Statistical distribution of Chinese names*

Guo Jin-Zhong(郭金忠), Chen Qing-Hua(陈清华)†, and Wang You-Gui(王有贵)

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

(Received 6 May 2011; revised manuscript received 14 June 2011)

This paper studies the statistical characteristics of Chinese surnames, first names and full names based on a credible sample. The distribution of Chinese surnames, unlike that in any other countries, shows an exponential pattern in the top part and a power-law pattern in the tail part. The distributions of Chinese first names and full names have the characteristics of a power law with different exponents. Finally, the interrelation of the first name and the surname is demonstrated by using a computer simulation and an exhibition of the name network. Chinese people take the surname into account when they choose a first name for somebody.

Keywords: surname, first name, full name, distribution

PACS: 89.65.Cd, 01.75.+m **DOI:** 10.1088/1674-1056/20/11/118901

1. Introduction

Everyone has his or her own name. In most cultures, a person's name is composed of a surname and a first name. These two components have quite different characteristics. Usually, the person's surname is inherited from his or her parents and in most cases from his or her father. Thus a family share the same surname and it is also called the family name. The first name or the given name is given by free choice as opposed to the inherited surname, so it can be recognized as an individual symbol of one person.

Since one's surname is inherited from his or her ancestor mostly, it thus has many genetic characteristics. So surnames are used to study the genes and genetics migration,[1,2] the human population growth,[3,4] the demographic structure and the cultural evolution.[5,6] Likewise, first names can also be used to study a variety of interesting scientific problems. Some scholars focus on the relationship between the given name and the personality, intelligence, school achievement.[7,8] And many scholars put enthusiasm into the first name, because they consider it as the most important object of individual choice behaviour.[9,10] Recently, the baby name in drifting is recognized as a mechanism for cultural change and understanding the relationship between individual decision making and collective outcomes.[11,12]

Usually, in those works, some meaningful inferences are drawn from the statistical analysis, which are necessary for further investigations. In fact, some researchers concentrate on the statistical characteristics of names. Many of them look into the distributions of the surnames in different countries and regions including the United States,[13] Japan,[14] Republic of Korea,[15] and some European countries.[16] It is found that the surnames follow a power-law distribution in those cases. Those empirical studies have triggered considerable theoretical attempts.[17,18] The studies on the first names are fewer than that on the surnames. There are several researchers who have studied the first name distributions in England[19] and in the U.S.[11,12] They claim that the first names in those countries also show the power-law feature. There are rare paper concerning the full names.[20]

All those researchers put their efforts into the study of name's distributions in the U.S. and other developed countries. China possesses one fifth of the world population and has a unique culture and a unique name structure. But the statistical characteristics of Chinese names have not yet received sufficient attention they deserved. Only a few pioneering scholars have carried out some simple statistic analyses on the top 100 Chinese surnames.[21,22] Up to now, there is still no systematic study yet on the statistics of Chinese first names and full names. Due to the absence of reliable statistics on Chinese names, no more extensive researches have been done.

This paper will examine the distributions of Chi-

*Project supported by the National Natural Science Foundation of China (Grant No. 61174165) and the Fundamental Research Funds for the Central Universities, China

†Corresponding author. E-mail: qinghuachen@bnu.edu.cn

© 2011 Chinese Physical Society and IOP Publishing Ltd

nese surnames, first names and full names based on a presentative sample. The present paper is organized as follows. In Section 2, we explain the data sources and check the sample with the whole population. The distributions of Chinese surnames, first names and full names are presented in Section 3. A result of computer simulation is given and compared with the real data in Section 4. A summary and conclusion is offered in the last section.

2. Data source

The data of all Chinese full names is not available to the public. In order to make a reliable statistic analysis, we need a representative sample of the complete data. This kind of sample is found in the donation records from the CRCF (Chinese Red Cross Foundation). The full names of people who have donated their money to the CRCF for the Wenchuan earthquake victims through the CCB (China Construction Bank) from May 12th to June 21st are collected. The original record of the sample contains 250477 items, among which some names are found to be invalid, as they are firms, anonymities and so on. We delete them and obtain 221739 valid items finally. The refined data is used as an effective sample for our following analysis. The total number of surnames in the sample is 1076, including 17 compound surnames.

Although our sample is very small in respect to the large population in China, it can be confirmed that the data are representative. The enthusiasm of donation was evoked all over the nation at that time and the donations of people were transferred to the CRCF in various ways. The donors who remitted money through the CCB were occasional. In addition, the possibility that a person donated his money two or more times through the same bank during that short period can be neglected. So the data we used can be regarded as an approximately random sampling from the whole population in mainland China. This fact has been confirmed by the donations' consistency with the wealth distribution.[23] The representativeness of our sample can be further validated by a comparison between the statistic result from our sample and that from the 2000 Census data. The comparison is presented in Fig. 1, where the distributions of surnames are illustrated seperately. It can be seen that there are little difference between the two cases. Therefore, we are convinced that the data can be used to demonstrate the distribution of Chinese names properly.

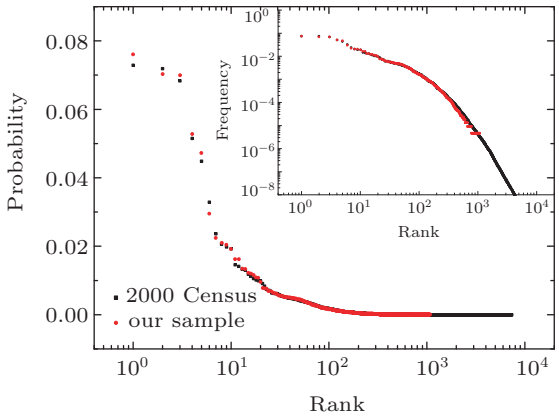


Fig. 1. (colour online) The Zipf plot of surname distributions of our sample (circles) and the 2000 Census (squares). The insert shows the data in the log–log scale.

3. Distribution of names

3.1. Distribution of surnames

The surnames in our sample exhibit extreme inequality in populatrity. For example, the surnames of Wang (王), Zhang (张), and Li (李) are the most popular ones, and their populations (the shares of the total population) are 16865 (7.6%), 15588 (7.03%) and 15509 (6.99%), respectively. In contrast, there are only 6 (0.003%) persons using the surname of Bao (保) and 296 surnames, such as Huyan (呼延) and Qian (乾), are applicable to only one person. Obviously, a smaller number of surnames are shared by the majority of Chinese people, while the majority of surnames are attributed to only a few persons. The skewed character of Chinese surnames suggests a power law distribution, which is a dominant pattern of the surname statistics in other countries.[18] So, we firstly plot the surname distribution of our sample in double logarithmic scales in Fig. 2.

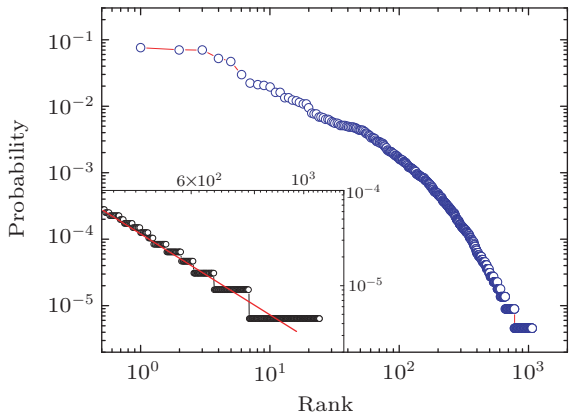


Fig. 2. (colour online) Distribution of surname in double-logarithmic scales. The distribution of small Chinese surnames after the 400th is given in the insert, which shows the power-law feature.

Obviously, the overall distribution of Chinese surnames differs from the power-law distribution. However, the tail part with bigger ranks of the distribution displays a straight line pattern in the double-log scales, which is shown in the insert of Fig. 2. This means that the surname distribution has a power-law tail. When we re-scale the distribution in a semi-logarithmic plot, we can find that the mass of the top part with smaller ranks presents an obvious exponential distribution, as shown in Fig. 3.

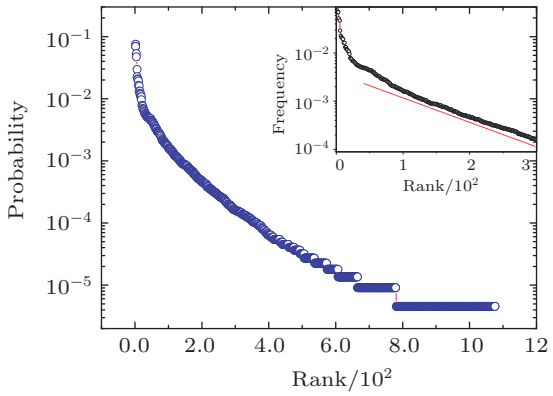


Fig. 3. (colour online) Distribution of the surnames in single-logarithmic scales. The middle part of the distribution is shown in the insert, which exhibits the exponential feature.

The preceding mixed feature of exponential and power law characters presented in the Chinese surname distribution curve indicates a truncated exponential function of $P = A\,\mathrm{e}^{Br}r^{C}$,[24] where r is the rank, A, B and C are fitting parameters. We make the fitting with the formula and obtain $P(r) = 0.0913e^{-0.0425r}r^{-0.415}$. The coefficient of determination of our fitting (R^2) is 0.95.

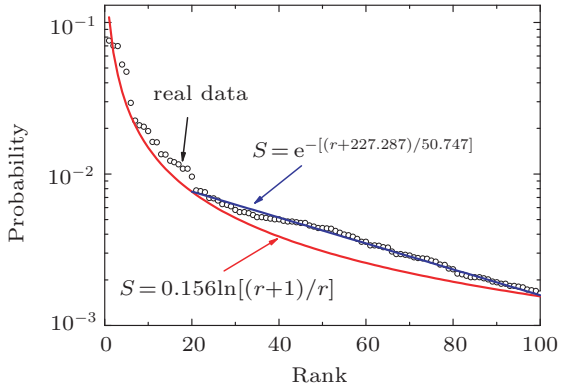


Fig. 4. (colour online) Distribution of the top 100 surnames and the fittings.

In the previous studies on Chinese surnames, Yuan proposed a formula of $S = 0.156\ln[(r+1)/r]$ to describe the distribution of the top 100 Chinese surnames.[21] We pick up the top 100 surnames of our sample and check the fitness of the form. As shown in Fig. 4, the fitting result is good enough to make us certain of the validity of our sample once more. However, the exponential character is still dominant in this part.

3.2. Distribution of first names

For Chinese people, the first name usually consists of a single character or two characters. In our sample, the first names of 53779 (24.25%) persons contain only one character and the others (167960, 75.75%) have two characters. There are totally 66923 varieties of first names in our sample, including 1901 (2.841%) with a single character and 65022 (97.159%) with two characters. In average, each single-character first name is shared by 28.3 (i.e. 53779/1901) individuals, while each two-character one only corresponds to 2.6 (i.e. 167960/65022) individuals. A single-character first name replicates 10 times more than a two-character one.

The popularity of one-character first name can be illustrated in Fig. 5, where the proportions of two kinds of first names are displayed respectively versus their ranks accordingly. From the figure, it can be seen that the single-character proportion in each bin of 100 ranks is almost continually decreasing as the rank goes up. Within the top 100 ranks, the single-character first names account for 78%, at the end of the ranks, the ratio decreases to about 1.6%.

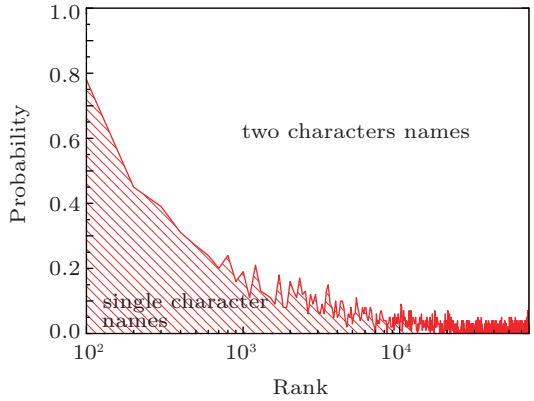


Fig. 5. (colour online) Proportions of the first names with different lengths of characters. It is counted within every 100 ranks.

The Zipf plot of Chinese first names, including both single-character and two-character ones, is shown in Fig. 6. Obviously, the first names in the upper part of the distribution are shared by many people, while those in the tail part are shared by only a few people.

For example, there are 1050 persons named Wei(伟), which accounts for 0.474% of the whole sample and for the two-character ones, Hai-Yan (海燕) is the most popular first name corresponding to 346 (a share of 0.156%) individuals. In contrast, some first names, such as Lu (麓) and Xiao-Fu (孝富), is possed by only one person. From Fig. 6, the main part of distribution can be characterized as a power law function with the Zipf exponent of 0.8572. It is consistent with the preference attachment in the naming procedure.[11,17]

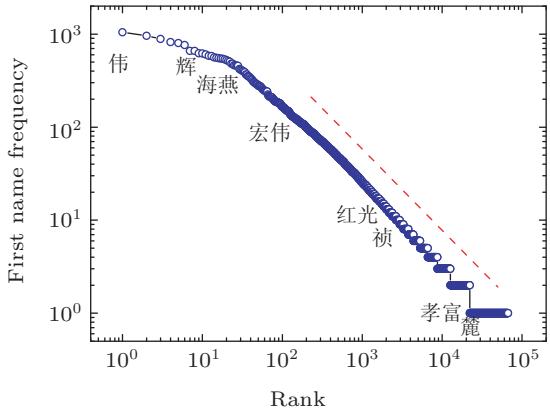


Fig. 6. (colour online) Distribution of Chinese first names in double-logarithmic scales.

3.3. Distribution of full names

A Chinese full name usually consists of a surname and a first name, and the length usually varies from 2 to 4. The statistics show that the three-character full names are most popular in our sample and they cover 167868 (75.7%) persons. 53689 (24.2%) full names possess two characters. Furthermore, very few people (less than 0.1%) have four-character full names.

There are total 166320 varieties of full names in our sample, including 22947 (13.8%) with two characters, 143197 (86.1%) with three characters and 176 (0.1%) with four characters. The proportions of these three kinds of full names versus their ranks are shown in Fig. 7. It is obvious that the proportion of the two-character full names decreases as the rank goes up, while that of the three-character increases accordingly. This pattern is very similar to that of the first names shown in Fig. 5. But there are two significant differences. First, the decreasing speed of the two-character full name is slower than that of the single-character first name. Second, the probability of the two-character full name is always bigger than that of the single-character first name in the corresponding places. For example, the top 100 full names are all two-character, while the corresponding proportion of the single-character first name is only 78%. Even in the last 100 full names, the two-character proportion is still markedly greater than 10%. These differences can be attributed to the feasibility of reduplication of the single-character first name. In fact, a single-character first name is shared by 12.1 types of surnames and a two-character one is shared only by 2.2.

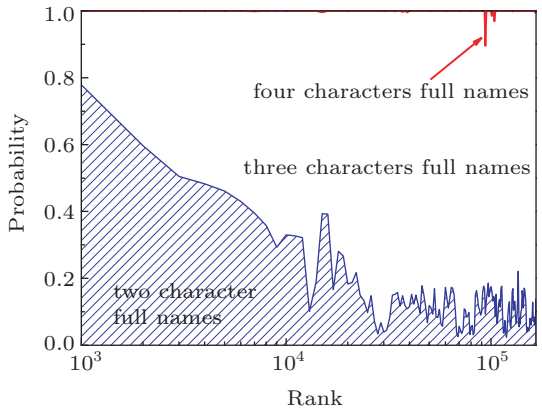


Fig. 7. (colour online) Proportions of full names with two characters, three characters and four characters within every 1000 ranks.

The distribution of the Chinese full names including all two-character, three-character and four-character ones, is presented in Fig. 8. The top part of the distribution possesses a majority of the population, while the tail part possesses a minority. For example, Wang Jing (王静) is the most popular name in our sample and it takes up to a share of 0.052%. There are 40 (0.018%) individuals with the name of Gao Feng (高峰) and 4 (0.0018%) persons are named Li Jian-Ming (李建明). Many full names (account for 144875) like Liang Xue-Yun (梁雪芸) are only used by one person in our sample. From the Zipf plot, we can see that a straight line fits the data well in the log–log scales, which indicates that the distribution of

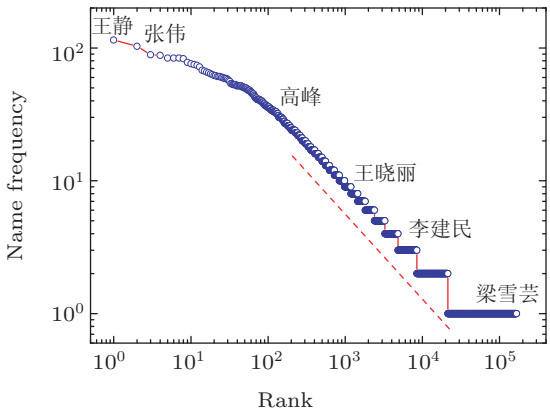


Fig. 8. (colour online) Distribution of the Chinese full names in log–log scales.

the Chinese full name has a power law feature. The Zipf exponent of this distribution is 0.5546 and it is

smaller than that for the first names. In other words, the inequality of the full names is weaker than that of the first names.

4. Correlation of first names and surnames

4.1. Random matching simulation

In Section 3, the statistics of Chinese surnames, first names and full names are analysed and presented respectively. It is found that the distributions of the first names and the full names are similar and both exhibit power law patterns. A question then arises in our hands: whether the full names can be reproduced by a random matching process of the given surnames and the first names.

We discuss this problem by using a computer simulation carried out in exchanges of names among people. There are 221739 agents ranging in a sequence. We initially name them with all full names in our sample accordingly and split all names into first names and surnames. Then, we resort the first names randomly and combine the first names and the surnames to form new name series. Finally, we collect the full names generated and make a statistical analysis.

A simulation result is shown in Fig. 9, where the empirical result is also plotted. The distribution of the produced full names also has a power-law part, which is similar to the real one. However, we can find some distinct deviations from each other. First, the frequency of the popular full names in the real data is greater than that in the simulation result. Actually, the frequency of top three full names in the real data is about 25.8% bigger than that in the simulation result. Second, the number of varieties of full names in the simulation is larger than that of our sample. There are 166320 types of full names in the real data, while we obtain 172231 types of full names in the simulation.

Based on the mean field analysis, if we reproduced the random matching process for sufficient times, each surname and each first name would have chances to form a full name. So the types of full names could reach 72009148, which is 433 times of that of the real names. However, there are only 18407 types that have a mean frequency not less than 1. In average, the full names of the top 3 types are Wang Wei (王伟), Li Wei (李伟) and Zhang Wei (张伟), and the sum of the frequencies is 227, which is 35.2% smaller than that of the top 3 real names. In reality, the top 3 types of full names are Wang Jing (王静), Zhang Wei (张伟) and Li Jing (李静), and the total frequency is 307.

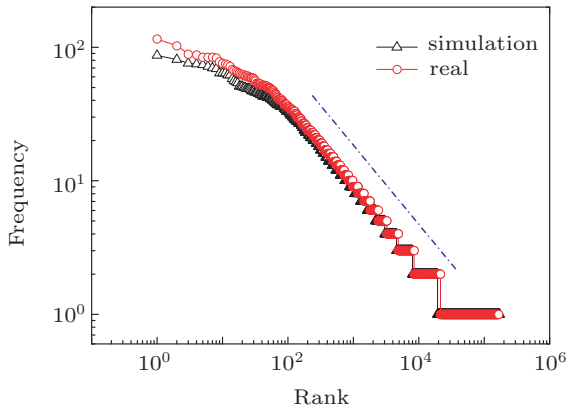


Fig. 9. (colour online) Distribution of Chinese full names in log–log scales.

The deviations of the real case from the simulations may result from name preferences and avoidances. When Chinese people are engaged in naming a child, they prefer a meaningful name. Some unpleasant assemblies in the random matching are excluded in reality.

4.2. Name network

Furthermore, we can discuss the difference between the real and the simulation results based on the concept and the technology of the complex network, which has been applied in diverse fields.[25–28] In the network, the surnames can be considered as one kind of nodes and the first names as another kind. If a full name consists of a surname and a first name, an edge between them will be made. Then we set the link weight by the frequences of the full names. Because connections can only exist between a surname node and a first name node, it is a bipartite network,[29,30] as shown in Fig. 10.

In the sketchy graph shown in Fig. 10, there are four surnames and five first names. The edge weight between Wang (王) and Jing (静) is 115. It means that there are 115 persons named Wang Jing (王静) in our sample. Obviously, there is only one Huyan Ping (呼延平). Besides, the degree of one node (i.e., the number of links from this node) also has meanings. It implies the ability of one node to form full name with the others.

According to the approach introduced previously, we randomly match the surnames with the first names. The random matching is carried out for 3000 times

and we record the frequencies of all kinds of full names after each turn. Thus, we can get frequency series for all full names. Based on these data, the average frequency and the confidence interval at a certain level for each full name can be given. For example, the average frequency of Wang Jing (王静) is 73.1 (it is very close to the value of 73.2 by the mean field analysis) and the 95% confidence interval is $[57, 89]$. It means that no less than 2850 runs (95% of 3000 runs) have the frequency in the range of [57, 89]. And if we narrow down the interval, we cannot have 2850 runs with the frequency in the new range. For convenience, we simply note the interval as 73.1 ± 16. We present the results of the random matching in the weighted bipartite network shown in Fig. 11. Because the surname Huyan (呼延) almost never matches with any first name in all the simulations, it can be regarded as an isolated node in the plot.

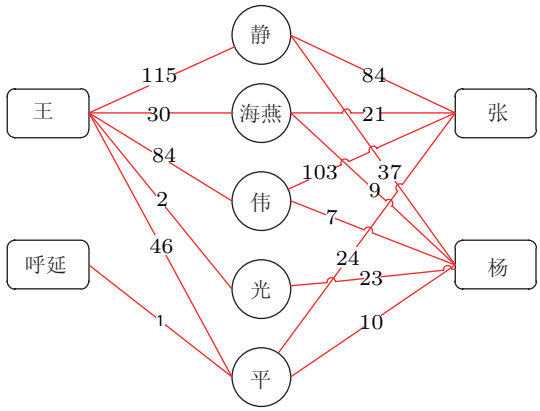


Fig. 10. (colour online) Network of real Chinese names in our sample.

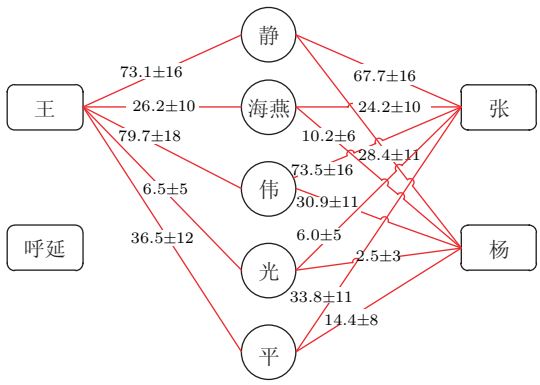


Fig. 11. (colour online) Chinese name network abtained with random matching.

Based on the comparison of Figs. 11 and 10, it is obvious that the real frequencies of some full names occur in the 95% confidence interval of the simulation. However, the frequencies of other full names in reality may exceed the 95% confidence interval. For example, the actual frequencies of Wang Jing (王静), Zhang Wei (张伟) and Yang Guang (杨光) are larger than those in the simulation result. This means that these names are favoured by Chinese people. Furthermore, there are 86 persons whose first name is Guang (光) in our sample and 23 (about 27%) of them are named as Yang Guang (杨光). This dominance of Yang Guang (杨光) as a full name can be explain by that it has the same pronunciation as that for 'sunshine' in Chinese.

However, actual numbers for Wang Guang (王光), Zhang Guang (张光), and Yang Wei (杨伟) are significantly less than the average numbers in the random matching result. The avoidances result from that Guang (光) sometimes means bare or nothing left and Yang Wei (杨伟) sounds like impotence. Chinese people would avoid taking such names with negative meanings or implications. So, it is clear that Chinese people take the surname into account when they choose a first name for somebody.

5. Conclusion

In this paper, the statistical characteristics of Chinese surnames, first names and full names are presented and analysed based on a credible sample. The distribution of Chinese surnames exhibits a mixing pattern of exponential and power laws. This is different from what many researchers have found in other countries. However, the distributions of Chinese first names and full names show some power law features.

In order to examine the dependency of first names on surnames, we mimic the formation process of full names by using a random matching simulation. Starting from the given amount of surnames and first names in our sample, a little larger number of kinds of full names are produced. We find that the distribution of those artificial full names has a power law character, which is very similar to the real ones.

We also find the top part of the simulation results is lower than that of the reality. We argue that the deviation may result from name preferences and aversions. Furthermore, based on the comparison of the name networks of the real data and the result of the random matching, we find that there are clear phenomena of name preferences and avoidances, which implies the interrelation between the surname and the first name. It is obvious that Chinese people would prefer names that have notable meaning and avoid taking names with negative meanings or implications.

References

[1] Lasker G W 1985 *Surnames and Genetic Structure* (Cambridge: Cambridge University Press)

[2] Sykes B and Irven C 2000 *Am. J. Hum. Gene.* **66** 1417

[3] Manni S F, Toupance B, Sabbagh A and Heyer E 2005 *Am. J. Phys. Anthropol.* **126** 214

[4] Sato K and Oguri A 2007 *Jap. J. Ind. Appl. Math.* **24** 119

[5] Manrubia S C and Zanette D H 2002 *J. Theor. Biol.* **216** 461

[6] Scapoli C, Goebl H, Sobota S, Mamolini E, Rodriguez-Larralde A and Barrai I 2005 *J. Theor. Biol.* **237** 75

[7] Schonberg W B and Murphy D M 1974 *J. Soc. Psych.* **93** 147

[8] Busse T V and Seraydarian L 1978 *Psy. in Scho.* **15** 29

[9] Smith-Bannister S 1997 *Names and Naming Patterns in England* (Oxford: Oxford University Press) pp. 1538–1700

[10] Perrin F, Maquet P, Peigneux P, Ruby P, Degueldre C, Balteau E, Del Fiore G, Moonen G, Luxen A and Laureys S 2005 *Neuropsychologia* **43** 12

[11] Hahn M W and Bentley R A 2003 *Proc. R. Soc.* B: *Bio. Sci.* **270** 120

[12] Gureckis T M and Goldstone R L 2009 *Top. Cogn. Sci.* **1** 651

[13] Zanette D H and Manrubia S C 2001 *Physica* A **295** 1

[14] Miyazima S, Lee Y, Nagamine T and Miyajima H 2000 *Physica* A **278** 282

[15] Kim B J and Park S M 2005 *Physica* A **347** 683

[16] Scapoli C, Mamolini E, Carrieri A, Rodriguez-Larralde A and Barrai I 2007 *J. Theor. Biol.* **71** 37

[17] Luca A D and Rossi P 2009 *Physica* A **388** 3609

[18] Baek S K, Kiet H A T and Kim B J 2007 *Phys. Rev.* E **76** 046113

[19] Galbi D A 2003 *Names* **50** 275

[20] Harada M, Sato S and Kazama K 2004 *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* USA p. 306

[21] Yuan Y D 2002 *Chinese Surnames* (Shanghai: East China Normal University Press) (in Chinese)

[22] Wu H, Chou C and Tseng J 2011 *Comp. Phys. Comm.* **182** 201

[23] Chen Q, Wang C and Wang Y 2009 *Eur. Phys. Lett.* **88** 38001

[24] Hernández-Pérez R, Angulo-Brown F and Tun D 2006 *Physica* A **359** 607

[25] Strogatz S H 2001 *Nature* **410** 268

[26] Newman M 2010 *Networks: An Introduction* (New York: Oxford University Press)

[27] Gao L F, Shi J J and Guan S 2010 *Chin. Phys.* B **19** 010512

[28] Li X M, Zeng M H, Zhou J and Li K Z 2010 *Chin. Phys.* B **19** 090510

[29] Zhou T, Ren J, Medo M and Zhang Y C 2007 *Phys. Rev.* E **76** 046115

[30] Chen H B, Fan Y, Fang J Q and Di Z R 2009 *Acta Phys. Sin.* **58** 1383 (in Chinese)