Scale-Free Download Network for Publications *

HAN Ding-Ding(韩定定)[1], LIU Jin-Gao(刘锦高)[1], MA Yu-Gang (马余刚)[2]**,
CAI Xiang-Zhou(蔡翔舟)[2], SHEN Wen-Qing(沈文庆)[2]

[1]*Department of Electronic Engineering, East China Normal University, Shanghai 200062*
[2]*Shanghai Institute of Applied Physics, Chinese Academy of Sciences, PO Box 800-204, Shanghai 201800*

(Received 27 April 2004)

The scale-free power-law behaviour of the statistics of the download frequency of publications has been reported, for the first time to our knowledge. The data of the download frequency of publications are taken from a well-constructed web page in the field of economic physics (http://www.unifr.ch/econophysics/). The Zipf-law analysis and the Tsallis entropy method were used to fit the download frequency. It was found that the power-law exponent of rank-ordered frequency distribution is $\gamma \sim 0.38 \pm 0.04$*, which is consistent with the power-law exponent* $\alpha \sim 3.37 \pm 0.45$ *for the cumulated frequency distributions. The preferential attachment model of Barabasi and Albert network has been used to explain the download network.*

PACS: 89. 20. Hh, 89. 75. Hc, 89. 75. Da

Recently the complex network has become a hot research field, especially for its feature of statistical mechanics. The rapid growth of the internet stimulates physicists to investigate the rules of networks. In a pioneering work of Barabasi and Albert, they found that the degree of node of Internet routes, URL (universal resource locator)-linked networks in the WWW (World-Wide Web) satisfies the power-law distribution[1,2] (also called scale-free networks).

The power-law behaviour of rank distribution is believed to be related to Zipf's law, which was found by Zipf in the early part of the last century.[3] Originally, Zipf made his remarkable observations about some basic linguistic laws. More precisely, if we order the words appearing in a text from the most to the least frequent ones, we can plot the number of times of those words appearing as a function of the rank. Zipf showed that, except for the words with extremely low rank, an inverse power law emerges (the so-called Zipf law). That is, the frequency x

$$x \simeq R^{-\gamma}, \tag{1}$$

where γ is a Zipf law exponent and R represents the rank. Zipf's law for scale free networks is different from the predictions of pure random networks introduced by in Refs. [4,5] For the former, Barabasi and Albert proposed a preferential attachment model (BA model) to give the scale-free law of the link of Internet[6] and Tsallis explained the statistical feature of the complex network using a non-extensive entropy (known as Tsallis' entropy[7]) approach.[8] The original BA model predicts the probability distributions $p(k) \simeq k^{-\alpha}$, where k is the degree of network node and $\alpha = 3$ (the corresponding rank-ordered law yielding $\gamma = 1/(\alpha - 1) = 1/2$.[9]) Extended and modified models based on the BA model have been developed in order to obtain $\alpha = 2 - 4$ more precisely to fit realistic systems.[10] Recently complex networks and/or Zipf's law have been explored in a broad range of sciences: physics, electronics, computing sciences, geology, sociology, economics, linguistics, biology and many others. For instance, complex networks have been observed for WWW and Internet, movie actor collaboration network, science collaboration graph, cellular networks, ecological networks, phone call networks, citation networks, networks in linguistics, power and neural networks, protein folding and interaction network, earthquake network, firm growth and bankruptcy and gene expression (for a review, see Refs. [6,11]), even for the fragment hierarchical distribution in nuclear dissociation[11] and hadronic production process,[12] etc.

The scale-free networks related to scientific publications have been also explored; it was shown that the citation network of scientific references[8,9] and the collaboration graph of the co-authorship of publications[13] satisfies the power law distribution for rank-order distribution. Redner exhibited and discussed the distributions of citations related to two quite large data sets, namely (i) 6716198 citations of 783339 papers, published in 1981 and cited between 1981 and June 1997, which have been catalogued by the Institute for Scientific Information (ISI), and (ii) 351872 citations, as of June 1997, of 24296 papers cited at least once and which were published in Physical Review D in volumes 11 through 50 (1975–1994). In his study, Redner addressed the citations of publications, in variance with Laherrere and Sornette,[14]

* Supported partially by the National Natural Science Foundation of China under Grant Nos 19725521 and 10328509, and the Major State Basic Research and Development Programme of China under Contract No G200077400.

** To whom correspondence should be addressed. Email: ygma@sinr.ac.cn

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 $\alpha \sim 3$, whereas, for relatively small values of x, the data were reasonably well fitted with a stretched exponential, i.e. $N(x) \propto \exp[-(x/x_0)^{\beta}]$, β and x_0 being the fitting parameters ($\beta \simeq 0.44$ and 0.39 for the ISI and the PRD data respectively).

To be helpful to expose these differences in the citation distribution, Redner constructed the Zipf plot,[3] in which the number of citations of the kth most-ranked paper out of an ensemble of M papers is plotted versus rank k. By this definition, the Zipf plot is closely related to the cumulative large-x tail of the citation distribution and hence it is well suitable for determining the large-x tail of the citation distribution. The integral nature of the Zipf plot also smoothes the fluctuations in the high-citation tail and thus facilitates quantitative analysis. For the above-mentioned data set, he found that the Zipf law exponent γ (see Eq. (1)) close to 1/2, which is consistent with the power law exponent $\alpha = 3$ ($\alpha = 1 + 1/\gamma$) for the distribution of citations.

In this work, we report that the rank-ordered download frequency of the papers in a web page can also be described by the Zipf law. The data set we are using here comes from a well constructed web page[15] in the field of economical physics (the so-called econophysics) by Zhang since 1998. The scale free download network is explored and the quantitative information about this complex network has been extracted by the Zipf law and Tsallis' non-extensive entropy. The preferential attachment network model of Barabasi and Albert is used to explain the mechanism of network.

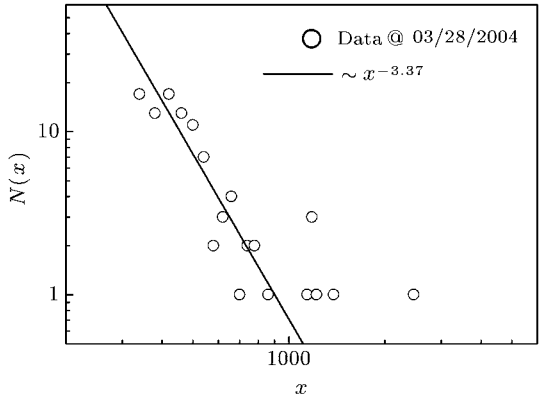

Fig. 1. Distribution of the 100 top download papers till 03/28/2004 (http://www.unifr.ch/econophysics). Here x represents the download frequency, and $N(x)$ represents the numbers of the papers which have been downloaded for x times.

In terms of the frequency of the download of a paper, the rank can be defined from the most downloaded paper ($R = 1$) to the less downloaded paper. The distribution of the downloaded frequency is shown in Fig. 1. Roughly speaking, the download distribution can be fitted by the power law distribution: $N(x) \sim x^{-\alpha}$. The last few points have a large fluctuation beyond the good fit below $x \sim 1000$. The extracted exponent $\alpha \sim 3.37 \pm 0.45$.

Since we have values of the rank-ordered frequency, we can make the Zipf plot for the download distribution. Figure 2(a) shows 12 Zipf plots for 12 selected dates that are represented by the different symbols, which are formatted by year–month–date. The time of these plots spans from 28 September 2003 to 28 March 2004. In order to minimize the fluctuation of these plots due to the statistics and network growth, we average 12 data set points and make the Zipf plot in Fig. 2(b). The Zipf power law (Eq. (1)) has been used to fit Fig. 2(b) and the extracted exponent $\gamma \sim 0.38 \pm 0.04$. This value leads to $\alpha = 1 + 1/\gamma = 3.63 \pm 0.28$, which is in a reasonable agreement with $\alpha = 3.37 \pm 0.45$ from the download distribution of Fig. 1. It is in the range of 2–4 for various realistic networks.[10]

On the other hand, Tsallis proposed a reasonable explanation[8] and well fitted real data sets by the non-extensive entropy theory.[7] In the non-extensive entropy theory, the probability distribution function is given by the expectation being constant, i.e.

$$p(x_k) \sim \frac{1}{[1 + (q-1)\lambda(x_k - \langle x_k \rangle)]^{\frac{q}{q-1}}}, \qquad (2)$$

where $\langle x_k \rangle$ denotes the mathematical expectation of x_k; λ is the factor similar to Lagrange multipliers, and q is the characteristic parameter related to the exponent. When q approaches to 1, Tsallis entropy becomes the Boltzmann–Gibbs entropy and $p(x_k)$ approaches an exponential distribution function.

In the rank-ordered statistics, the rank of x_k, $R(x_k)$, and the cumulative distribution function are equivalent to a simple relation, it reads[9]

$$R(x_k) \propto \int_{x_k}^{\infty} p(x)dx = 1 - \int_{0}^{x_k} p(x)dx. \qquad (3)$$

By integrating $p(x)$ in Eq. (2) one can obtain the following result:

$$R \propto \frac{1}{\lambda}[1 + (q-1)\lambda(x - \langle x \rangle)]^{-\frac{1}{q-1}}; \qquad (4)$$

or representing x as a function of $Rank$ yields

$$x = x_0 + \frac{b}{(R - R_{FS})^{q-1}}, \qquad (5)$$

where x_0 and b are the fitted parameters, and parameter R_{FS} is introduced here to take the finite size effect

into account.

By using Eq. (4) we fit the data of average download frequency (Fig. 2(b)) and extract the parameter q and $Rank_{FS}$. The dotted line represents this fit with the parameter $q = 1.351 \pm 0.006$ and $R_{FS} = 0.60$. From q we deduce $\frac{q}{q-1} = 3.846 \pm 0.068$. This exponent is very close to the exponent of $\alpha = 3.37 \pm 0.45$ in Fig. 1 as well as $\alpha = 3.63 \pm 0.28$ deduced from the γ value of the Zipf law fit to the rank-ordered distribution.

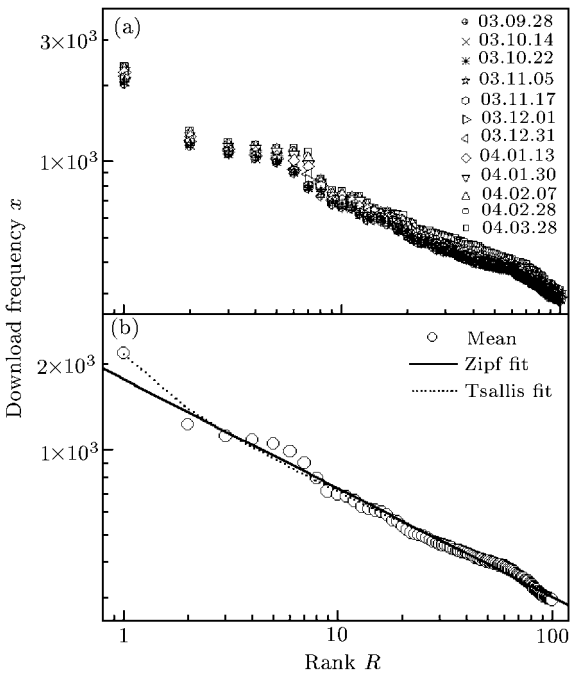

Fig. 2. Rank-ordered (Zipf-type) plot for the download frequency of http://www.unifr.ch/econophysics web page with the numbers representing year–month–date.

The scale-free characteristic of statistics of the download frequency could be interpreted by the BA model.[6] In the linear BA model, the growth of the network can be constructed by two steps. Firstly, the earlier growth process: starting with a small number of vertices (download papers) whose visitors are interested in them from a huge reference base of the econophysics web page, at every time step some visitors add some new vertices and a rank web page of the vertices was initially constructed. Secondly, the preferential attachment process: each visitor of the econophysics web page can freely access the rank web page of the papers. The higher the rank of the downloaded papers, the more probability a visitor would like to download, and the more frequency this leads to in statistics. In this mechanism of the BA model, it is natural that such a kind of preferential attachment process will result in a power-law or Zipf law distribution of the downloaded frequency.

In conclusion, the scale-free power-law behaviour has been observed in the download frequency distribution of the papers in an econophysics web page. From the download frequency distribution, it can be described by the power-law with the exponent $\alpha = 3.37 \pm 0.45$, which is consistent with the description of the Zipf law for the rank-ordered download frequency with a scale-free power law exponent $\gamma \sim 0.38 \pm 0.04$. This Zipf law parameter is not far from the exponent from the rank-ordered citation distribution.[9] It may be indicative of a similar mechanism for both networks, which can be explained by the preferential attachment process of the BA model. On the other hand, the download frequency is also considered in the framework of the non-extensive Tsallis entropy theory, which gives us the non-extensive Tsallis entropy index $q = 1.351 \pm 0.006$ and leads to $\frac{q}{q-1} = 3.846 \pm 0.068$, which is also in good agreement with the above α parameter.

References

[1] Barabasi AL and Albert R 1999 *Science* **286** 509
[2] Albert R, Jeong H and Barabasi A L 1999 *Nature* **401** 130
[3] Zipf G K 1949 *Human Behavior and the Principle of Least Effort* (Cambridge: Addison-Wesley)
[4] Marsili M and Zhang Y C 1998 *Phys. Rev. Lett.* **80** 2741
[5] Erdos P and Renyi A 1960 *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17
[6] Albert R and Barabasi A L 2002 *Rev. Mod. Phys.* **74** 47
[7] Tsallis C 1988 *J. Stat. Phys.* **52** 479
[8] Tsallis C and Albuquerque M P 2000 *Eur. Phys. J.* B **13** 777
[9] Redner S 1998 *Eur. Phys. J.* B **4** 131
[10] Dorogovtsev S N and Mendes J F F 2002 *Adv. Phys.* **51** 1079
[11] Ma Y G 1999 *Phys. Rev. Lett.* **83** 3617
Ma Y G 2000 *Chin. Phys. Lett.* **17** 340
[12] Wilk G and Wlodarczyk Z 2004 *Preprint* arXiv:hep-ph/0403244
[13] Newman N E J 2001 *Proc. Natl. Acad. Sci.* U.S.A. **98** 404
[14] Laherrere J and Sornette D 1998 *Eur. Phys. J.* B **2** 525
[15] http://www.unifr.ch/econophysics/