ELSEVIER

Volume 387, issue 7 1 March 2008 ISSN 0378-4371

PHYSICA A

STATISTICAL MECHANICS AND ITS APPLICATIONS

Recognized by the European Physical Society

EPS

Available online at
ScienceDirect
www.sciencedirect.com

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

Available online at www.sciencedirect.com

ScienceDirect

Physica A 387 (2008) 1628–1636

PHYSICA A

www.elsevier.com/locate/physa

q-exponential fitting for distributions of family names

Hiroaki S. Yamada[a,*], Kazumoto Iguchi[b]

[a] *Yamada Physics Research Laboratory, 5-7-14 Aoyama, Niigata 950-2002, Japan*
[b] *KazumotoIguchi Research Laboratory, 70-3 Shinhari, Hari, Anan, Tokushima 774-0003, Japan*

Received 27 March 2007; received in revised form 23 September 2007
Available online 6 November 2007

Abstract

We study the applicability of the q-exponential function for the distribution of family names. We mainly focus on the rank-size distribution of Japanese family names. The result supports the fact that the q-exponential distribution is relevant to the distribution of family names that is understood until now to obey power-law distribution (Zipf law).

Keywords: q-exponential property; Family names; Zipf's law; Scale-free network

1. Introduction

Power-law distribution has been found in wide fields such as biology, sociology and economics and so on. Parete law for the distribution of income [1,2], Zipf's law for city-size and word frequency [3], and Omori's law for earthquakes [4,5] are typical classical examples. Such examples are too many to enumerate even in recent technological society as commonly seen in the distribution of documents on the Web site [6]. The reasons for the emergence of such power-law distributions have been studied for a long time. To explain the occurrence of Zipf's law, Simon proposed a mechanism based on the stochastic process [7–13], a long time ago. Two decades ago, self-organized criticality (SOC) has also been introduced by Bak et al. in order to explain the frequency-size or frequency-intensity distribution of the dissipative systems [14].

Recently, instead of random networks, the scale-free networks (i.e., complex network structures with a power-law distribution in the number of degrees at the site) have been found by Barabási and co-workers from studying the growth of the internet geometry and topology [15–21]. Here, the link distribution functions also obey a power law. More recently, a generalized thermostatistical formalism based on a power-law entropic measure (Tsallis entropy) has attracted much attention in various complex systems and critical phenomena [22–24]. The optimization of the Tsallis entropy gives a q-exponential function as the non-equilibrium distribution function [24]. In the natural progression, the q-exponential fits for earthquakes [25], market traded volume distribution [26,27], empirical degree distribution in feedback networks [33–35] and so on, have been successfully done. Note that the q-exponential function has power-law tail in an asymptotic limit, and is equivalent to the Zipf–Mandelbrot distribution in the cases with q larger than

[*] Corresponding author. Tel.: +81 25 267 1941; fax: +81 25 267 1941.
E-mail address: hyamada@uranus.dti.ne.jp (H.S. Yamada).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.11.002

unity. In this way, the power-law distribution can be ubiquitously seen and play a very important role in many branches of science.

From this context, in this paper we investigate the applicability of the q-exponential function for the rank-size distribution of family names. The statistical properties of the distributions of family names have been reported for several countries [28–31]. Miyazima et al. showed that the frequency distribution of family names obeys a power-law distribution by using the data for some Japanese local cities [28]. They found that in the rank-size distribution of the family names the power-law behavior changes from a power law $P(k) \sim k^{-\gamma}$ with an exponent $\gamma \approx 0.67$ in the low-frequency regime to another one with a different exponent $\gamma \approx 1.33$ in the high-frequency regime. Satou and Seno have also obtained almost the same result for distribution of family names in Japan and some other countries [32]. However, the origin of the crossover (or singular point) between both the different power laws is not clear.

For such power-law behavior of the distributions, explanations based on Galton–Watson branching processes and the Simon models have been performed [28–31]. These are microscopic approaches in the sense that the theory is based on the rate equation that the distribution functions of family names follow. Although there is the advantage that we can treat analytically the solution of the equation, its applicability is certainly limited. For example, many such theories are applied only to the asymptotic case, i.e. quasi-stationary state, and to the case for tails of distributions. Indeed, when we apply the fitting of power-law decay to the frequency distributions of family names of the Japanese and the Korean, the low-frequency region and the high-frequency region have different exponents to each other, but it is impossible to explain at the same time those exponents by the microscopic approach based on the rate equations described above.

We explore the q-exponential properties of the distribution of family names in the Japanese society. The result supports the fact that the q-exponential distribution is relevant to the distribution of family names over whole frequency regime. The estimated power exponent in the high-frequency regime is larger than the one observed by Miyazima et al. Furthermore, we show numerical results of the fit to data for family names in other countries.

The organization of the paper is the following. In Section 2, we present the simple explanation for the q-exponential distribution and the relation to the power-law distribution by using scale-free network. In Section 3, we apply the q-exponential fit to data set of family names. In the last section, we will give the summary and discussion.

2. q-exponential properties of distribution functions

In this section, we give a brief introduction to the q-exponential properties and the relation to the power-law distribution.

2.1. q-exponential properties

The q-exponential function has been introduced through nonextensive statistical mechanics which has been extremely successful for critical phenomena, complex systems, and nonergodic systems [22–24]. The distribution $P(k)$ can be derived by extremizing the q-entropy,

$$S_q[P(k)] = \frac{1 - \int_0^\infty \mathrm{d}k[P(k)]^q}{q-1}, \tag{1}$$

with the constraints as,

$$\int_0^\infty \mathrm{d}k P(k) = 1, \qquad \frac{\int_0^\infty \mathrm{d}k k[P(k)]^q}{\int_0^\infty \mathrm{d}k[P(k)]^q} = k_0. \tag{2}$$

Then we can obtain $P(k) \propto \exp_q\{-k/k_0\}$, where the q-exponential function is defined as,

$$\exp_q\{x\} \equiv [1 + (1-q)x]^{\frac{1}{1-q}}. \tag{3}$$

Moreover, it easily found that the q-exponential function is the solution of the nonlinear equation,

$$\frac{\mathrm{d}y(x)}{\mathrm{d}x} = y(x)^q. \tag{4}$$

According to White et al. [33], we adapt a fitting function as

$$P(k) = P_0 k^{\delta} \exp_q(-k/k_0), \tag{5}$$

where P_0 is the normalization constant. This ansatz as the fitting form has been successfully used for the distribution of the trading volumes and distribution of degrees in some complex networks. Note that in the limit $k \to \infty$, $P(k)$ approaches the Parete distribution or the Zipf's rank-size distribution such as $P(k) \sim P_0 a k^{-b}$, where

$$a = \left[\frac{k_0}{q-1}\right]^{\frac{1}{q-1}}, \qquad b = \frac{1}{q-1} - \delta. \tag{6}$$

Thus at least the function of the form (5) has the power-law behavior in the tail. The number of the adjusting parameters is three, i.e. the q-exponential parameter q, the characteristic value k_0, and the complement power index δ.

2.2. *Relation to scale-free property*

We shows the simple relationship between the q-exponential property and the scale-free distribution in the network growth. We consider the algebraic preferential-attachment (PA) probability:

$$\Pi_i = \frac{k_i^{\alpha}}{\sum_{j=1}^{N-1} k_j^{\alpha}}, \tag{7}$$

where k_i is the number of degree at the ith site and $-\infty < \alpha < \infty$. For a large network, Π_i can be regarded as a continuous rate of change of k_i. Then we obtain the following equation for the growth of degree k_i as,

$$\frac{\mathrm{d}k_i}{\mathrm{d}t} = \bar{m}\frac{k_i^{\alpha}}{\mu_\alpha t}, \tag{8}$$

where $t(=\sum_j k_j^0) = \sum_j 1$ denotes the evolution time that is equal to the number of nodes, and $\sum_j k_j = 2\bar{m}t$, $\sum_j k_j^{\alpha} = \mu_\alpha t$. It follows that Eq. (8) has a q-exponential type solution when compared to Eq. (4). We can obtain the solution under the condition $k_i(t_i) = \bar{m}$ as,

$$k_i(t) = \bar{m} \exp_\alpha\left[-\frac{\bar{m}^{\alpha}}{\mu_\alpha} \ln\frac{t_i}{t}\right] \tag{9}$$

$$= \bar{m}\left[1 - (1-\alpha)\frac{\bar{m}^{\alpha}}{\mu_\alpha} \ln\frac{t_i}{t}\right]^{\frac{1}{1-\alpha}}. \tag{10}$$

Since the degree distribution $P(k)$ is defined by $P(k) = -\frac{\mathrm{d}}{\mathrm{d}k}\left(\frac{t_i}{t}\right)$ [17], the degree distribution becomes

$$P(k) = \frac{\mu_\alpha}{\bar{m}}\frac{1}{k^{\alpha}} \exp\left[-\frac{\mu_\alpha}{\bar{m}}\left(\frac{k^{1-\alpha}-1}{1-\alpha} - \frac{\bar{m}^{1-\alpha}-1}{1-\alpha}\right)\right]. \tag{11}$$

This is the result Eq. (11) of Liu et al. [21].

It is shown that in the limit of $k \to \infty$ the asymptotic form of $P(k)$ becomes $P(k) \sim k^{-3}$ when $\alpha = 1$ and $P(k) \sim \exp(-k/m)$ when $\alpha = 0$.

We can regard a society consisting of individual persons with a family name as a bipartite network: family names and individual persons. Then the degree distribution of family names is the rank-size distribution. In the following section we try to apply the fitting based on the q-exponential function to the rank-size distributions of family names.

3. Numerical results

In this section, we show the numerical results of the q-exponential fitting for the data for some empirical size-rank distributions of family names. We used the Levenberg–Marquardt algorithm for nonlinear least-square fitting of the

Table 1
The SW diversity S and evenness J for distribution of family names in some countries

Country	N_t	K	K_{cut}	S	J
Japan	125	270 000	10 000	10.71	0.80
USA*	230	1500 000	88 799	9.96	0.60
Norway	4.2	4 000	3 200	10.16	0.87
Manx	0.075	449_+	449	7.30	0.82
Korea	50	280_+	280	4.81	0.58
$China$*	1300	619_+	619	5.62	0.61

The other columns denote the total population N_t, the number of family names K, and the cut-off K_{cut} which are used for the calculation of the SW entropy S and the evenness J in the distribution. The unit of the total population is millions. We cited some data of the marked countries by asterisk ($*$) from Ref. [32].

parameters of the q-exponential function from Eq. (5) [36]. The Levenberg–Marquardt algorithm is a standard method to optimize the cost function by using the effective combination of the steepest decent method and the inverse Hesse method. We obtained the data for fitting from some web sites [37].

3.1. *Some basic properties of distribution of Japanese family names*

Before performing the q-exponential fit to the data, we give some statistical properties of the distribution compared to the ones in other countries. We use the *Shannon–Wiener (SW) diversity* $S[n_i]$, i.e. the Shannon entropy in informatics, and the *evenness* $J[n_i]$ for the distribution. The evenness express uniformity of the size for each rank. They are commonly used to measure biodiversity of ecological data. We consider size n_i of the rank i. SW entropy and evenness are defined as,

$$S[n_i] = -\sum_{i}^{K} p_i \log_2 p_i, \tag{12}$$

$$J[n_i] = \frac{S}{\log_2 K}, \tag{13}$$

where $p_i \equiv n_i/N$, and N and K are the total number of persons and the number of family names used as data, respectively. The $\log_2 K$ denotes the *maximum capacity* of the information, i.e. entropy in a uniform distribution. It has been reported that K is proportional to the total population by using data from some cities in Japan [28]. Note that in regarding p_i as a continuous variable, p_i corresponds to $P(k)$ in Section 2.

In Table 1 we give the SW entropy and the evenness for some countries. Fig. 1 shows plots of entropy as a function of evenness. It is found that in the case of Japan high entropy and high evenness are a typical feature as compared with other countries. In the last section we give a comment on the rich diversity of the distribution of Japanese family names.

3.2. *q-exponential fit to the distribution of the Japanese family names*

Fig. 2 shows family name frequencies as a function of the family size (Zipf plots) for Japan in recent times. In Fig. 2(a) and (b), the number of individual persons and the number of households are used as a measure of frequencies, respectively. The curves in Fig. 2 are numerical q-exponential functions given by the nonlinear least-square fitting of the data. The estimated parameters $\{q, k_0, \delta\}$ are summarized in Table 2. As shown in Fig. 2(c) and (d), the data clearly show the power-law dependence with $\gamma \approx 1.5$ in the tail of the distribution. The result supports the fact that the q-exponential distribution is relevant to the distribution of family names over the whole frequency regime.

Let us change the region of the parameter fitting. In Fig. 3 we show Zipf plots and the estimated q-exponential curves of the data for the number of individual persons. The fitting ranges and the estimated parameters are summarized in Table 2. The value of the power-law index γ based on the estimated parameters gradually increases as we use the data in the range of tail in the distribution. The value is larger than the value $\gamma \approx 1.33$, which was estimated by Miyazima et al. [28] and Satou et al. [32].

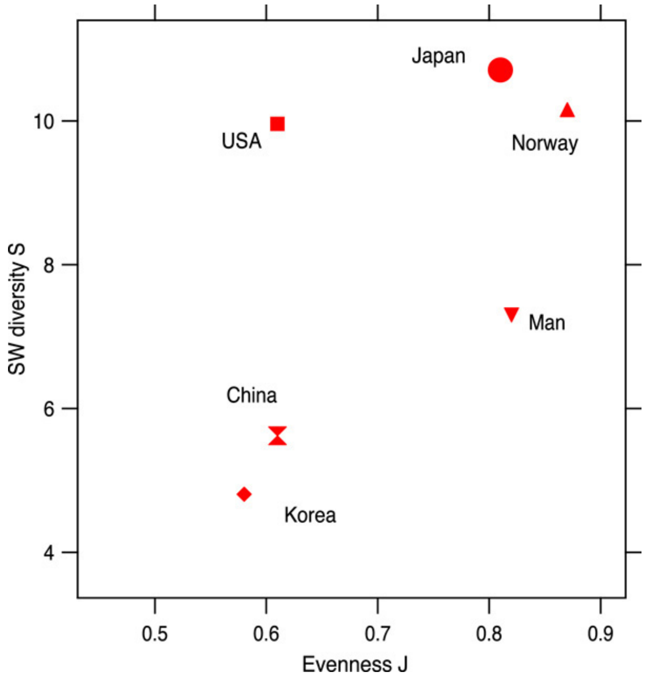

Fig. 1. (Color online) Plots of SW entropy S as a function of evenness J for some countries.

Table 2
Parameters for the best fit to the q-exponential functions for the rank-size distributions of the Japanese family names

Range	q	k_0	δ	b
1–10 000	2.07	562.7	−0.606	1.54
1–10 000	2.03	524.8	−0.594	1.56
1–1000	2.16	147.6	−0.414	1.28
1–7000	1.99	288.1	−0.471	1.49
2000–7000	1.68	338.1	−0.134	1.62
5000–10 000	1.66	393.0	−0.154	1.66

The first row shows the result for the household unit. The others are ones for individual units. The left column denotes the fitting range. The effective digit is two. And $b = \frac{1}{q-1} - \delta$.

Table 3
Parameters for the best fit to the q-exponential function for the rank-size distributions of family names in some countries shown in Fig. 4

Country	q	k_0	δ	b
USA	2.81	1251	−0.63	1.18
Norway	2.68	5.71	−0.25	0.84
Manx	1.63	28.3	−0.31	1.88
Korea	1.05	20.5	−0.34	2.34

3.3. *q-exponential fit to the distribution of the family names in other countries*

Generally, the distribution of the family names strongly depends on the country due to the effect of each culture and own history. We would like to check the applicability of the q-exponential fitting to the data for various countries.

Fig. 4 shows Zipf plots of family names for some countries in recent times. The curves in the Fig. 4 are numerical q-exponential functions by the nonlinear least-square fitting of the data. The estimated parameters $\{q, k_0, \delta\}$ are summarized in Table 3. It was found that almost all the cases are well fitted by the q-exponential function and well observed the power-law behavior in the tail of the distribution except for the Korean case. As pointed out in other references, Korea has relatively small number of family names, i.e. the variety of family names is extremely low.

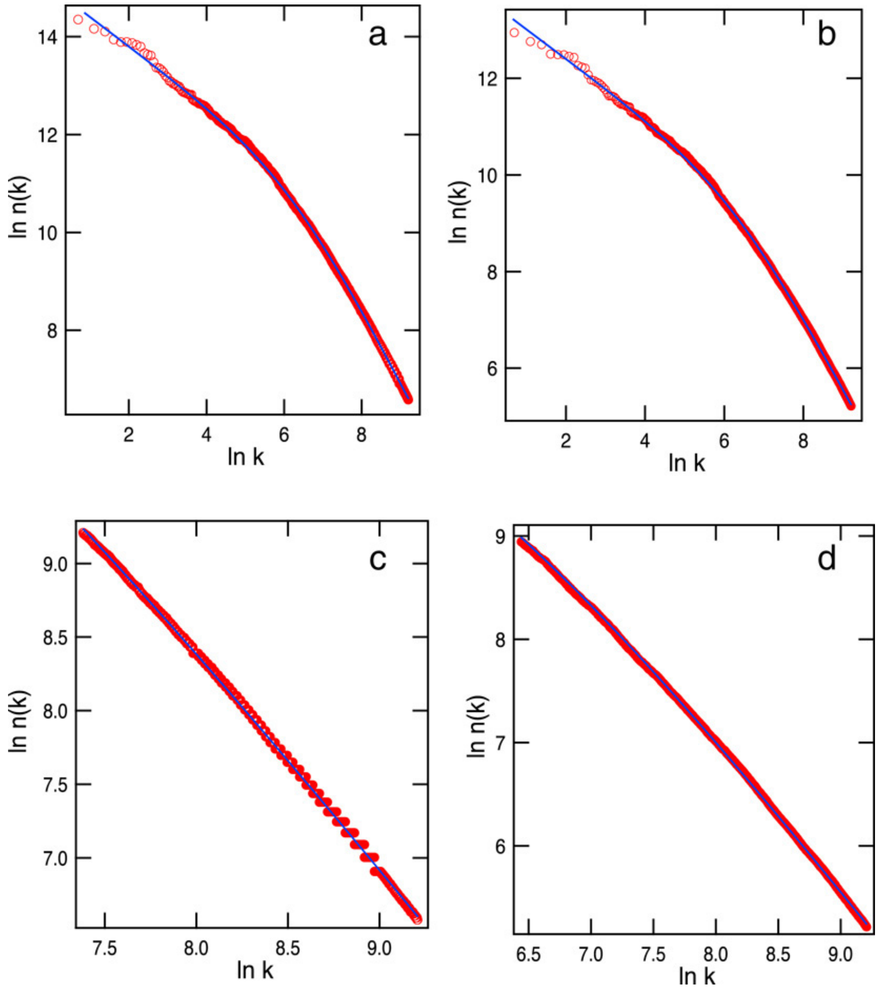

Fig. 2. (Color online) Log–log plots of the rank-size distributions of family names in the Japanese society (denoted by circles) and fit to the q-exponential function (denoted by full line). In the panels (a) and (b), the frequencies are counted in the individual unit and the household unit, respectively. The panels (c) and (d) denote the expanded tails of the distributions in panels (a) and (b), respectively. In both cases we used $N = 10\,000$. The estimated parameters are given in Table 2.

The other interesting point is a similarity of the distribution forms between Japan and Manx, although the population and the number of family name are quite different in the scale. The similarity might be based on the properties of the isolated natural islands that the immigration rate from outside is relatively low.

4. Summary and discussion

We investigated the applicability of the q-exponential function for the rank-size distribution of family names in Japan. The result supports the fact that the q-exponential distribution is relevant to the distribution of family names over the whole frequency regime. The estimated power exponents in the high-frequency regime are larger than those observed by Miyazima et al. Furthermore, we showed numerical results of the fit to data for family names in other countries.

As mentioned in the introduction, the family name distribution is sometimes modeled by branching processes including death rate, birth rate and mutation rate corresponding to the change of family name by marriage. On the other hand, there are some intrinsic properties in the Japanese family names, such as historical, social, ethnic and cultural diversity of family names in Japan. We list up them.

- Japan is the isolated natural islands surrounded by the oceans.
- The living people are consisting of only several kinds of races (probably, such as Ainu people, Okinawa people, Korean-originated people, Chinese-originated people and Yamato-originated people).

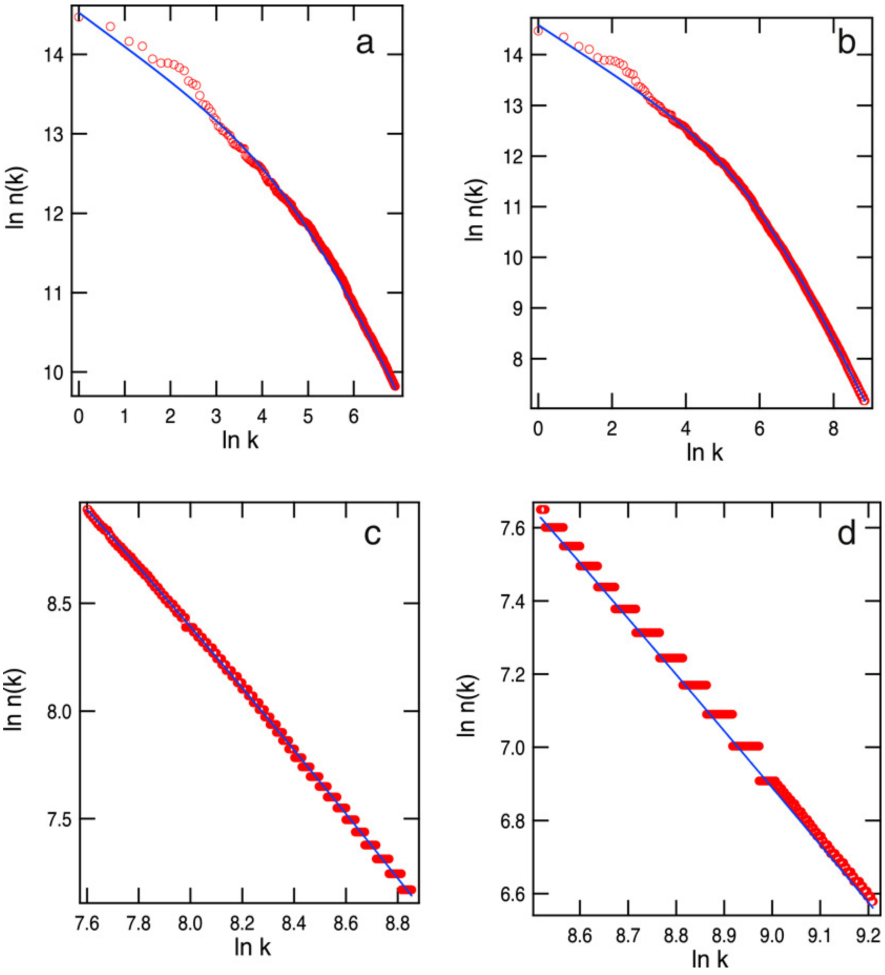

Fig. 3. (Color online) Log–log plots of the rank-size distribution of family names in Japanese society and fit to the q-exponential function. The different ranges for the rank in the individual unit are used for the fit. (a) 1–1000, (b) 1–7000, (c) 2000–7000, and (d)500–10 000. The estimated parameters are given in Table 2.

– Most of the Japanese family names were created about 125 years ago. (Even in recent times, instead of family names, house names are used to distinguish the persons in rural Japan.)

– Most of the Japanese family names were adopted from the names of natural objects, i.e. trees, flowers, animals, landscapes, seasons, etc.

We think that the last item is strongly related to the individual reason that Japanese family names have rich diversity as seen in Section 3.1. It can make a variety of combinations. For example, “Yamada” is made of “Yama” with the meaning of *mountain* and “da” with the meaning of *rice field* in Japanese, while “Iguchi” is made of “I” with the meaning of *well* and “guchi” with the meaning of *mouth* or *entrance*.

We have shown that rank-size distribution of family names can be fitted very well by the q-exponential function derived by the entropy approach in a more natural way that the entire frequency distributions can be fitted without specifying the frequency regions. Especially, it is interesting that the global fitting by the q-exponential function is possible for the frequency distributions of family names of the Japanese, which are realized under such unique conditions. Such function form of the frequency distributions suggests that it is determined not by precise conditions but by global conditions such as Tsallis' nonextensive entropy maximization.

In this paper we dealt with the application of the q-exponential distribution function to the rank-size distributions of family names only. Obviously, it is possible to apply it to income distribution, city-size distribution, and so on [38,39].

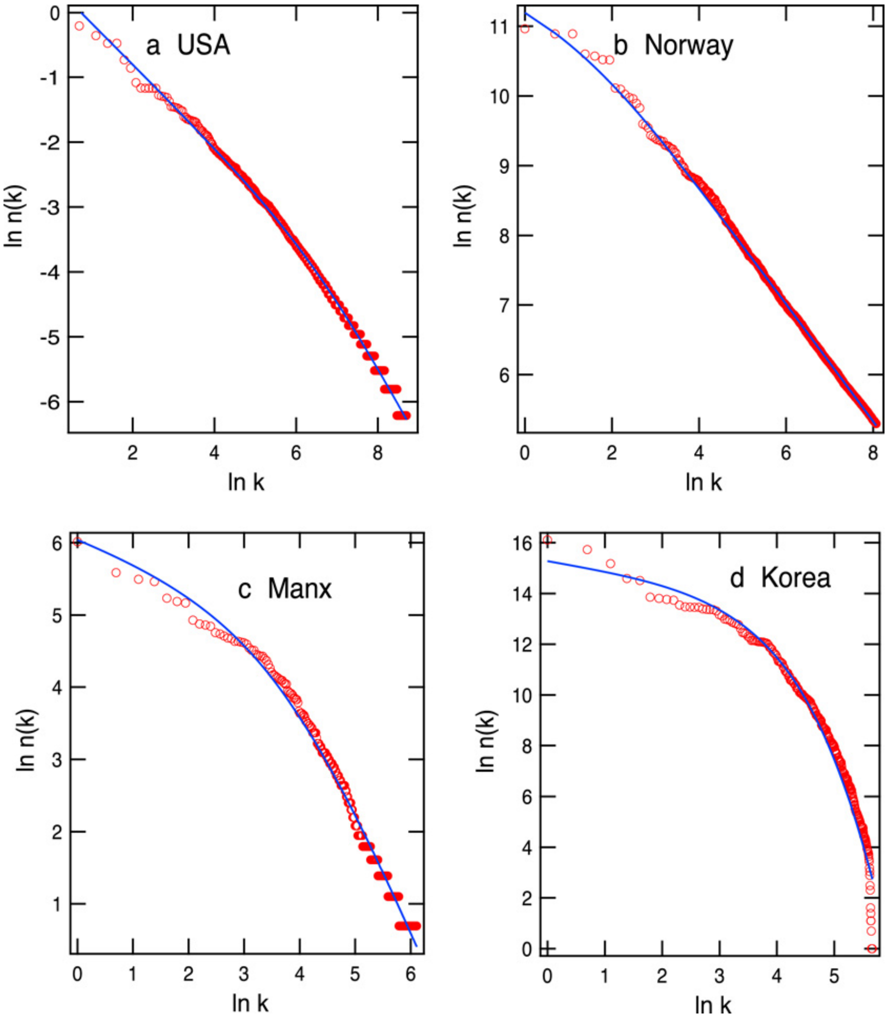

Fig. 4. (Color online) Log-log plots of the rank-size distributions of family names in some countries and fit to the q-exponential function. (a) USA in frequency in percent ($N = 1 - 6000$). (b)Norway($N = 1 - 3206$). (c) Manx ($N = 1 - 449$). (d) Korean($N = 1 - 228$). The estimated parameters are given in Table 3.

Acknowledgments

We would like to thank Shu-ichi Kinoshita and Dr. Jun Hidaka for sending us many relevant papers. We would also like to thank Sachiho Yamada for marshalling some basic data about Family names. K.I. would like to thank Kazuko Iguchi for her financial support and encouragement.

References

[1] V. Parete, Le Cours d'Economie Politique, Macmillan, London, 1897.
[2] B.B. Mandelbrot, Fractals and Scaling in Finance, Springer, Berlin, 1997.
[3] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, Reading, MA, 1949.
[4] F. Ohmori, J. Coll. Sci. Lmp. Univ. Tokyo 7 (1894) 111.
[5] B. Gutenberg, C.F. Richter, Bull. Seism. Soc. Am. 34 (1944) 185.
[6] L.A. Adamic, B.A. Huberman, Glottometric 3 (2002) 143–150.
[7] H.A. Simon, On a class of skew distribution functions, Biometrika 42 (1955) 425–440.
[8] H.A. Simon, Models of Man, Wiley, New York, 1957.
[9] D. Sornette, Phys. Rev. E 57 (1998) 4811.
[10] P.R. Jelenkovi, J. Tan, Proceedings of Forty-Fourth Annual Allerton Conference, Allerton House, UIUC, Illinois, USA, 27–29 September, 2006.
[11] E. Teramoto, Mathematical Ecology: Suuriseitaigaku, Asakura shoten, Tokyo, 1997 (in Japanese); Mathematics of Random phenomena: Random-na-genshou-no-Suugaku, Yoshioka shoten, Kyoto, 1990 (in Japanese).

[12] W.J. Reed, Econo. Lett. 74 (2001) 15–19.

[13] X. Gabaix, Quart. J. Econom. 114 (1999) 739–767.

[14] P. Bak, C. Tang, K. Wiesenfeld, Phys. Rev. Lett. 59 (1987) 381–384;
P. Back, How Nature Works: The Science of Self-Organized Criticality, Springer-Verlag Telos, 1999.

[15] R. Albert, A.-L. Barabási, Rev. Modern Phys. 74 (2002) 47–97.

[16] A.-L. Barabási, E. Bonabeau, Sci. Am. (May) (2003) 60–69.

[17] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Networks: From Biological Nets to the Internet and WWW, Oxford University Press, NY, 2003.

[18] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Phys. Rep. 424 (2006) 175–308.

[19] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 85 (2000) 4629–4632.

[20] S. Bornholdt, H. Ebel, Phys. Rev. E 64 (2001) 035104(R)-1-4.

[21] Z. Liu, Y.-C. Lai, N. Ye, Phys. Rev. E 66 (2002) 036112-1-7.

[22] C. Tsallis, R.S. Mendes, A.R. Plastino, Physics A 261 (1998) 534.

[23] S. Abe, Y. Okamoto (Eds.), Nonextensive Statistical Mechanics and Its Applications, Springer-Verlag, Heidelberg, 2001.

[24] J.P. Boon, C. Tsallis (Eds.), Overview: nonextensive statistical mechanics: new trends, new perspectives, Europhys. News (6) (2005) (special issue). See also a web site, http:/tsallis.cat.cbpf.br/biblio.htm, for the related full bibliography.

[25] S. Abe, N. Suzuki, Europhys. Lett. 65 (2004) 581–586.

[26] R.N. Mantegna, H.E. Stanley, An Introduction to Econophysics: Correlations and Complexity in Finance, Cambridge University Press, 1999.

[27] S.M.D. Queiros, Europhys. Lett. 71 (2005) 339. cond-mat/0502337.

[28] S. Miyazima, Y. Lee, T. Nagamine, H. Miyajima, Physica A 278 (2000) 282–288.

[29] D.H. Zanette, S.C. Manrubia, Physica A 295 (2001) 1;
S.C. Manrubia, D.H. Zanette, J. Theor. Biol. 216 (2002) 461.

[30] W.J. Reed, B.D. Hughes, Physica A 319 (2003) 579–590.

[31] B.J. Kim, S.M. Park, Physica A 347 (2005) 683–694.

[32] Y. Satou, H. Seno, Mathematical Ecology of Conservation and Extinction of Family Names: Sei-no-keisyou-to-Zetumetu-no-Suuriseitaigaku, Kyoto Univ. Press, Kyoto, 2003 (in Japanese).

[33] D.R. White, N. Kejžar, C. Tsallis, D. Farmer, S. White, Phys. Rev. E 73 (2006) 016119-1-8. cond-mat/0508028.

[34] G. Wilk, Acta Phys. Pol. B 36 (2005) 2513–2522.

[35] M.D.S. de Meneses, et al., Prog. Theor. Phys. Suppl. 162 (2006) 131–137.

[36] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing, Cambridge University Press, 1996.

[37] http://www.myj7000.jp-biz.net/, http://www2s.biglobe.ne.jp/suzakihp/index40.html, http://www.census.gov/genealogy/names/dist.all.last, http://www.ssb.no/english/subjects/00/navn-en/etternavn-100.html, http://www.isle-of-man.com/manxnotebook/famhist/fnames/sn1881.htm, http://kosis.nso.go.kr/.

[38] W.G. Song, H.P. Zhang, T. Chen, W.C. Fan, Fire Safety J. 38 (2003) 453–465.

[39] R.N. Onody, P.A. de Castro, Phys. Rev. E 70 (2004) 037103-1-4.