ruelle thermodynamical formalism for symbolic dynamics, the nonextensive thermostatistics proposed some years ago by one of us [4], Denisov *deduced* the Zipf-Mandelbrot law. To be more precise, it is clear that, unless one uses a specific model, there is no way to deduce the *precise values* for the exponent and the additive constant. What Denisov deduced, from very generic entropic arguments, was the *form* of the law. In this sense, the approach is very analogous to those which succeed associating Gaussians to normal diffusion, and Levy or Student's $t$-distributions to anomalous superdiffusion (see, for instance, [5,6] respectively). Finally, it is important to stress here that, although the present problematic was historically triggered in Linguistics, the same kind of considerations are equally relevant to DNA sequences, artificial languages, and a variety of other stochastic, deterministic or mixed processes.

Here we focus on an interesting analysis of data concerning citations of scientific publications. More precisely, Redner [7] recently exhibited and discussed the distributions of citations related to two quite large data sets, namely (i) 6 716 198 citations of 783 339 papers, published in 1981 and cited between 1981 and June 1997, that have been catalogued by the Institute for Scientific Information (ISI), and (ii) 351 872 citations, as of June 1997, of 24 296 papers cited at least once and which were published in Physical Review D (PRD) in volumes 11 through 50 (1975-1994). In his study, Redner addressed the citations of publications, in variance with Laherrere and Sornette [8], who addressed, in a similar study, the citations of authors. If we denote by $x$ the number of citations and by $N(x)$ the number of papers that are cited $x$ times. The main results of the study were that, for relatively large values of $x$, $N(x) \propto 1/x^\alpha$ with $\alpha \simeq 3$, whereas, for relatively small values of $x$, the data were reasonably well fitted with a stretched exponential, *i.e.*, $N(x) \propto \exp[-(x/x_0)^\beta]$, $\beta$ and $x_0$ being the fitting parameters ($\beta \simeq 0.44$ and 0.39 for the ISI and the PRD data respectively); see Figures 1a and 1b.

[a] e-mail: tsallis@cbpf.br

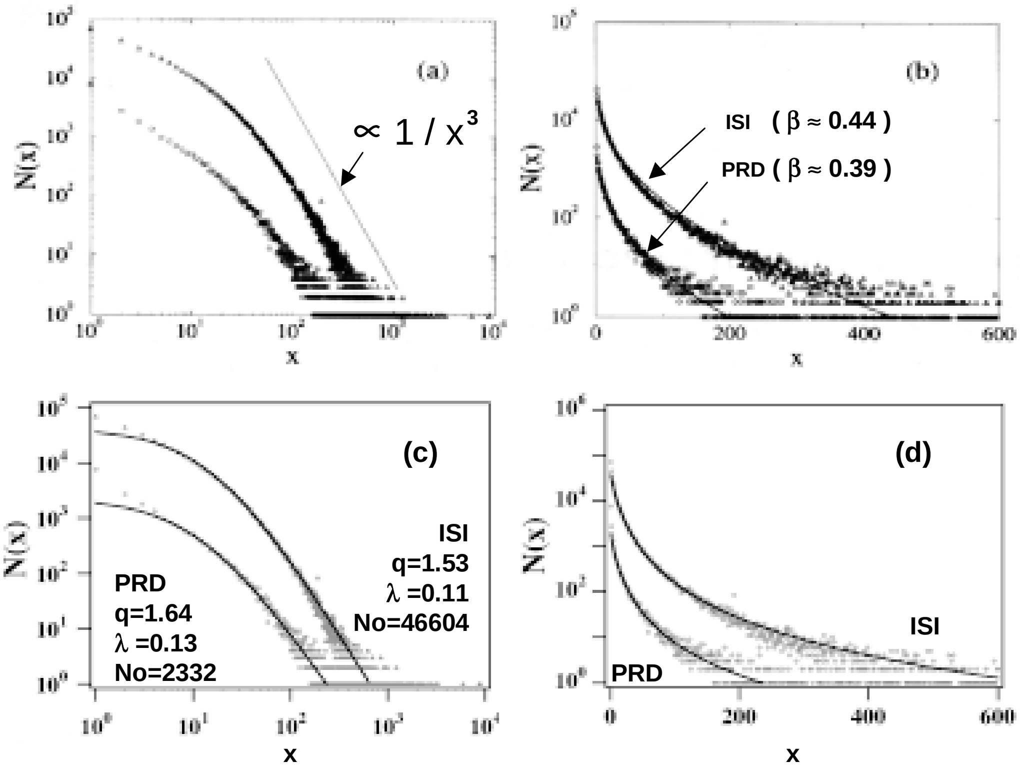


**Fig. 1.** ISI and PRD distributions of citations (experimental data and fittings). From [7]: (a) log-linear plot and (b) log-log plot. Present work: (c) log-linear plot and (d) log-log plot (Eq. (9) has been used with the values for $(q, \lambda, N_0)$ indicated in the figure).

Since a stretched exponential by no means asymptotically provides an inverse power law, the author concluded that large $x$ and low $x$ phenomena are of *different nature* (in the author's words, "These results provide evidence that the citation distribution is not described by a single function over the entire range of citation count"). While the phenomenon exhibited by Redner is of great interest, we must disagree with his conclusion. It is the central purpose of our present effort to develop arguments within the nonextensive statistical mechanics mentioned above [4], and along the lines of Denisov, which will lead to a *single* function $N(x)$ having, like the stretched exponential, only two fitting parameters. This function is of the power-law type and will turn out to fit both ISI and PRD experimental data sensibly better than the forms described above.

Before presenting our arguments for this specific problem, let us briefly introduce the nonextensive formalism we are referring to. If the physical system we are focusing on involves long-range interactions or long-range microscopic memory or (multi)fractal boundary conditions, it can exhibit a quite anomalous thermodynamic behavior, which might even be untractable within Boltzmann-Gibbs (BG) statistical mechanics. To overcome at least some of these pathological situations, an entropic form $S_q$ has been proposed [4] which yields a generalization of standard statistical mechanics and thermodynamics. This entropy is defined as follows:

$$S_q \equiv k \frac{1-\sum_i p_i^q}{q-1} \qquad \left(\sum_i p_i = 1;\; q \in \mathcal{R}\right) \tag{1}$$

where $k$ is a positive constant (from now on taken to be unity, without loss of generality). In the limit $q \to 1$, we recover the usual BG entropy, *i.e.*, $S_1 = -\sum_i p_i \ln p_i$. A property which characterizes the above generalized entropic form is the following: if we have two independent systems $A$ and $B$ such that $p_{ij}^{A+B} = p_i^A\, p_j^B$, then

$$S_q(A+B) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B). \tag{2}$$

Consequently, if $q > 1, = 1$ or $< 1$, $S_q$ is *subextensive, extensive or superextensive*. Optimization of this entropy with appropriate constraints provides equilibrium distributions which are of the power-law type, and which recover the exponential Boltzmann factor only in the limit $q \to 1$. This thermostatistics has provided interesting insights onto a variety of physical systems such as two-dimensional turbulence in pure-electron plasma [9], self-gravitating systems [10], cosmology [11], solar neutrinos [12], Levy [5] and correlated [6] anomalous diffusions, inverse bremsstrahlung absorption in plasma [13], quantum scattering [14], one-dimensional maps [15], a variety of self-organized critical models [16], long-range interaction conservative systems [17], processing of EEG signals of epileptic humans and turtles [18], among others (see [19] for a review). To theoretically study such complex systems within this nonextensive formalism, some approaches (besides, naturally, the usual analytic and numerical methods) are now available such as the generalizations of (i) Kubo's linear response theory, (ii) Feynman's perturbation theory as well as the Bogoliubov's inequality (basis of the variational method), (iii) Green's functions, and (iv) Feynman's path integral (respectively generalized, in the realm of nonextensivity, in [20–23]).

Let us now focuse on our specific problem, namely the distributions of citations. The corresponding entropic form is given by

$$S_q = \frac{1-\sum_{x=1}^{\infty} p_x^q}{q-1} \,. \tag{3}$$

The optimization of this entropy with the corresponding constraints [24], namely

$$\sum_{x=1}^{\infty} p_x = 1 \tag{4}$$

and

$$\langle x \rangle_q \;\equiv \frac{\sum_{x=1}^{\infty} x\, p_x^q}{\sum_{x=1}^{\infty} p_x^q} = \text{const.}, \tag{5}$$

yields

$$p_q(x) = \frac{[1-(1-q)\lambda x]^{\frac{1}{1-q}}}{\sum_{y=1}^{\infty}[1-(1-q)\lambda y]^{\frac{1}{1-q}}} \tag{6}$$

where, unless $q = 1$, $\lambda$ is *not* (see [24]) the Lagrange parameter associated with constraint (5), but can nevertheless be determined through that constraint. This distribution is expected to be an excellent approximation for $x$ not too small (say, not below 5), but departures would not be surprising while approaching unity. Indeed, we have deduced equation (6) through generic entropic considerations and not by using a specific model. Also, for precisely the same reason, $q$ and $\lambda$ are to be considered as free parameters within the present phenomenological approach.

Equation (6) implies that the so-called *escort* distribution is given by

$$P_q(x) \equiv \frac{[p_q(x)]^q}{\sum_{y=1}^{\infty}[p_q(y)]^q} = \frac{[1-(1-q)\lambda x]^{\frac{q}{1-q}}}{\sum_{y=1}^{\infty}[1-(1-q)\lambda y]^{\frac{q}{1-q}}}$$

$$= \frac{\left\{\sum_{y=1}^{\infty}\frac{1}{[1+(q-1)\lambda y]^{\frac{q}{q-1}}}\right\}^{-1}}{[1+(q-1)\lambda x]^{\frac{q}{q-1}}} \,. \tag{7}$$

This escort distribution is to be identified (see [24]) with the above introduced experimental distribution $N(x)$, hence

$$N(x) = N(1)\frac{[1+(q-1)\lambda]^{\frac{q}{q-1}}}{[1+(q-1)\lambda x]^{\frac{q}{q-1}}} \tag{8}$$

or, equivalently,

$$N(x) = \frac{N_0}{[1+(q-1)\lambda x]^{\frac{q}{q-1}}} \tag{9}$$

where we have simplified the notation by introducing $N_0$. The fittings of both ISI and PRD data series using this functional form are exhibited in Figures 1c and 1d. We can appreciate that they are considerably better (in both precision and completeness) than those appearing in [7]. In particular, we have obtained, for the ISI series, $q \simeq 1.53$, hence $q/(q-1) \simeq 2.89$, which is clearly compatible with the approximate exponent 3 advanced in [7].

As a summarizing conclusion, we suggest that, in variance with what is stated in [7], the present interesting linguistic-like phenomenon revealed by Redner appears to emerge from one and the same reason for practically the *entire* range of citation score $x$. Furthermore, this reason appears to be deeply related to thermostatistical nonextensivity. However, by no means this conclusion can be considered as the complete analysis of the present empirical fact. Indeed, we still need to understand what peculiarity of the nonlinear dynamics (*i.e.*, what physical mechanism) of this phenomenon is responsible for the specific values of $q$ which fit the experimental data. As a supplementary bonus, we might be able, along this line, to understand why a stretched exponential form does *not* fit the entire experimental range when citations *per paper* are focused, whereas it appears to be satisfactory when citations *per scientist* are focused instead [8]. In other words, specific microscopic models are of course very welcome in order to achieve a more concrete insight, and also for addressing the exceedingly small values of $x$, which are out of the scope of the present phenomenological approach. Naturally, the predictive power of the theory would consequently be enriched.

CNPq, FAPERJ and PRONEX/FINEP (Brazilian Agencies) are acknowledged for partial support.

## References

1. G.K. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley, Cambridge-MA, 1949).
2. B.B. Mandelbrot, *The fractal geometry of nature* (Freeman, San Francisco, 1983).

3. S. Denisov, Phys. Lett. A **235**, 447 (1997).
4. C. Tsallis, J. Stat. Phys. **52**, 479 (1988); E.M.F. Curado, C. Tsallis, J. Phys. A **24**, L69 (1991) [Corrigenda: **24**, 3187 (1991); **25**, 539 (1992)]. A regularly updated bibliography is available at `http://tsallis.cat.cbpf.br/biblio.htm`.
5. D.H. Zanette, P.A. Alemany, Phys. Rev. Lett. **75**, 366 (1995); C. Tsallis, S.V.F. Levy, A.M.C. de Souza, R. Maynard, Phys. Rev. Lett. **77**, 5442 (1996); **77**, 5442(E) (1996); C. Tsallis, D. Prato, Phys. Rev. E **60**, 2398 (1999).
6. A.R. Plastino, A. Plastino, Physica A **222**, 347 (1995); C. Tsallis, D.J. Bukman, Phys. Rev E **54**, R2197 (1996); A.M.C. Souza, C. Tsallis, Physica A **236**, 52 (1997).
7. S. Redner, Eur. Phys. J. B **4**, 131 (1998).
8. J. Laherrere, D. Sornette, Eur. Phys. J. B **2**, 525 (1998).
9. B.M. Boghosian, Phys. Rev. E **53**, 4754 (1996).
10. A.R. Plastino, A. Plastino, Phys. Lett. A **174**, 384 (1993); **193**, 251 (1994).
11. V.H. Hamity, D.E. Barraco, Phys. Rev. Lett. **76**, 4664 (1996); D.F. Torres, H. Vucetich, A. Plastino, Phys. Rev. Lett. **79**, 1588 (1997); **80**, 3889(E) (1998).
12. G. Kaniadakis, A. Lavagno, P. Quarati, Phys. Lett. B **369**, 308 (1996); P. Quarati *et al.*, Nucl. Phys. A **621**, 345c (1997).
13. C. Tsallis, A.M.C. de Souza, Phys. Lett. A **235**, 444 (1997).
14. D.B. Ion, M.L.D. Ion, Phys. Rev. Lett. **81**, 5714 (1998); **83**, 463 (1999).
15. C. Tsallis, A.R. Plastino, W.-M. Zheng, Chaos, Solitons and Fractals **8**, 885 (1997); M.L. Lyra, C. Tsallis, Phys. Rev. Lett. **80**, 53 (1998); U. Tirnakli, C. Tsallis, M.L. Lyra, Eur. Phys. J. B **11**, 309 (1999).
16. F.A. Tamarit, S.A. Cannas, C. Tsallis, Eur. Phys. J. B **1**, 545 (1998); A.R.R. Papa, C. Tsallis, Phys. Rev. E **57**, 3923 (1998).
17. C. Anteneodo, C. Tsallis, Phys. Rev. Lett. **80**, 5313 (1998).
18. L.G. Gamero, A. Plastino, M.E. Torres, Physica A **246**, 487 (1997); A. Capurro, L. Diambra, D. Lorenzo, O. Macadar, M.T. Martin, C. Mostaccio, A. Plastino, J. Perez, E. Rofman, M.E. Torres, J. Velluti, Physica A **265**, 235 (1999).
19. C. Tsallis, in *Nonextensive Statistical Mechanics and Thermodynamics*, eds. S.R.A. Salinas, C. Tsallis, Braz. J. Phys. **29**, 1 (1999). (See `cond-mat/9903356` or `http://sbf.if.usp.br/WWW_pages/Journals/BJP/Vol29/Num1/index.htm`).
20. A.K. Rajagopal, Phys. Rev. Lett. **76**, 3469 (1996).
21. E.K. Lenzi, L.C. Malacarne, R.S. Mendes, Phys. Rev. Lett. **80**, 218 (1998).
22. A.K. Rajagopal, R.S. Mendes, E.K. Lenzi, Phys. Rev. Lett. **80**, 3907 (1998).
23. E.K. Lenzi, L.C. Malacarne, R.S. Mendes, Physica A (2000), in press.
24. C. Tsallis, R.S. Mendes, A.R Plastino, Physica A **261**, 534 (1998).