Physica A 390 (2011) 3938–3943

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Chinese family name distributions in multiple scales

Jiawei Chen, Liujun Chen, Yan Liu *, Dahui Wang, Yougui Wang

Department of Systems Science, School of Management, Beijing Normal University, Beijing 100875, PR China

ARTICLE INFO

Article history:
Received 2 September 2010
Received in revised form 8 March 2011
Available online 22 June 2011

Keywords:
Family name distribution
Truncated power law
Multiple scales

ABSTRACT

The distribution of Chinese family names is investigated based on data of the fifth national census of China in 2000, including 7329 Chinese family names and 1.28 billion people. The cumulative distribution function (CDF) of family name in multiple scales is presented and the correlation between distribution and scale is discussed. The fitting results show that CDFs at the scales of country, province, city and county all follow stretched exponential truncated power law, while the fitting parameters vary greatly. A characteristic index is proposed to measure the extent the distribution is stretched-exponential-like. The index increases as the scale goes down, indicating that the distribution at lower scale is more stretched-exponential-like, and the distribution at higher scale is more power-law-like. Such correlation between distribution and scale is verified by investigating how the distribution changes as the scale is expanded county by county. Furthermore, the correlation is qualitatively explained by artificially constructing multiple scales with a partition process.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Family name distribution, which can reveal living habits such as habitation and migration, has been investigated in many countries and regions [1–11]. The frequency distributions of family name in most countries have a form of power-law, $f(n) \propto n^{-\gamma}$, where n is family size defined as the number of people sharing the same family name, and γ is a scaling exponent. The exponents are around 2 for most countries, for example, $\gamma = 1.75$ for Japan, $\gamma = 1.94$ for the United States and $\gamma = 2.16$ for Norway [3]. For Korea, the distribution depends on whether the regional origin of family name is considered, that is, the cumulative distribution without regional origin is logarithmic and that with regional origin is power law [4]. For China, researches based on the top 100 popular family names show that the Zipf plot is exponential and it is maintained since the Song Dynasty [3,5]. However, the conclusion for China is limited since only the top 100 popular names are included. In fact, there are 7329 family names according to the fifth national census of China in 2000. What is the distribution for all Chinese family names? Does the distribution of less popular family names show similar features with that of popular ones?

Although there are many researches on family name distribution, most of them are based on data of country or city, and few involve multiple scales. China has a large population of 1.28 billion inhabiting an area of 9.6 million square kilometers, which is divided into 31 provinces (including municipalities), which are subdivided into 341 cities, and further subdivided into 2861 counties. What are the family name distributions in the scales of country, province, city and county respectively? What are the differences among the distributions in these scales? Is there any explicit correlation between distribution and scale? These questions will be addressed in this paper.

The rest of the paper is organized as follows. In Section 2, we fit the cumulative distribution in the scales of country, province, city and county respectively. To discriminate the distributions in multiple scales, we put forward a characteristic

* Corresponding author.
E-mail address: bnuliuyan@bnu.edu.cn (Y. Liu).

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.06.014

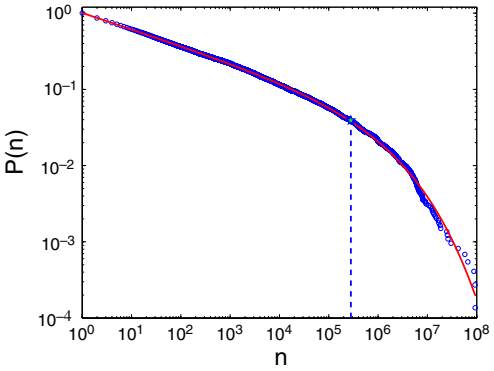

Fig. 1. Log–log plot of $P(n)$ vs n for China. Data (circles) can be well fitted by $P(n) = a \cdot n^{-b} \cdot e^{-(n/n_c)^d}$ (line) with fitting parameters $a = 1.014 \pm 0.011$, $b = 0.2132 \pm 0.0021$, $n_c = (1.148 \pm 0.0700) \times 10^6$, $d = 0.3585 \pm 0.0045$. The asymptotic p-value of KS test is 0.85. The dashed line presents $n_0 = 2.81 \times 10^5$ and $P(n_0) = 0.039$.

index in Section 3. In Section 4, we measure the distributions in the four scales by the index, and accordingly we suppose and verify a general correlation of distribution with scale. A qualitative explanation is given in Section 5.

2. Family name distributions in multiple scales

The data set in this work is based on the fifth national census of China in 2000 (the populations in Hong Kong, Macao and Taiwan are excluded). The data are obtained from the identity information system of China which constructed by the National Citizen Identity Information Center. The data set lists family size of each family name in all county-level administrative divisions (county for short). Detailed information on family name distribution for all the 31 provinces and the median values for the cities and counties are shown in Table 1.

Cumulative distribution of family names represents the proportion of family names whose sizes are no less than n, denoted by $P(n)$. Empirical cumulative distribution function (CDF) for Chinese family names evidently follows a truncated power law as shown in the log–log plot (Fig. 1). In order to include the rapidly decreasing tail, a stretched exponential is adopted as the crossover function. So we fit the data with a stretched exponential truncated power-law function[12] in the following form,

$$P(n) = a \cdot n^{-b} \cdot e^{-(n/n_c)^d} \tag{1}$$

where b is the power exponent, n_c is the cutoff size of power law behavior and d is the stretch parameter of the stretched exponential crossover function. The goodness-of-fit of CDF shows that it can pass the Kolmogorov–Smirnov (KS) test [13]. The fitting result indicates that CDF of Chinese family names is a truncated power law with a stretched exponential tail, implying that the upper part of the distribution is power-law-like while the lower part is stretched-exponential-like as shown in Fig. 1.

At the other three scales, all CDFs of family names can be well fitted by stretched exponential truncated power law functions just as Eq. (1). But the fitting parameters vary greatly among provinces, among cities and among counties, so the corresponding median values are presented in Table 1. Different parameters of CDF can represent quite different distributions even if they are all stretched exponential truncated power laws. The cutoff size of power law n_c is often treated as a crossover point and is used to characterize crossover behavior of truncated power law functions. By comparing n_c at the scale of country, province, city and county, we can see a difference of one or two orders of magnitude between two adjacent scales in turn as shown in Column 6 of Table 1. Considering that the population sizes also have a difference of one order of magnitude between two adjacent scales as shown in Column 3, the n_c needs to be re-scaled to characterize the difference of CDFs. Furthermore, the values of power exponent b are also quite different at the four scales as shown in Column 5. So an integrated index including more than one parameters is needed to discriminate the stretched exponential truncated power law functions in multiple scales.

3. A characteristic index for stretched exponential truncated power law distribution

Stretched exponential truncated power law distribution shows a crossover from power law to stretched exponential behavior. Whether the whole distribution function is more power-law-like or more stretched-exponential-like can be judged by the relative importance of the power law factor n^{-b} and the stretched exponential factor $e^{-(n/n_c)^d}$. According to

$$\frac{dP(n)}{dn} = a \cdot n^{-b-1} \cdot e^{-(n/n_c)^d} \cdot \left(-b - d \cdot \left(\frac{n}{n_c} \right)^d \right) \tag{2}$$

Table 1
Family name distributions at the scales of country, province, city and county. N denotes the population size, N_f the number of family name, p the asymptotic p-value of KS test.

Items	N_f	$N(10^6)$	a	b	$n_c(10^3)$	d	n_0	$P(n_0)$	p
Country	7329	1280	1.01	0.21	1148	0.36	281000	0.039	0.85
Province (median)	2991	37.45	0.64	0.18	67.41	0.37	8795	0.09	0.70
Tibet	2470	2.54	1.16	0.42	67.41	0.90	29151	0.01	1.00
Guangdong	2991	79.08	0.62	0.24	793.36	0.60	169616	0.02	0.46
Hubei	4058	59.44	0.56	0.23	300.80	0.53	61546	0.03	0.50
Hainan	1798	8.08	0.64	0.25	76.19	0.50	19354	0.03	0.80
Sichuan	4330	86.60	0.83	0.24	236.78	0.43	60923	0.03	1.00
Guangxi	2872	49.17	0.58	0.22	293.01	0.51	56446	0.03	0.48
Fujian	2162	33.68	0.57	0.22	308.15	0.52	60574	0.04	0.63
Anhui	4451	65.15	0.83	0.24	165.33	0.42	45981	0.04	1.00
Hunan	3331	67.61	0.48	0.20	356.67	0.53	56626	0.04	0.43
Guizhou	3333	38.49	0.57	0.22	158.90	0.48	30461	0.04	0.14
Jiangxi	2607	43.75	0.53	0.20	262.87	0.59	41048	0.05	0.55
Shan1xi	3326	32.99	0.64	0.20	53.15	0.36	11019	0.06	0.65
Henan	4282	100.94	0.67	0.19	105.29	0.34	19992	0.06	0.95
Zhejiang	2415	45.89	0.55	0.18	206.59	0.54	27436	0.06	0.39
Jiangsu	3116	72.66	0.61	0.17	123.46	0.41	14245	0.09	0.90
Xinjiang	3500	18.95	0.85	0.21	19.67	0.37	4377	0.09	0.93
Chongqing	2360	31.98	0.51	0.16	86.97	0.46	8795	0.09	0.41
Shangdong	3429	92.89	0.63	0.16	70.04	0.33	7729	0.10	0.57
Yunnan	4350	42.69	0.90	0.19	10.59	0.28	2573	0.10	1.00
Shan3xi	3329	37.10	0.60	0.15	17.92	0.32	1579	0.13	0.79
Hebei	3405	68.73	0.66	0.15	20.66	0.28	1986	0.13	0.90
Heilongjiang	2764	37.45	0.64	0.14	18.83	0.31	1451	0.14	0.93
Shanghai	1614	13.68	0.78	0.17	29.15	0.43	3246	0.14	0.99
Neimenggu	3634	23.52	0.77	0.16	5.22	0.28	700	0.15	0.87
Jilin	2718	26.57	0.59	0.13	10.54	0.30	612	0.17	0.62
Liaoning	2698	41.92	0.60	0.12	16.09	0.31	738	0.18	0.70
Gansu	3191	25.96	0.70	0.12	3.45	0.27	168	0.23	0.63
Beijing	1997	11.89	0.90	0.08	0.71	0.25	8	0.58	0.97
Qinghai	2905	5.01	0.78	0.07	0.26	0.26	2	0.79	0.57
Tianjin	1657	9.10	0.88	0.04	0.23	0.22	1	1.00	0.90
Ningxia	2044	5.86	0.72	0.03	0.15	0.23	1	1.00	0.68
City (median)	1330	3.19	0.67	0.13	3.25	0.35	252	0.21	0.86
County (median)	614	0.37	0.91	0.08	0.26	0.33	3	0.66	0.97

the relative importance should be measured by the relative magnitude of b and $d \cdot (\frac{n}{n_c})^d$. So the effect of the power law factor on the distribution and that of the stretched exponential factor are equal when $n = n_c \cdot \left(\frac{b}{d}\right)^{\frac{1}{d}}$, which is denoted by n_0. That is, for $n < n_0$, the power law factor has more effect, while for $n > n_0$, the stretched exponential factor has more effect. Since

$$P(n_0) = a \cdot \left(n_c^d \cdot \frac{b}{d} \cdot e\right)^{-\frac{b}{d}} \tag{3}$$

represents the proportion of family names whose sizes are no less than n_0, it can be used to measure the extent the distribution is stretched-exponential-like. So $P(n_0)$ can be taken as a characteristic index of the stretched exponential truncated power law distribution.

In order to verify the validation of $P(n_0)$ as a characteristic index, we divide the whole data of the country into two parts by the separatrix of $n_0 = 2.81 \times 10^5$ and $P(n_0) = 0.039$ as shown in Fig. 1. We fit the part of $n < n_0$ with a power law $P(n) = a' n^{-b'}$, the fitting parameters are $a' = 1.342$, $b' = 0.27$. The fitting can pass the KS test with $p = 0.19$. For the part of $n > n_0$, we fit it with a stretched exponential $P(n) = a' \mathrm{e}^{-(\frac{n}{n_c})^{d'}}$, the fitting parameters are $a' = 1.47$ and $d' = 0.54$. The fitting can also pass the KS test with $p = 0.30$. It indicates that the distribution of the 96.1% less popular family names follows a power law, and the distribution of the 3.9% popular family names follows a stretched exponential law. So n_0 can be taken as the crossover point from power law to stretched exponential and $P(n_0)$ can be used as an characteristic index. Specifically, the larger $P(n_0)$ is, the more the whole distribution is stretched-exponential-like, and the smaller $P(n_0)$ is, the more the whole distribution is power-law-like.

4. Correlation between distribution and scale

By the characteristic index $P(n_0)$, we can discriminate family name distributions in multiple scales and describe the correlation between distribution and scale. As it is shown in Table 1, the value of $P(n_0)$ for the country is 0.039, while the

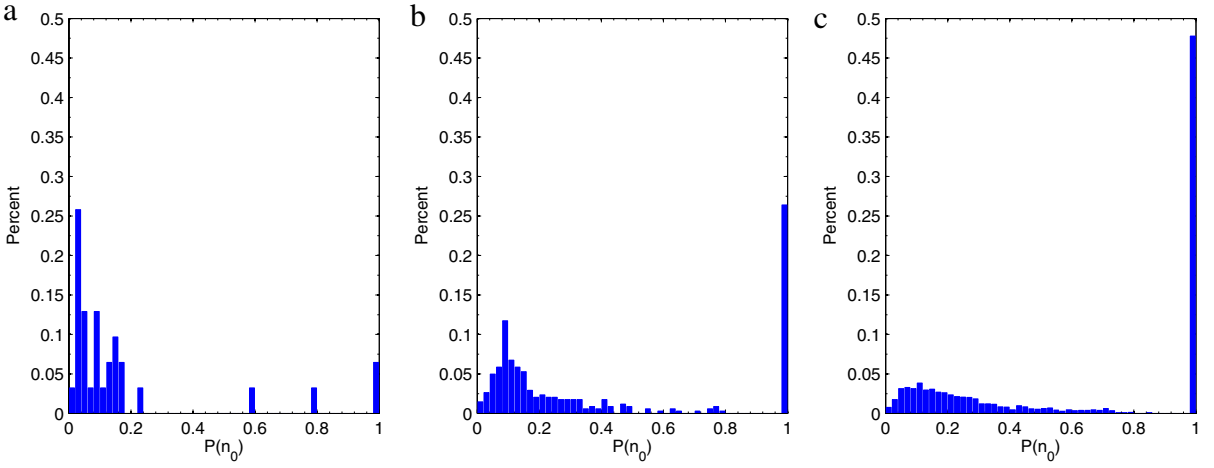

Fig. 2. Histograms of $P(n_0)$ for provinces, cities and counties. (a) For 31 provinces, there are 22 provinces with $P(n_0) > 0.039$. (b) For 341 cities, there are 328 cities with $P(n_0) > 0.039$ and 90 cities whose distributions are stretched exponential since their $P(n_0) = 1$. (c) For the 2861 counties, there are 2795 counties with $P(n_0) > 0.039$ and 1367 counties whose distributions are stretched exponential with $P(n_0) = 1$.

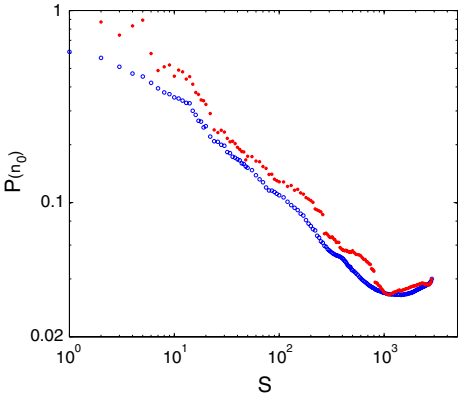

Fig. 3. Log–log plot of $P(n_0)$ vs S. Circles are the values of a single fitting and dots are the average values of 50 fittings.

median values of $P(n_0)$ at the scales of province, city and county are 0.085, 0.21 and 0.66 respectively. That is, the values of $P(n_0)$ at the lower scale are larger than that at the higher scale. The conclusion can also be seen from the histogram of $P(n_0)$ as shown in Fig. 2, where the histogram for cities are more skewed to the right than that for provinces, and that for counties are more skewed to the right than that for cities. So the extent that the distribution is stretched-exponential-like increases in a descending order from country to province, city and county. It is worth noting that the above conclusion is statistical, not for each province, city or county.

A supposition can be inferred from the preceding results that the distribution of family names at lower scale would be more stretched-exponential-like, and that at higher scale would be more power-law-like. The scale used above means the level of administrative division and there is a hierarchical relationship among them. That is, the country consists of provinces, and each province consists of cities and so on.

To verify the correlation between distribution and scale more generally, we investigate how the distribution changes as the scale is expanded county by county. We start from a randomly chosen county as region R_1. Then we expand this region by adding the nearest county to get region R_2. We continue the process and get region R_3, R_4, and so on, until the final one R_{2861} which includes all the Chinese counties. We define scale of a region as the number of inclusive counties and let S denote it, then we have $S \in [1, 2861]$. The corresponding CDFs at all these scales are well fitted by a stretched exponential truncated power law function, and the corresponding values of $P(n_0)$ are obtained. The dependence of $P(n_0)$ on scale S is shown in Fig. 3, from which we can find the supposition holds. In other words, the higher the scale is, the more the distribution is power-law-like.

Table 2
Family name distributions in constructed multiple scales. m denotes the number of divisions, the variables with prime denote those for one division, and the others are the same as Table 1.

m	N'_f	$N'(10^6)$	a	b	$n_c(10^3)$	d	n'_0	$P(n'_0)$	p
1	7329	1279.36	1.01	0.21	1147.60	0.35	280125	0.039	0.85
10	5400	127.94	0.85	0.22	124.67	0.35	30940	0.051	0.98
50	3950	25.59	0.79	0.21	20.64	0.34	4687	0.077	0.98
100	3399	12.79	0.79	0.20	8.71	0.33	1813	0.097	0.99
200	2913	6.40	0.79	0.19	3.46	0.32	633	0.133	1
400	2496	3.20	0.80	0.17	1.29	0.31	192	0.184	1
800	2146	1.60	0.83	0.15	0.43	0.30	47	0.270	1
1600	1836	0.80	0.89	0.12	0.12	0.28	7	0.452	1
3200	1559	0.40	1.07	0.08	0.02	0.25	1	1.000	0.99

5. A qualitative explanation

To give a qualitative explanation for such a correlation, we artificially construct family name distribution in multiple scales by successively dividing the whole Chinese population into smaller and smaller parts. We expect this correlation between distribution and scale remains during this partition process.

The population of each family name is divided into m parts almost equally, so the scale of each part is $1/m$. In particular, when a family size n is less than m, they will be randomly allocated to the m parts. Then, we can have the corresponding CDF of this scale. The CDFs at all the scales can be well fitted by stretched exponential truncated power law function, and the fitting results are shown in Table 2. It is obvious that the characteristic index $P(n_0)$ monotonically increases with m, indicating that the correlation really remains.

This evidence can be qualitatively explained by analyzing the relationship between CDF for the whole country $P(n)$ and that for any one part $P(n')$. After partition, the new family size in a part is $n' = \lfloor \frac{n}{m} \rfloor$ or $\lceil \frac{n}{m} \rceil$. Consequently, the new CDF of family names $P(n')$ can be obtained by a transformation of $P(n)$. Firstly, $P(n)$ is shifted to the left by $\lg m$ in the log–log plot, since the average family size in a part is $1/m$ of that of the whole country. Then, the parts of $n < m$ are cut off for they are mingled into $n' = 1$. Finally, the remaining part is stretched upward with a multiplication of $\lg \frac{N_f}{N'_f}$ at $n' = 1$ to compensate the loss of the number of family names resulting from cutting. This transformed distribution curve is quite similar to the actual $P(n')$. Given a larger m, we need to shift $P(n)$ more to the left and upper during the transformation. As a result, the part of power law shrinks more and $P(n'_0)$ is much greater than $P(n_0)$.

This analysis can also infer that the correlation between distribution and scale is mainly caused by the specific shape of CDF for the whole country. As long as the CDF for a larger scale takes the form of a truncated power law, a partition process will lessen the ratio of the power-law part to the CDF for a small scale. In contrast, when family names follow solely a power-law distribution, the distribution should keep the same as the whole population is divided into smaller parts as demonstrated in the case of Japan [1]. We performed a partition process for a supposed power-law distribution of family names and verified this supposition.

6. Conclusion

In this paper, we focus on the family name distribution for all the 7329 Chinese family names. The fitting results show that CDFs at the scales of country, province, city and county all follow a stretched exponential truncated power law. At a given scale, the part of less popular family names ($n < n_0$) can be well fitted by a power law, while the part of popular family names ($n > n_0$) can be well fitted by stretching exponential. Thus we have a crossover point n_0 which separates the CDF of family names into two parts. The proportion of the stretched exponential part can be measured by a characteristic index $P(n_0)$. Estimating the values of $P(n_0)$ for the scale of country, province, city and county respectively, we find that family name distribution at lower scale is more stretched-exponential-like, and distribution at higher scale is more power-law-like. A general supposition is put forward that $P(n_0)$ decreases as the scale is expanded part by part. This kind of correlation is verified by successively dividing the whole population into smaller and smaller parts. This scale-dependency of family name distribution will be meaningful to other issues of statistics and worthy for further research.

Acknowledgments

This work is supported by NSFC under the grant No. 70771012, No. 70601002, and the Fundamental Research Funds for the Central Universities. We are grateful to Professor Yida Yuan for his warm heart in sharing the data.

References

[1] S. Miyazima, Y. Lee, T. Nagamine, H.S. Miyajima, Physica A 278 (2000) 282.
[2] D.H. Zanette, S.C. Manrubia, Physica A 295 (2001) 1.

[3] S.K. Baek, H. Kiet, B.J. Kim, Physical Review E 76 (2007) 046113.
[4] B.J. Kim, S.M. Park, Physica A 347 (2005) 683.
[5] Y. Yuan, C. Zhang, Chinese Surnames: Community Heredity and Population Distribution, East China Normal University Press, Shanghai, 2002, pp. 21–57 (in Chinese).
[6] M. Dzierzawa, M.J. Omero, Physica A 287 (2000) 321.
[7] S.C. Manrubia, D.H. Zanette, Journal of Theoretical Biology 216 (2002) 461.
[8] W.J. Reed, B.D. Hughes, Physica A 319 (2003) 579.
[9] C. Scapoli, H. Goebl, S. Sobota, E. Mamolini, A. Rodriguez-Larralde, I. Barrai, Journal of Theoretical Biology 237 (2005) 75.
[10] A. Luca, P. Rossi, Physica A 388 (2009) 3609.
[11] H.S. Yamada, K. Iguchi, Physica A 387 (2008) 1628.
[12] E. Bonabeau, L. Dagorn, P. Freon, Proceedings of the National Academy of Sciences 96 (1999) 4472.
[13] F.J. Massey, Journal of the American Statistical Association 46 (253) (1951) 68.