Formation and Statistical Properties on Name Networks

Lei Zhang[1], Yong Hu[2], Jianxiang Cao[3]

[1,3]School of Computer Science, Communication University of China, Beijing, 100024, China.

[2]BeiXin Kuka (Tianjin) Communication Science and Technology, Galaxy Kuka Group of USA.

[1]zhanglei2012@cuc.edu.cn, [3]jxcao@126.com

Keywords: complex network, small word effect, scale-free, power-law distribution, disassortative mixing.

Abstract. In this paper, we investigate name network based on complex network theory. We study statistical properties of the network and find that name network show some important characteristics: not only small word effect, scale-free and the power-law distribution, but also disassortative mixing. The origin and formation of such properties are investigated from a macroscopic Chinese culture and philosophy perspective. It is interesting to find Chinese culture and philosophy may shape the formation and structure of names. According to the study, we can give some insight of names. Furthermore, understanding name network can help to knowledge Chinese culture. The hub nodes in the network show close relationship with Chinese culture. Learning the hub nodes, which effectively explain the origin of names.

1 Introduction

In recent years, network analysis approach has been successful in different fields including brain structures [1], protein-protein interaction networks [2], social interactions [3], and the Internet and WWW [4], and language networks [5]. Two main features (small world effect [6] and scale free features [7]) seem to be shared by most complex networks, both natural and artificial. Small-world network is often characterized according to two graph measurements of the network: clustering coefficient C and characteristic-shortest path length L [8]. If the clustering coefficient is significantly higher than a value expected for a random network, and the characteristic shortest-path length is lower than a value expected for a regular network, then the network is a small world. Scale free networks are those networks whose connectivity distributions are in a power-law form ($P(k) \sim k^{-r}$) that is independent of the network scale [7, 9]. Different from an exponential network, a scale-free network is in homogeneous in nature: most of the nodes have very few link connections and yet a few nodes have extremely high number of connections.

In Chinese rich traditional culture, the research of name has vital significance. As we know, the study of names have experienced for a long time. Early in the Pre-Qin period, the book “The origin of world” has recorded the pedigree from Emperor to the end of Qin Dynasty. In modern life, everyone has a specific appellation: name, its value can be described by symbol. In ancient society, the development and evolution of names can be described as all-inclusive. From its complex form and the influence of ritual secular, we can know that names are an important part of the Chinese nation Material life and Spiritual life. Furthermore, names have played an important role in political, culture and social activities. After thousands of years of accumulation, the culture of names contains abundant cultural information, national resources, historical accumulation, social semiotics, civilization, ideological implications, Family identity, training results and so on.

Today, scientific and academic have made great advancement, exploring properties of names still has its important research significance. The complex cultural phenomenon of names is not isolated; it is closely related to many disciplines and can be studied from different parts. We can study the number of surname and the characteristics of names from Modern Chinese Character. When people choose Chinese characters form a name, we can also study the psychological domination from Sociolinguistics and Psycholinguistics. In new period, large number of study for names reflect the development and prosperity of academic and culture. The research of names has not simply limited to the name itself, but it's theoretical or combined with other disciplines such as languages, sociology, history and genetics. Furthermore, it is gradually formed a new interdisciplinary name of science.

Names are formed mainly based on Chinese characters. We construct the networks in the following ways: (1) the Chinese characters correspond to nodes of the network, and (2) an directed link exists between Chinese characters if they can form a name. Except for the similar results such as small world and scale free as in [9, 10], some more interesting results including power-law distribution, disassortative mixing, scale-free and Chinese culture related hub nodes are reported. These features play an important role in formation and organization of names.

In this paper we argue that this constructed name network exhibit not only the small-world effect, but also scale-free effect. The paper has not only found the global organization feature of names, but also has given insight about how to knowledge names. The paper is organized as follows. First, we focus on the statistical properties of name network. Second, we draw the conclusions in last section.

2 Name networks construction principles

In Chinese the basic unit of name is Chinese characters. Most names consist of two or three Chinese characters, which are called two-character names or three-character names. In this section, we calculate the statistical properties of the name network. The analysis based on the properties will be given. The results are compared with the completely random networks and random networks with the same degree distribution.

The name network is constructed in the following ways: (1) single Chinese characters are served as nodes of the network; (2) connections exist between two Chinese characters if they can form a name. Let us consider the network of name, $G_L = (V_L, E_L)$, where $V_L = \{w_i\}, (i = 1, \cdots, N_L)$ is the set of N_L nodes and $E_L = \{w_i, w_j\}$ is the set of edges or connections (formed names) between Chinese characters. Here, $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge between Chinese characters w_i and w_j. Two connected Chinese characters are adjacent and the degree of a given Chinese character is the number of edges that connect the given character with other characters.

3 Formation and Statistical Properties on Name Networks

We will investigate properties of the resulting networks. We quantify the structural properties of these networks by their characteristic path length L, clustering coefficient C, and degree distribution $P(k)$. The characteristic path length, L, is the path length averaged over all pairs of nodes. The path length $d(i, j)$ is the number of edges in the shortest path between nodes i and j. The clustering coefficient is a measure of the cliqueness of the local neighborhoods. For a node i with k_i neighbors, then at most $\frac{k_i(k_i-1)}{2}$ edges can exist among them. The clustering coefficient C_i of the node i is defined as $\frac{2e_i}{k_i(k_i-1)}$, where e_i is number of existing links between the k_i neighbors. The clustering coefficient, C, is the average of C_i over all the nodes in the graph. The degree of a vertex in a network is the number of edges incident on (connected to) that vertex. We define $P(k)$ to be the fraction of vertices in the network that has degree k.

Name network which contains 1709 nodes (Chinese characters) and 29440 edges (names)(see Table 1 and Table 2). Our analysis focuses on four properties: sparsity, short path-lengths, degree distributions and assortative coefficient.

	n	m	ra	L	L_{random}
Name Network	1709	29440	2.017%	3.0568	3.5796

Table 1: Statistics properties for the name network.

(n: the number of nodes, m: the number of edges of the network, ra: the ratio between the edge number of the network and the complete graph with the same nodes, L: the average shortest path length, L_{random}: the average shortest path length with random graph of same size and density.)

	C	C_{random}	r	γ_{in}	γ_{out}	γ
Name Network	0.6188	0.0047	-0.1058	3.0910	2.9280	2.7220

Table 2: Statistics properties for the name network.

(C: clustering coefficient, C_{random}: the clustering coefficient for a random graph of same size, r: the assortative coefficient, γ_{in}: power law exponent for the in-degree distribution of the network, γ_{out}: power law exponent for the out-degree distribution of the network, γ: power law exponent for the degree distribution of the network.)

3.1 Sparsity

In order to make the paper clear and easy to understand, name network which contains 1709 nodes and 29440 edges will be mainly discussed. The network has 1709 nodes, and the average short path-lengths L is about 3.0586 (See Table 1). Given the size of the network and the number of connections, it can be observed that the network is sparse: on average, a node is connected to only a very small percentage of other nodes. The total edges of the network, i.e. names are 29440. The total edges of the complete graph with 1709 nodes are $C_{1709}^2 = \frac{1709 \times 1708}{2 \times 1}$. The ratio between them is 2.017%. From the ratio 2.017%, we can conclude that the combination of names is sparse.

3.2 Short Path-Lengths

The network displays very short average path-length, for example, 3.0568 (see Table 1). These short path lengths and small diameter are well-described by random graphs of equivalent size and density, consistent with Watts's findings for their small-world networks [8]. The short path length means it takes about just 3 steps, any information in the net can be reached. If we regard a shortest directed path between two nodes as a meaningful sequence, and the shorter path can express the same meaning as long path, then without doubts, the shorter path can save time, energy and room to store, remember and process information. That agrees with the least effort principle [11]. Finally, if names and Chinese characters in the network can be treated as concepts, short average path implies concepts reveal close relationships among them in compact way which is helpful for information processing.

3.3 Degree Distribution

Figure.1, Figure.2 and Figure.3 display the degree distributions for the Chinese character nodes of each network in log-log coordinates, together with the best-fitting power functions (which appear as straight lines under the log-log scaling).

For the directed network, power functions fit the degree distributions almost perfectly. The exponents γ of the best-fitting power laws (corresponding to the slopes of the lines in Figure.1, Figure.2 and Figure.3) are quite similar in all cases, varying between 2.722 and 3.091 (see Table 1). The high-connectivity Chinese characters at the tail of the power-law distribution can be considered as the "hubs" of the name network.

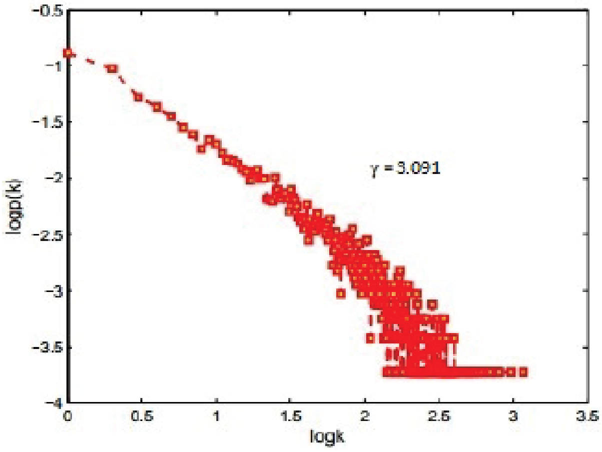

Figure 1: Illustration of log-log in-degree distribution of name network.

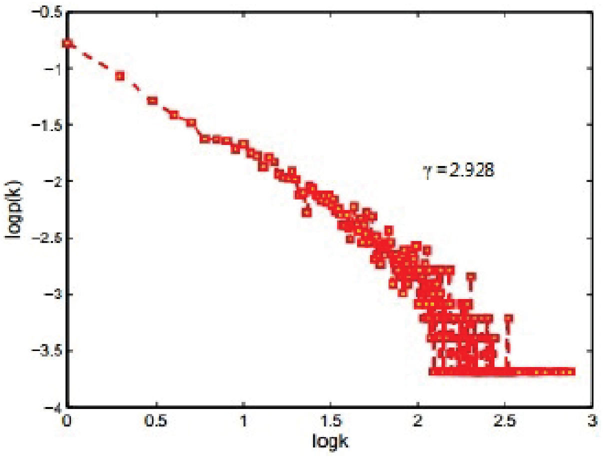

Figure 2: Illustration of log-log out-degree distribution of name network.

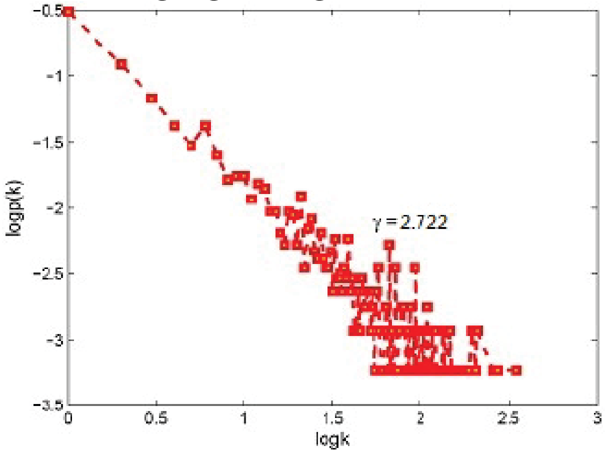

Figure 3: Illustration of log-log degree distribution of name network.

3.4 Assortative coefficient

Assortative mixing is a bias in favor of connections between network nodes with similar characteristics and disassortative mixing is a bias in favor of connections between dissimilar nodes. Of particular interest is the phenomenon of assortative mixing by degree, meaning the tendency of nodes with high degree to connect to others with high degree, and similarly for low degree.

Social networks display specific features that put them apart from biological and technological ones. One of these features is assortative mixing [12, 13, 14, 15, 16]if the networks' assortative coefficient $r \geq 0$. r is defined as

$$r = \frac{M^{-1}\sum_i j_i k_i - [M^{-1}\frac{1}{2}\sum_i (j_i + k_i)]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - [M^{-1}\frac{1}{2}\sum_i (j_i + k_i)]^2} \quad (1)$$

Where j_i, k_i are the degrees of the vertices at the ends of the i th edge, with $i = 1, \cdots, M$, and M is the number of the networks' edges.

The assortative coefficient $r = -0.1058$ displays unlike other social networks, the name network display disassortative feature: higher degree nodes tend to be connected to the low degree nodes.

4 Conclusion

In this paper, we study the special cultural phenomena of names from a new perspective. According to the study, we can know that name network display some important features: not only small-world effect, but also scale-free effect. First, this is because there are some limitations of Chinese characters selection if they can form name, not all Chinese characters can be used for names. Second, people will have certain randomness for the name chosen, but under the influence of the Chinese culture and the concept of social constraints, which people tend to have a kind of psychological merit in the process of selected Chinese characters. And those who can express people's pursuit of the truth, goodness and beauty Chinese characters will become the preferred, this makes the network exists some high degree nodes. The high degree nodes and important edges between hubs should be knowledge first. Because of the role of the hub, it should be helpful to learn other Chinese characters in an efficient, tight and integrated way. Since names strengthen the hubs, the hub-related learning will be very helpful to understand the names better.

We also find that there are a lot of low degree nodes in name network, which is mainly due to the development of society. In order to reduce the probability of rename, people begin to use the remote Chinese characters. This will cause a lot of trouble to individuals and society. At present, the problem of scholar’s common concern is how to standardize name. In the future we will focus on the research in that standardize name. In addition, making the name easy to remember and innovation should be considered.

5 Acknowledgement

The author would like to thank Prof.Li for helpful suggestions and comments. This research work was supported by Engineering Planning Project for Communication University of China (3132014XNG1449 and XNG1436); Guangzhou Research Institute of Communication University of China Common Construction Project "Sunflower"-the Aging Intelligent Community; and The Construction Project of Professional Postgraduate Degree Education Case Library in Communication University of China (ZYKC201448).

References

[1] T. B. Achacoso and W. S. Yamamoto. AY's Neuroanatomy of C. elegans for Computation. CRC Press, Boca Raton, FL, 1992.

[2] H. Jeong, S. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, **411**, pp. 41-42, 2001.

[3] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* , **98**, pp. 404-409, 2001.

[4] X. Li and G. Chen. A local-world evolving network model. *Physica A*, **328(1-2)**, pp. 274-286,

2003.
[5] M. Steyvers and J. B. Tenenbaum. The largescale structure of semantic networks: statistical analyses and a model of semantic growth. *Cogn. Sci*, **29(1)**, pp. 41-78, 2005.
[6] S. H. Strogatz. Exploring Complex Networks. *Nature*, **410**, pp. 268-276, 2001.
[7] R. Albert and A.L. Barabasi. Statistical mechanics of complex networks. *Rev. Modern Phys*, **74**, pp. 47-97, 2002.
[8] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, **393**, pp. 440-442, 1998.
[9] K.Yamamoto and Y.Yamazaki. A network of two-Chinese-character compound words in the Japanese language. *Physica A: Statistical Mechanics and its Applications*, **388(12)**, pp. 2555-2560, 2009.
[10] J. Wang and L. Rong. An Empirical Study on Chinese Word-Word Network Based on Complex Network Theory. *Journal of Dalian Maritime University*, **4**, pp. 15-18, 2008.
[11] R. F. I Cancho and R. V. Sole. Least effort and the origins of scaling in human language. *Procs. Natl. Acad. Sci. USA*, **100**, pp. 788-791, 2003.
[12] M. E. J. Newman. Why social networks are different from other types of networks. *Phys. Rev. E*, **68**, pp. 036122, 2003.
[13] M.E.J. Newman. Assortative Mixing in Networks. *Phys. Rev. Lett*, **89**, pp. 208701, 2002.
[14] M.E.J Newman. Mixing pattern in Networks. *Phys. Rev. E*, **67**, pp. 026126, 2003.
[15] M. Catanzaroa, G. Caldarellib, and L. Pietronero. Social network growth with assortative mixing. *Physica A*, **338**, pp. 119-124, 2004.
[16] A.P. Quayle, A.S. Siddiqui, and S.J.M. Jones. Modeling network growth with assortative mixing. *European Physical Journal B*, **50**, pp. 617-630, 2006.