Names | A Journal of Onomastics

The Use of a Matching Preference Index to Empirically Examine Distribution Imbalances in Beijing Citizens' Names

Ziming Zhao
School of Systems Science, Beijing Normal University, Beijing, CHINA

Xiaomeng Li
School of Systems Science, Beijing Normal University, Beijing, CHINA

Qinghua Chen
School of Systems Science, Beijing Normal University, Beijing, CHINA
New England Complex Systems Institute, Cambridge, MA, USA
Department of Chemistry, Brandeis University, Waltham, MA, USA

Abstract
Personal names contain considerable meaningful information about biological and social characteristics of the name-bearer. They also routinely contain important data about cultural preferences in the naming process. Access to this level of information has been limited in the past by a lack of access to large-scale empirical data. As this investigation demonstrates, by utilizing a reliable large-scale sample of Beijing citizens, it is possible to empirically demonstrate onomastic imbalances in the occurrence of Chinese surnames, given names, and full names. In particular, this paper explores the matching imbalance between Chinese surnames and given

ans-names.pitt.edu
ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

names, a phenomenon which has as yet received scant attention in onomastic literature. As this article demonstrates, our innovative quantitative approach makes it possible to reveal statistically significant differences between real names and "random-matching names" that reflect a matching imbalance and imply the probable existence of underlying cultural preferences in Chinese naming processes. The key to this approach is generating a matching preference index (MPI) for names in a dataset. Alongside explaining how this approach is used, this paper offers possible reasons to explain why specific names have higher or lower MPI rankings. As this paper argues, one of the main reasons for these empirical differences may be found in special associations name-givers have within Chinese culture.

KEYWORDS: Chinese names, random matching, MPI, Zipf's Law, anthroponymy

Introduction

Background Information on Naming

Personal names serve more functions than simply providing identification. They also often contain meaningful information, such as social psychological processes (Allen et al. 1941), esthetic values (Finch et al. 1944), ethnicity-specific characteristics (Nick 2017 2013), and ideology (Bloothooft & Groot 2008). For this reason, research on names has been an attractive area of study in many fields. In the field of onomastics, mining hidden naming patterns is extremely important for investigating the historical migration of people and the evolution of cultures and civilizations (Shi et al. 2019). As onomastic research gradually enters the era of big data, to establish a micro-to-macro bridge, statistical analysis has become increasingly important (Castellano et al. 2009).

The study of surnames via statistical analysis and mathematical modeling has become a very active method of large-scale research in this century. A series of studies have found that there is a significant imbalance in the use of surnames in various countries: in the United States (Zanette & Manrubi 2001; Tucker 2001; Hanks & Tucker 2000), Canada (Tucker 2002), Japan (Kamada & Mizuguchi 2020; Hayakawa et al. 2012; Miyazima et al. 2000), Korea (Kim & Park 2005), and some European countries (Mateos & Tucker 2008; Scapoli et al. 2007). Statistically speaking, surnames in these countries largely follow Zipf or power law distributions. Zipf's law is an empirical law formulated using mathematical statistics that refers to the fact that for many types of data studied in the physical and social sciences, the rank-frequency distribution is an inverse relation. The Zipf distribution is one of a family of related discrete power law distributions (Newman 2005).

In contrast to surnames, given names are often more contingent upon the personal choices of name-givers such as parents or guardians. Thus, studies of given names frequently take into account personal preferences and social-psychological phenomena. Studies have suggested that there are sex and age differences in the incidences of and preferences for given names (Joubert 1985; Finch et al. 1944).

Chinese Names and Naming

In the present work, we focus on the widely observed imbalanced distribution of Chinese names. In general, a Chinese surname consists of a single surname and compound surname inherited from one parent (usually the father). By comparison, the given names in China tend to be more flexible in the alternatives possible. In China, naming a baby is a highly complex psychological process. Under the influence of Confucianism, Chinese people attempt to select suitable names by thoughtfully and carefully considering pronunciation and semantics (Li 2007). Chinese given names are indicative of culture, heritage, religion, as well as parental hopes and dreams (Zheng 2011; Ji 2009; Li 2007). In addition to these considerations, the perceived beauty of the name-bearer's potential name as a whole is an integral and indispensable consideration (CCC1982 2015; Bessmum 2008). Previous qualitative analyses have shown that Chinese people often consider whether the surname and given name match across many features such as tone, component structure, and meaning (Gao 2012; Zheng 2011; Li 2007). The meaning of a Chinese character used for a personal name is context-dependent related to its homonym (Gao 2012).

Homonyms are closely linked to China's history and culture. Through association, they can add meaning to Chinese personal names, making them more or less attractive to speakers (Li 2007). For example, the pronunciation of *潘峰* 'Pan Feng' is the same as *攀峰* which symbolizes 'climbing mountains' (Ji 2009). The positive association of this image makes the name 'Pan Feng' more attractive to Chinese speakers. The pronunciation of *杨光* 'Yang Guang' is the same as the pronunciation of 'sunshine' which has a positive association in China. By comparison, the pronunciation of *马伟* 'Ma Wei' is the same as for 'horse tail', which has a less positive association, and makes the name 'Ma Wei' less desirable than 'Yang Guang' (Guo et al. 2011). Preference for certain names is a common cultural phenomenon (Colman et al. 1980). However, where Chinese names are concerned, there has been comparatively little quantitative investigation of this topic due, in part, to the lack of sufficient empirical data (Li 2007).

DOI 10.5195/names.2021.2314 Vol. 69, Issue 3, Summer 2021 ISSN: 0027-7738 (print) 1756-2279 (web)

In our study, we present an innovative, quantitative approach to reveal statistically significant differences between real names and "random-matching names" that reflects a matching imbalance and implies the probable existence of underlying cultural preferences in Chinese naming processes. For our purposes, the term "name" will be used, unless otherwise noted, to mean a "full name" composed of both a given name and a surname. In addition, the surname precedes the given name in our study.

Materials and Methods

Research Purpose and Hypotheses

This study systematically focuses on the imbalanced distribution of Chinese surnames, given names, and full names. The usage imbalance of Chinese full names cannot be explained by the simple combination of imbalances witnessed in Chinese surnames and given names. To gain more information about this phenomenon, the current investigation compares real names and "random-matching names." Furthermore, this study calculates the average matching preference of each name statistically using a matching preference index (MPI).

Data Source: The Beijing Municipal Commission of Transport

To ease traffic congestion and save energy effectively, Beijing implemented a license plate lottery system on January 26, 2011. To apply for these licenses, potential private car buyers were required to register on a website (Beijing Municipal Commission of Transport 2011) by providing their personal information, including their names. The vehicle license plate lottery thus serves as a valuable onomastic data source. Our data were downloaded from the Beijing Municipal Commission of Transport website. The dataset contained 961,765 records established between January and November.

The final onomastic dataset is largely representative of the population and is highly reliable. According to the Beijing Statistical Yearbook, Beijing had approximately 20 million residents in 2011. Thus, our data account for approximately 5% of all Chinese residents, which is a large sample.

Data Processing

For research purposes, we divided Chinese names into surnames and given names. First, we collected 81 of the most common compound surnames on the website (Online Chinese Dictionary 2011). Using this list, we were able to distinguish between single and compound surnames. For example, according to information provided in the Dictionary, the name *欧阳建* 'Ouyang Jian' is most likely composed of *欧阳* 'Ouyang' which is a surname and *建* 'Jian' which is a given name. This division is far more likely than one which would identify *欧* 'Ou' as a surname and *阳建* 'Yangjian' as a given name. Our data include 1,203 single surnames and 59 compound surnames. Table 1 shows a list of the ten most common surnames.

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

Table 1. The Ten Most Common Surnames

			2010 Census			
Our Sample (Beijing)			China		Fangshan District, Beijing	
Rank	Surname	%	Surname	%	Surname	%
1	王 'Wang'	10.19	王 'Wang'	7.10	王 'Wang'	10.3
2	张 'Zhang'	9.18	李 'Li'	6.96	张 'Zhang'	9.1
3	李 'Li'	8.43	张 'Zhang'	6.42	李 'Li'	8.4
4	刘 'Liu'	6.81	刘 'Liu'	5.16	刘 'Liu'	7.3
5	赵 'Zhao'	3.36	陈 'Chen'	4.26	赵 'Zhao'	3.5
6	杨 'Yang'	3.02	杨 'Yang'	2.97	杨 'Yang'	3.4
7	陈 'Chen'	2.85	黄 'Huang'	2.16	陈 'Chen'	2.9
8	孙 'Sun'	1.98	赵 'Zhao'	2.03	高 'Gao'	1.8
9	马 'Ma'	1.73	周 'Zhou'	1.88	孙 'Sun'	1.7
10	高 'Gao'	1.62	吴 'Wu'	1.78	马 'Ma'	1.5
Total		49.17		40.72		49.9

Sources: 2010 Census of China (Wu &Yang 2014); 2010 Census of the Fangshan District, Beijing (Census Office of Fangshan 2011).

As shown in the table above, the proportions and rankings of the ten most common surnames in our dataset are similar to those of the 2010 census of China (Wu & Yang 2014), the 2010 census of the Fangshan district in Beijing (2011). Our distribution of names also matches other population data from China and Beijing reported elsewhere (Youxifeng 2020; Lele 2020; China Global Television Network Society 2020; Bjname 2007; Chen 2006). This similarity provides statistical evidence that our data represent a good sample of Beijing citizens and are, therefore, reliable for quantitatively investigating the matching imbalance.

Display and Measurement of Usage Imbalance

In our work, we quantified usage imbalance statistically, mainly through graphs and isonymy indexes. In our graphic analyses, we used the same method employed for the Zipf distribution: we sorted the surnames, given names, and full names in order from the most to the least frequent. Taking the position of the sequence as the vertical axis and the frequency as the horizontal axis, we drew a distribution curve and determined the distribution characteristics. Our findings showed that the distribution curve decreases quickly and monotonically, which indicates serious maldistribution. Double logarithmic coordinates are used to display our findings.

The isonymy index represents the proportion of two randomly chosen people who have the same original surname. Isonymy is frequently used to display and measure usage imbalance (Liu et al. 2012; Yuan & Zhang 2002). The formula for isonymy is presented below:

$$I = \sum_{k=1}^{S} p_k^2$$

In the above formula, S represents the number of surnames and p_k represents the probability of the k-th surname. A higher isonymy index indicates a greater degree of maldistribution. We used the same method to measure and display the usage imbalance of a given name and full name. The index represents the proportion of two randomly chosen people who have the same given name and full name. For comparison, we calculated the isonymy indexes of each surname, given name, and each full name. The results are presented and discussed in the next section.

Random-Matching Process

In order to reveal whether real names combine in a special relationship, rather than matching randomly, we employed another process. Figure 1 shows the random-matching process. Figure 1 (a) shows the original names. In Figure 1 (b), we split the Chinese names into segments of surname and given name. All of the segments are thoroughly mixed, and each segment of the surname combines with a randomly chosen segment of a given name. Figure 1 (c) shows the random-matching process.

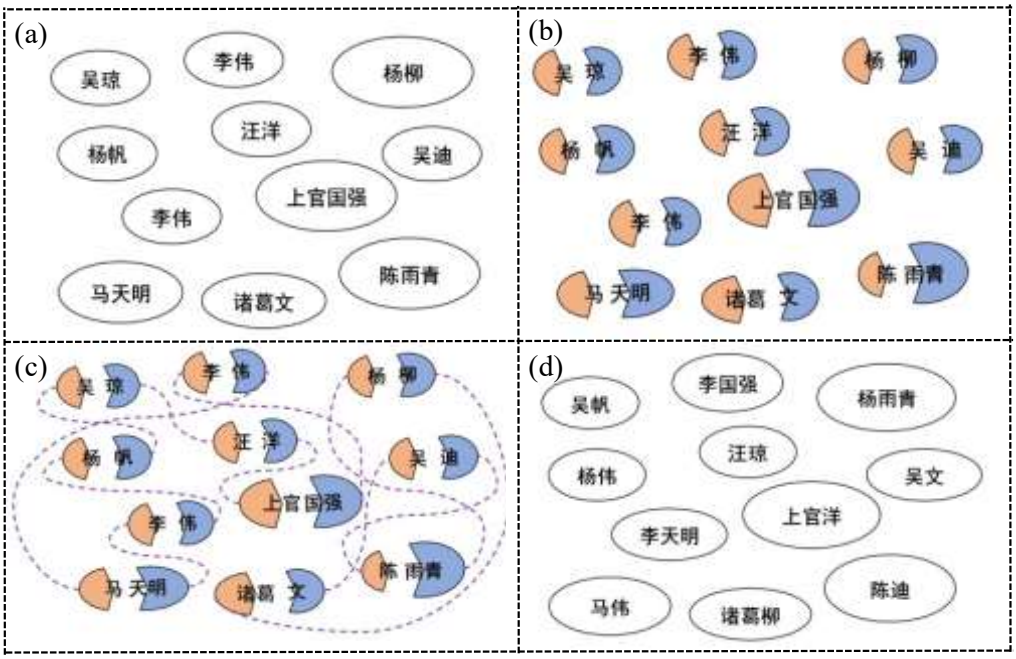

Figure 1. The Process of Random Matching

After the random-matching process, we still have 961,765 full names. These names are not real but were generated based on this process. As shown above in Figure 1 (d), these names are referred to as "random-matching names." However, like the usage imbalance of surnames, given names, and full names, the probability of random-matching names is also imbalanced. The frequency of random-matching names is the theoretical frequency of people not having a preference for any name. Therefore, to measure the matching preference of each name, we compared the frequency (or probability) of real names and random-matching names.

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

Matching Preference Index

Based on the random-matching principle, we defined the MPI for a specific full name as follows:

$$MPI = \frac{p_{real\ name} + \varepsilon}{\bar{p}_{matching\ name}} = \frac{N_{real\ name} + \varepsilon N}{\bar{N}_{matching\ name}} = \frac{P_{real\ name} + \varepsilon}{P_{real\ surname} \times P_{real\ givenname}}$$

$p_{real\ name}$	the probability of a real name
$\bar{p}_{matching\ name}$	the average probability of the name in repeated random-matching processes
$N_{real\ name}$	the real number for a name
$\bar{N}_{matching\ name}$	the average number for a name in repeated random-matching processes
ε	a tunable parameter for distinguishing the difference in matching preferences in the case of $p_{real\ name} = 0$

Here, we use $\varepsilon = 1/N = 1/961765 = 1.04 \times 10^{-6}$. A higher MPI indicates a greater preference for the name, and a lower MPI indicates a lower preference.

Results and Discussion

The Usage Imbalance of Surnames, Given Names, and Full Names

Table 2. The Ten Most Common Given Names and Their Meanings

Rank	Given Name	Meaning	%	Examples
1	伟 'Wei'	'Great'	0.76	张伟 'Zhang Wei' (11.07%); 王伟 'Wang Wei' (10.05%); 李伟 'Li Wei' (9.91%)
2	磊 'Lei'	'Open and upright'	0.68	王磊 'Wang Lei' (12.99%); 张磊 'Zhang Lei' (12.82%); 刘磊 'Liu Lei' (5.35%)
3	超 'Chao'	'Beyond'	0.59	王超 'Wang Chao' (12.43%); 张超 'Zhang Chao' (8.68%); 李超 'Li Chao' (8.05%)
4	静 'Jing'	'Quiet'	0.59	李静 'Li Jing' (10.02%); 张静 'Zhang Jing' (9.37%); 王静 'Wang Jing' (9.26%)
5	涛 'Tao'	'Wave'	0.51	王涛 'Wang Tao' (9.90%); 张涛 'Zhang Tao' (9.48%); 刘涛 'Liu Tao' (6.99%)
6	鹏 'Peng'	'Roc, a large bird'	0.50	王鹏 'Wang Peng' (11.86%); 张鹏 'Zhang Peng' (11.82%); 李鹏 'Li Peng' (9.06%)
7	颖 'Ying'	'Clever'	0.47	王颖 'Wang Ying' (11.38%); 张颖 'Zhang Ying' (10.74%); 刘颖 'Liu Ying' (10.32%)
8	杰 'Jie'	'Outstanding'	0.43	张杰 'Zhang Jie' (10.07%); 刘杰 'Liu Jie' (9.19%); 李杰 'Li Jie' (8.83%)
9	洋 'Yang'	'Ocean'	0.43	刘洋 'Liu Yang' (20.52%); 李洋 'Li Yang' (9.40%); 王洋 'Wang Yang' (9.21%)
10	军 'Jun'	'Army'	0.42	王军 'Wang Jun' (9.95%); 李军 'Li Jun'(9.19%); 张军 'Zhang Jun' (9.14%)

Note: For each given name, the three most common full names and their probabilities are listed in the last column.

DOI 10.5195/names.2021.2314 Vol. 69, Issue 3, Summer 2021 ISSN: 0027-7738 (print) 1756-2279 (web)

As shown in Tables 1 and 2, the ten most common surnames account for 49.17% of the total population's surnames; the ten most common given names account for more than 5% of the total population's given names. It indicates the usage imbalance of surnames and given names. Five percent may seem small, however, there are 122,711 given names in our dataset. In addition, from the last column of Table 2, for each given name, more than 25% of people choose the three listed names. The distribution of surnames is shown in Figure 2, where subplots (a), (b), and (c) are on linear, double logarithmic, and single logarithmic scales, respectively.

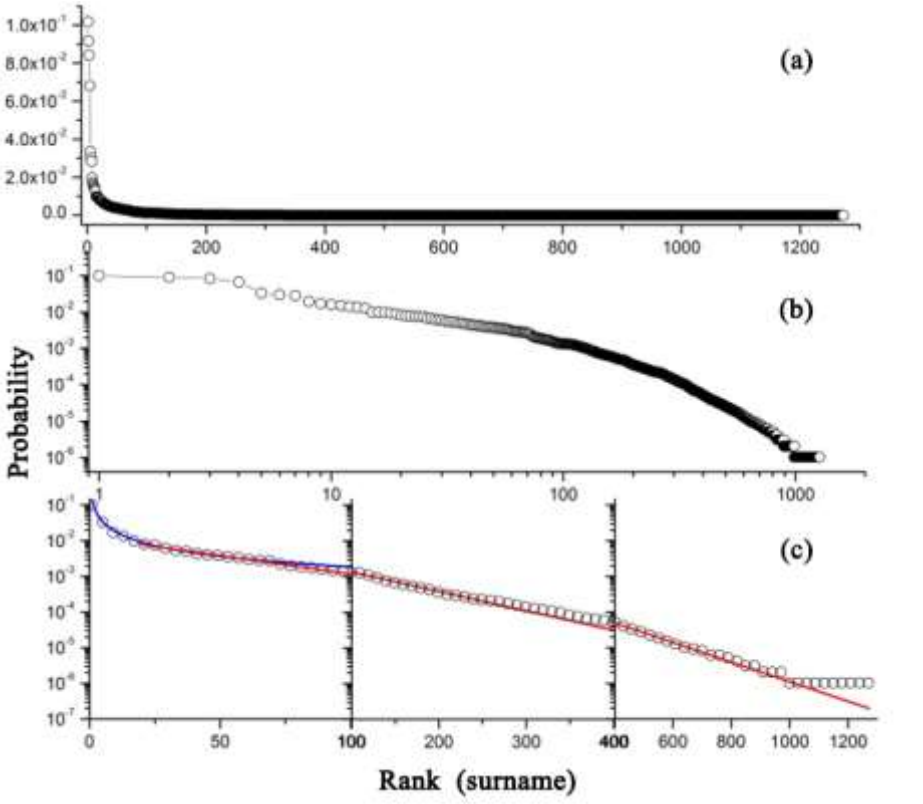

Figure 2. Statistical Distribution of Surnames

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

We found that the distribution is not a linear or power law, but rather a tripartite exponential distribution. The dash line shown in subplot (c) is $p = 0.19ln((r+1)/r)$, as suggested by Yuan (2002). The r indicates the ranking of each surname, and p indicates the probability of each surname. The distributions of given names and full names are presented in Figure 3, and are plotted with double logarithmic coordinates.

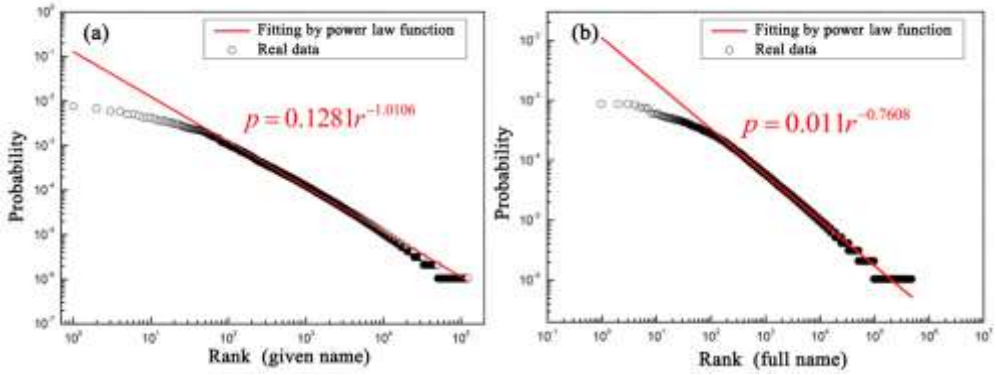

Figure 3. Statistical Distribution of Given Names and Full Names

Unlike the distribution of surnames, the distributions of both the given names and full names reflect power law distributions. In brief, based on distribution graphs, we were able to statistically demonstrate the usage imbalance of surnames, given names, and full names. All three distributions were found to exhibit significant usage imbalance (Gureckis & Goldstone 2009). Based on these findings, we then attempted to quantify the degree of usage imbalance of the surnames, given names, and full names. As shown in Table 3, the ratio of given names is 101.72, which implies that a considerable number of people use the same given name.

Table 3. Measurement of Usage Imbalance

	Surname	Given Name	Full Name
Types	1,262	122,711	478,126
Isonymy (real)	3.70×10^{-2}	8.29×10^{-4}	1.70×10^{-5}
Isonymy (balance)	8.00×10^{-4}	8.15×10^{-6}	2.09×10^{-6}
Ratio	46.25	101.72	8.13

By comparing the ratio, it was possible to demonstrate that the distribution of given names showed the greatest degree of usage imbalance, followed by the distribution of surnames, and the distribution of full names. However, as indicated earlier, the usage imbalance of Chinese full names could not be explained by the simple combination of the usage imbalance of surnames and given names. The difference among the three imbalances is indicative of the complexity of the naming mechanism. In the next sub-section, we discuss the matching imbalance using mathematical modeling and computer simulation.

DOI 10.5195/names.2021.2314 Vol. 69, Issue 3, Summer 2021 ISSN: 0027-7738 (print) 1756-2279 (web)

The Matching Imbalance between Surnames and Given Names

Based on the usage imbalance, our investigation revealed the matching imbalance in the naming process by comparing real names and random-matching names. This comparison is shown in Figure 4.

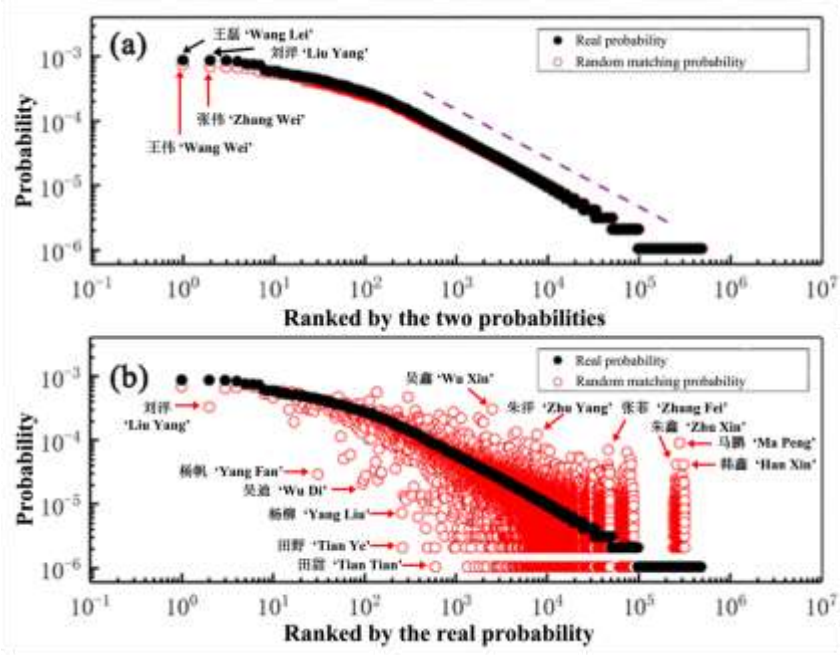

Figure 4. Statistical Distribution of Real Names and One Sample of Random-Matching Names

In Figure 4 (a), the two series are ranked separately. The random-matching names are considered to be the result of choosing a name without preference. From the perspective of the probability of names, there is no difference between random-matching names and real names. However, the probability of each specific name changes significantly.

In Figure 4 (b), the coordinate x denotes the same name, and the series is ranked by real probability. The significant difference between real probability and random-matching probability can be explained by people's preferences for certain names. If more people prefer to choose a specific name, the real probability of that name should be greater than the random-matching probability. The greater the preference, the larger the deviation. For example, more people have a preference for *田甜* 'Tian Tian' and *田野* 'Tian Ye' and fewer people have a preference for *马鹏* 'Ma Peng', *韩鑫* 'Han Xin', and *朱鑫* 'Zhu Xin'. Using the logic presented here, the deviations should therefore be greater for 'Tian Tian' and 'Tian Ye' than for 'Ma Peng', 'Han Xin', and 'Zhu Xin'. Based on this idea, we quantified the matching preference of each name through real probability and random-matching probability.

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

Quantifying the Matching Preference

In this sub-section, we quantify the matching imbalance through MPI. We also present the MPI's of some specific names and provide the associations that might help to explain their preference or avoidance in China. According to the contemporary Chinese dictionary consulted, all of the names in Table 4 have positive meanings.

Table 4. Samples of Preferred Full Names and Their Associations

Name	N_{real}	$\bar{N}_{matching}$	MPI	Associations
田甜 'Tian Tian'	67	1.05	63.91	Pronunciation is the same as 甜甜, which symbolizes 'sweetness and happiness'.
田野 'Tian Ye'	152	2.39	63.61	Means 'field' and symbolizes 'a bright future'.
吴琼 'Wu Qiong'	96	5.64	17.02	Pronunciation is the same as 无穷 'infinite and potential'.
杨柳 'Yang Liu'	152	9.14	16.64	Means 'willow' and symbolizes 'gentle, kind and caring'.
安然 'An Ran'	85	6.19	13.72	Means 'safe' and symbolizes 'peaceful'.
陈曦 'Chen Xi'	272	19.87	13.69	Pronunciation is the same as 晨曦, which means 'first rays of the morning sun' and symbolizes 'bring light, warmth, and happiness'.
吴迪 'Wu Di'	262	19.41	13.50	Pronunciation is the same as 无敌, which symbolizes 'invincible and powerful'.
杨帆 'Yang Fan'	417	32.41	12.86	Pronunciation is the same as 扬帆, which means 'set sail' and symbolizes 'struggle for the ideal'.
汪洋 'Wang Yang'	98	8.22	11.92	Means 'boundless ocean' and symbolizes 'infinite potential'.
刘洋 'Liu Yang'	838	278.26	3.01	Pronunciation is the same as 流洋, which means 'all rivers flow to the sea' and symbolizes 'all things tend in one direction'.

As shown in Table 4, when names are generated by the random-matching process and computer simulation, the frequency of each listed name is significantly lower than it is in reality. First, the MPI of each name in Table 4 is significantly greater than 1, which means that the real probability is significantly greater than the random-matching probability. For example, according to the random-matching process, theoretically, there should be only 1.05 people named *田甜* 'Tian Tian', however, there were 67 people who had this name in reality. That is, more people preferred to choose this name than would have been predicted by the random-matching process. Importantly, the MPI of *田甜* 'Tian Tian' was determined to be 63.91, which is significantly greater than 1. A possible reason for its preferred use is *田甜* 'Tian Tian' and *甜甜* are homonyms in Chinese, and the latter symbolizes 'sweetness and happiness' as indicated in Table 4.

DOI 10.5195/names.2021.2314 Vol. 69, Issue 3, Summer 2021 ISSN: 0027-7738 (print) 1756-2279 (web)

Table 5. Samples of Non-Preferred Full Names and Their Associations

Name	N_{real}	$\bar{N}_{matching}$	MPI	Associations
马鹏 'Ma Peng'	1	83.43	0.012	鹏 is 'a fabulous bird of enormous size', which symbolizes 'ambitiousness'; however, the pronunciation of the full name is the same as 马棚, which means 'horse stable'.
韩鑫 'Han Xin'	1	30.27	0.033	鑫 consists of three 金 'gold', which means 'prospering or good profit'; however, the pronunciation of the full name is the same as 寒心, which means 'bitterly disappointed'.
朱鑫 'Zhu Xin'	1	29.95	0.033	鑫 consists of three 金 'gold', which means 'prospering or good profit'; however, the pronunciation of the full name is the same as 诛心, which symbolizes 'bloody-minded and execrable'.
张菲 'Zhang Fei'	3	79.16	0.038	菲 means 'lush and fragrant'; however, the pronunciation of the full name is the same as 张飞, who is 'a rough man in the history of China'.
吴鑫 'Wu Xin'	2	41.28	0.048	鑫 consists of three 金 'gold', which means prospering or good profit; however, the pronunciation of the full name is the same as 无心, which means 'has no heart'.
朱洋 'Zhu Yang'	2	40.56	0.049	洋 means 'ocean', which symbolizes that the person is 'open and thoughtful'; however, the pronunciation of the full name is the same as 猪羊, which means 'pigs and sheep' and symbolizes 'takes orders from others'.
黄军 'Huang Jun'	3	33.76	0.089	军 means 'army', which symbolizes 'braveness'; however, the pronunciation of the full name is the same as 皇军, which refers to 'Japanese invaders during World War II'.
马峰 'Ma Feng'	3	33.52	0.089	峰 means 'peak', which symbolizes 'steady and outstanding'; however, the pronunciation of the full name is the same as 马蜂, which means 'hornet' and symbolizes 'dangerous'.
马洋 'Ma Yang'	9	70.63	0.127	洋 means 'ocean', which symbolizes 'open and thoughtful'; however, the pronunciation of the full name is the same as 马羊, which means 'horses and sheep' and symbolizes 'take orders from others'.
马伟 'Ma Wei'	18	126.00	0.143	伟 means 'great, robust, brilliant and extraordinary'; however, the pronunciation of the full name is the same as 马尾, which means 'horse tail' and symbolizes 'mediocre'.

When the names shown in Table 5 are generated by a random-matching process and computer simulation, the frequency of each listed name is significantly higher than they are in reality. For example, after the random-matching process, there should be 83.43 people named *马鹏* 'Ma Peng', however, there was only one person named *马鹏* 'Ma Peng' in reality. In other words, fewer people preferred to choose this name. Importantly, the MPI of *马鹏* 'Ma Peng' is 0.012, which is significantly less than 1. Although the character *鹏* 'Peng' means 'roc, a large bird in Chinese legend', which is a good symbol when combined with *马* 'Ma' the result is *马棚* 'stable', which is homonymic in Chinese for 'dirty'. As this and the other examples demonstrate, the MPI is effective for revealing how name preferences influenced homonymy. This finding is in accordance with Gao (2012). As shown in Figure 5, fitting a log-normal distribution model to the values of the MPI indicates that both preferred and non-preferred names are generally supposed to be extremely rare.

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

Figure 6 shows the preferred characteristics of full names plotted with double logarithmic coordinates. We used the horizontal axis to represent the random-matching probability of the full name and the vertical axis to represent the real probability of the full name. Clearly, more common names have a greater real probability. The farther the MPI is away from 1, the larger the size of the point.

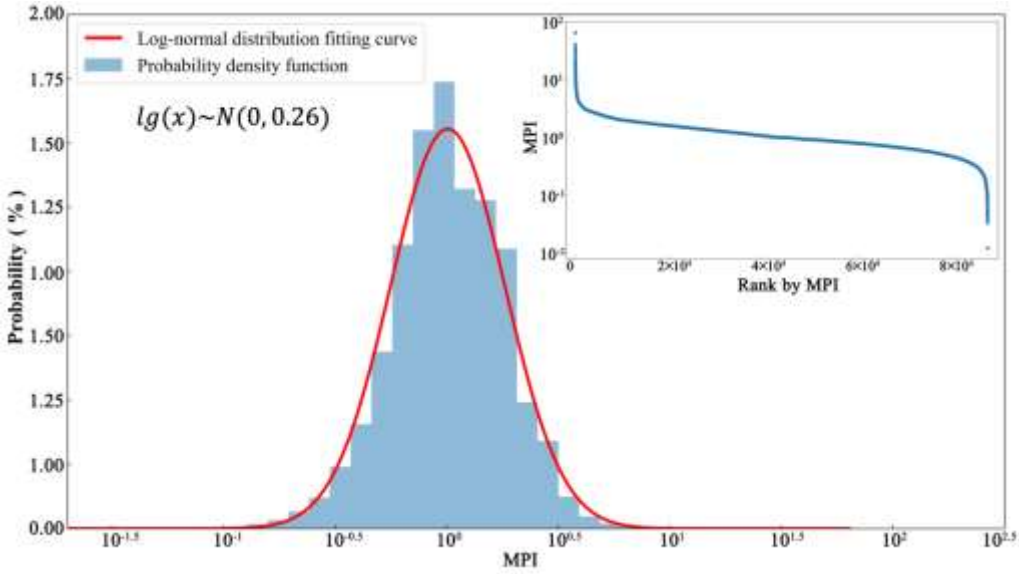

Figure 5. Distribution of the Matching Preference Index (MPI)

According to the definition of MPI, non-preferred names are located below the line MPI = 1, while preferred names are located above the line. As is shown in Figure 6, large points with light gray are located in the upper-right part of the figure, which means some of the uncommon given names are used in common full names when they are combined with particular surnames, such as *田甜* 'Tian Tian' and *田野* 'Tian Ye.' Large dots in dark gray are located in the lower-left part of the figure. This location means that some of the common given names are used in uncommon full names when they are combined with particular surnames (e.g., *马鹏* 'Ma Peng' and *韩鑫* 'Han Xin').

DOI 10.5195/names.2021.2314 Vol. 69, Issue 3, Summer 2021 ISSN: 0027-7738 (print) 1756-2279 (web)

Figure 6 provides a visual presentation of the MPI. As demonstrated in this graphic, both preferred and non-preferred names are effectively distinguished by the MPI.

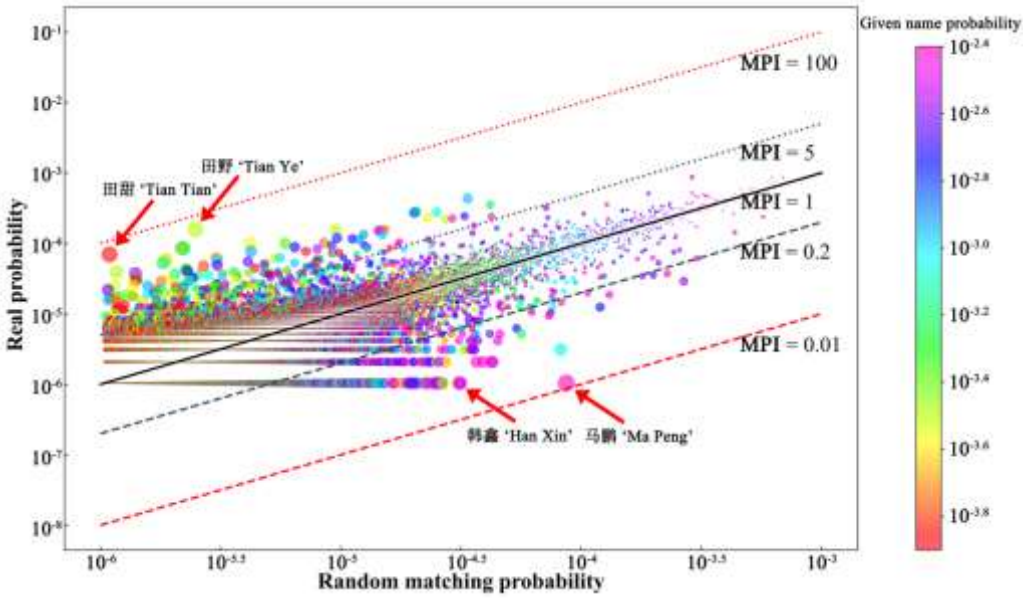

Figure 6. The Relationship Between the Matching Preference Index (MPI) and Given Name Probability

Conclusion

The effectiveness of our method is clearly indicated in the findings presented. Our approach reliably reproduced the imbalances in Beijing citizens' surnames, given names, and full names. Moreover, it empirically replicated the matching imbalance between surnames and given names. This method could be helpful in highlighting potentially overlooked yet culturally important phenomena in Chinese names. However, like any preliminary study, this investigation has some limitations, particularly with regard to its empirical scope and depth. Objectively speaking, the dataset used in this paper is not large enough to perform deeper analyses that are reliable for the entirety of China. Were this method enhanced with the integration of reliable large-scale demographic information, it would be more powerful. Though preliminary, the approach taken here could be extended to other languages and onomastic sets. A few other issues worthy of future exploration include the following:

1.) a comparison of Chinese names taken from different areas or eras;
2.) a comparison of names from different gender and age-sets to reveal contributing factors of matching imbalances of gender and age; and
3.) cross-national comparisons to determine whether (and if so, why) similar cultural phenomena exist in other countries. These issues will be the focus of our future research.

Acknowledgments and Funding Details

We appreciate the comments and helpful suggestions we received from Professors Zengru Di and Yougui Wang. We also wish to express our thanks for the assistance we received from Ziyao Li in obtaining and processing the data used in our investigation. This work was supported by the Humanities and Social Sciences Foundation of the Ministry of Education of China (20YJAZH010), and the National Key R&D Program of China under Grant number 2017YFC0803402.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

Allen, L., V. Brown, L. Dickinson, and K. C. Pratt. 1941. "The Relation of First Name Preferences to Their Frequency in the Culture." *The Journal of Social Psychology* 14, no. 2: 279–293.

Beijing Municipal Commission of Transport. *Beijing Vehicle License Plate Lottery System*. Accessed November 30, 2011. http://www.bjhjyd.gov.cn/.

Bessmum. "Matching First Name to Surname." *Mumsnet*. Accessed August 15, 2008. https://www.mumsnet.com/Talk/baby_names/587499-matching-first-name-to-surname.

Bjname. "Baby Names — An Introduction to the Ranking of 100 Surnames in Beijing (in Chinese)." *Bjname*. Accessed November 5, 2007. http://www.bjname.com/article/2007-11/1473.htm.

Bloothooft, Gerrit, and Loek Groot. 2008. "Name Clustering on the Basis of Parental Preferences." *Names* 56, no. 3: 111–163.

Castellano, Claudio, Santo Fortunato, and Vittorio Loreto. 2009. "Statistical Physics of Social Dynamics." *Reviews of Modern Physics* 81, no. 2: 591–646.

CCC1982. "Matching First Name to Surname." *Baby Name Wizard*. Accessed June 29, 2015. https://www.babynamewizard.com/forum/matching-first-name-to-surname.

Census Office of Fangshan. 2011. "After the Sixth Population Census, the Numbers are Amazing — An Interpretation of the Data of Fangshan District Population Census in Beijing (in Chinese)." *Data* 19, no. 8: 46.

CGTN Society. "Wang Remains the Most Common Surname in China: 2019 Report." *CGTN*. Accessed January 21, 2020. https://news.cgtn.com/news/2020-01-21/Wang-remains-the-most-common-surname-in-China-2019-report-NqJ4JWbpfi/index.html.

Chen, Jie. "Wang is the Most Common Surname in Beijing (in Chinese)." *Taihainet*. Accessed November 4, 2006. http://www.taihainet.com/news/txnews/cnnews/sh/2006-11-04/16345.html.

Colman, Andrew M., David J. Hargreaves, and Wladyslaw Sluckin. 1980. "Psychological Factors Affecting Preferences for First Names." *Names* 28, no. 2: 113–129.

Finch, M., H. Kilgren, and K. C. Pratt. 1944. "The Relation of First Name Preferences to Age of Judges or to Different Although Overlapping Generations." *The Journal of Social Psychology* 20, no. 2: 249–264.

Gao, bo. 2012. "Research on Chinese Naming Method from Sociolinguistic Perspective (in Chinese)." *Science & Technology Information* 34, no.1: 591.

Guo, JinZhong, QingHua Chen, and YouGui Wang. 2011. "Statistical Distribution of Chinese Names." *Chinese Physics B* 20, no. 11: 118901.

Gureckis, Todd M., and Robert L. Goldstone. 2009. "How You Named Your Child: Understanding the Relationship Between Individual Decision Making and Collective Outcomes." *Topics in Cognitive Science* 1, no. 4: 651–674.

Hanks, Patrick, and D. Kenneth Tucker. 2000. "A Diagnostic Database of American Personal Names." *Names* 48, no. 1: 59–69.

Hayakawa, Ryo, Yuta Fukuoka, and Tsuyoshi Mizuguchi. 2012. "Size Frequency Distribution of Japanese Given Names." *Journal of the Physical Society of Japan* 81, no. 9: 094001.

He, Hailun. 2014. "The Art of Naming in China and Translating Western Names into Chinese." *Literary Onomastics Studies* 16, no. 1: 46–50.

Ji, Yan. 2009. "Linguistics Concept about Chinese First Name." *Journal of Anqing Teachers College* 28, no. 7: 125–128.

Joubert, Charles E. 1985. "Sex Differences in Given Name Preferences." *Psychological Reports* 57, no. 1: 49–50.

Kamada, Ryohei, and Tsuyoshi Mizuguchi. 2020. "Heterogeneity of Japanese Names." *Journal of the Physical Society of Japan* 89, no. 7: 074802.

Kim, Beom Jun, and Sung Min Park. 2005. "Distribution of Korean Family Names." *Physica A: Statistical Mechanics and Its Applications* 347: 683–694.

Lele, "Release of the Most Common Surname in Beijing, and the Rank of the Ten Most Common Surnames in Beijing (in Chinese)." *360doc*. Accessed April 7, 2020. http://www.360doc.com/content/20/0407/00/21750475_904318150.shtml.

Li, Na. 2007. "A Linguistic Study of Chinese Naming (in Chinese)." Dissertation, Jinan, Shandong: Shandong Normal University.

Liu, Yan, Liujun Chen, Yida Yuan, and Jiawei Chen. 2012. "A Study of Surnames in China through Isonymy." *American Journal of Physical Anthropology* 148, no. 3: 341–350.

Mateos, Pablo, and Ken Tucker. 2008. "Forenames and Surnames in Spain in 2004." *Names* 56, no. 3: 165–184.

Miyazima, Sasuke, Youngki Lee, Tomomasa Nagamine, and Hiroaki Miyajima. 2000. "Power-Law Distribution of Family Names in Japanese Societies." *Physica A: Statistical Mechanics and Its Applications* 278, no. 1–2: 282–288.

Newman, MEJ. 2005. "Power Laws, Pareto Distributions and Zipf's Law." *Contemporary Physics* 46, no. 5: 323–51.

Nick, I. M. 2013. "A Question of Faith: An Investigation of Suggested Racial Ethnonyms for Enumerating US American Residents of Muslim, Middle Eastern, and/or Arab Descent on the US Census." *Names* 61, no. 1: 8–20.

Nick, I. M. 2017. "Names, Grades, and Metamorphosis: A Small-Scale Socio-Onomastic Investigation into the Effects of Ethnicity and Gender-Marked Personal Names on the Pedagogical Assessments of a Grade School Essay." *Names* 65, no. 3: 129–142.

Online Chinese Dictionary. *Ancient Chinese Surnames and Extant Surnames*. Accessed November 30, 2011. http://xh.5156edu.com/page/18485.html.

Scapoli, Chiara, Elisabetta Mamolini, Alberto Carrieri, Alvaro Rodriguez-Larralde, and Italo Barrai. 2007. "Surnames in Western Europe: A Comparison of the Subcontinental Populations through Isonymy." *Theoretical Population Biology* 71, no. 1: 37–48.

Shi, Yongbin, Le Li, Yougui Wang, Jiawei Chen, Yida Yuan, and Eugene Stanley H. 2019. "Regional surname affinity: A spatial network approach." *American Journal of Physical Anthropology* 168, no. 3: 428–437.

Tucker, D. K. 2001. "Distribution of Forenames, Surnames, and Forename-Surname Pairs in the United States." *Names* 49, no. 2: 69–96.

Tucker, D. K. 2002. "Distribution of Forenames, Surnames, and Forename-Surname Pairs in Canada." *Names* 50, no. 2: 105–132.

Wu, Jie, and Jianchun Yang. 2014. "Zhang, Wang, Li, Zhao Who Have the Most: Surnames Structure and Distribution Characteristics in the 2010 Census (in Chinese)." *China Statistics* 62, no.6: 21–22.

Youxifeng, "Release of the Most Common Surname in Beijing, and the Rank of the Ten Most Common Surnames in Beijing (in Chinese)." *Sohu*. Accessed September 21, 2020. https://www.sohu.com/a/419951994_120743672.

Yuan, Yi Da, and Cheng Zhang. 2002. *Chinese Surnames: Community Heredity and Population Distribution*. Shanghai: East China Normal University Press.

Zanette, Damián H, and Susanna C Manrubia. 2001. "Vertical Transmission of Culture and the Distribution of Family Names." *Physica A: Statistical Mechanics and Its Applications* 295, no. 1–2: 1–8.

Zheng, Xianru. 2011. "Psychological Analysis on Names and Benaming." *Journal of Longyan University* 29, no. 6: 81-85.

Note on the Contributors

Ziming Zhao is an M.A. student at the School of Systems Science at Beijing Normal University, Beijing, People's Republic of China. His research interests include complexity in social-economic systems and statistics.

Xiaomeng Li, Ph.D., is a researcher at the School of Systems Science at Beijing Normal University, Beijing, People's Republic of China. Her research focuses on complexity studies, especially in social and economic systems.

Qinghua Chen, Ph.D., is a professor at the School of Systems Science at Beijing Normal University, Beijing, People's Republic of China. His research focuses on complexity in social-economic systems.

ISSN: 0027-7738 (print) 1756-2279 (web) Vol. 69, Issue 3, Summer 2021 DOI 10.5195/names.2021.2314

Correspondence to: Dr. Qinghua Chen, School of Systems Science, Beijing Normal University, No.19 Xinjiekouwai St., Haidian District, Beijing 100875, China. Email: qinghuachen@bnu.edu.cn.