# An empirical study of Chinese word-word language directed network

Jianwei Wang, Lili Rong
Institute of Systems Engineering
Dalian University of Technology
Dalian, China
E-mail: wdut@yahoo.cn, llrong@dlut.edu.cn

Tao Jin
School of Electronic and Information Engineering
Dalian University of Technology
Dalian, China
E-mail: dlutjintao@hotmail.com

***Abstract*—In this paper, we explore the statistical properties of Chinese language network within the framework of complex network theory. Based on one of the common databases in China, i.e. Tshinghua Unisplendour's one, we construct Chinese word-word language directed network (CWWLDN). The node in CWWLDN represents a single character, and while a directed link between two nodes exists if two different characters appear in the same words. The empirical results show that CWWLDN possesses scale-free behavior and small-world effect. Two characteristics indicate that CWWLDN is similar to other previously studied language networks, which shows that different languages may have some common properties in their evolutionary processes. We believe that our research may shed some new light into the Chinese language evolution and find some potentially significant implications.**

***Keywords-component; complex network; scale free; small world; direct network***

## I. INTRODUCTION

Over the last few years, it has been suggested that a lot of social [1,2], technological [3], biological [4], and information [5,6] networks share the following three striking statistical characteristics: scale-free, high clustering coefficient and small average path length (APL). Scale-free shows that the degrees of nodes on the networks satisfy a power law distribution: $P(k) \sim k^{\gamma}$ , where $k$ is the number of links of a randomly chosen node in the network and $\gamma$ is the scaling exponent. High clustering coefficient indicates that if vertices A and B are linked to vertex C, then A and B are also likely to be linked to each other. Small APL implies that the expected number of edges needed to pass from one arbitrarily selected node to another one is low, that is, APL grows logarithmically with the number of nodes or slower.

In the past few years, the phenomenon of human language is widely studied from various points of view [7-10]. Moreover, with the expansion of complex network research, it is interesting not only for social scientists, anthropologists or philosophers, but also for those, interested in researching on complex networks. Any language is composed of many thousands of words linked together in an apparently fairly sophisticated way. Because words are a good example of simple elements that are combined to form complex structures, such as novels, poems, dictionaries, and manuals that are designed to transport or convey information, human languages have been much investigated from the perspective of complex network. Identifying and understanding the common network topology of languages is of great importance, not only for the study of languages themselves, but also for cognitive science where one of the most fundamental issues concerns associative memory, which is intimately related to the network topology. However, on the one hand, there were very few works about the topology study on the Chinese language, although it is one of the most widely used languages in the world. On the other hand, constructing the networks, the methods were often based on two ways [11-13]: (i) a node represents a word; (ii) a link exists between two words if they express similar concepts or appear simultaneously in a sentence. While the researches on a word as a node of networks and the relation between a word and a word are ignored usually. In fact, to the extent, the people's way of thinking is well discussed by this method.

In this paper, we define a single Chinese character as a word, such as "中", "国", and "人"; while the words which consists of two or more Chinese characters, such as "中国", "人民", and "网络". If we consider word-word relation as our method constructed network and treat this network as a finite directed network in which a single word is a node and two nodes are linked if they are neighbors, then we can analyze this network completely. Inspired these, Chinese word-word language directed network (CWWLDN) is proposed, which is based on the database of Tsinghua Unisplendour , one of the most common databases. We present a comprehensive analysis of the network characteristics including degree distribution, clustering coefficient, and average path length. Our results demonstrate that the degree distribution of CWWLDN follows a power-law distribution, which shows that it exhibits the scale-free property. And the small APL and high clustering coefficient indicates that it possesses the small-world effect. Our study makes it possible to investigate the complexity of the Chinese language within the framework of network theory.

This work was supported by the National Natural Science Foundation of China under Grant nos. 70571011 and 70771016.

978-1-4244-2013-1/08/$25.00 ©2008 IEEE

The rest of this paper is organized as follows: in section 2, we introduce the data source and present network construction method in detail. The statistical characteristics and results are proposed in Section 3. Finally, some summaries and conclusions are shown in Section 4.

## II. CONSTRUCTION OF CHINESE WORD-WORD LANGUAGE DIRECTED NETWORK

For better understanding the formation of Chinese language and analyzing various topology characteristics of it, it is necessary that we start with seeking for a typical database, and then construct Chinese word-word language directed network based on it.

### A. *The Database*

Chinese is one of the most widely used languages in the world. Chinese characters play an important role in its well-known civilization. One reason is that they bear some unique and elegant structures. Most western language characters are phonetic-based, while Chinese characters are mostly picture-based. In order to discuss characteristics of the structure of Chinese characters, in this paper, we make use of the database from Tsinghua Unisplendour, one of the most common databases. The chosen database is representative, since it includes 548,387 words, covering almost all words.

Among the 548,387 words, many words covered by three or more Chinese characters, have a certain degree of data redundancy, because of taking into account the relations between a word and a word, for example, "复杂网络", no relation between "杂" and "网". So we will tackle with the data in the database from Tsinghua Unisplendour in order to keep those words consisting of two Chinese characters.

### B. *Network Construction Method*

According to the principle to extracting words consisting of two characters, the Chinese characters studied in this paper contain 7440 characters and 178,971 words from Tsinghua Unisplendour' database.

The CWWLDN is established as follows:

(i) We define every one among the 7440 characters as a node;

(ii) Two nodes are connected if they appear in the same words.

Thus, the CWWLDN may shed some insight into the understanding of the structure characteristics of words formation in Chinese language. In Fig.1, we present a simple example of 45 words consisting of two Chinese characters and construct a small CWWLDN based on Fig.1.

心脏 中心 中国 国人 人民 民众 发展 展开 开心
发现 展现 开发 凤中 民风 高中 关中 关心 开关
国中 华中 口中 众口 人口 心口 开口 风口 中大
大小 中发 中方 中风 地方 大地 地界 土地 民心
民国 小人 大人 高人 高地 人心 中展 高开 民口

Figure 1. 45 words consisting of two Chinese characters.

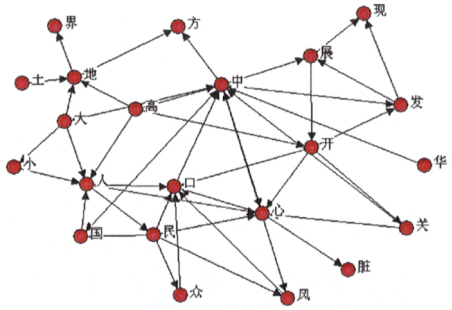


Figure 2. An example of the CWWLDN based on the 45 words shown in Fig.1.

Following the constructional method, we get the CWWLDN, which includes 7440 nodes and 178,971 edges, and contains one giant component of 7125 nodes and 178,770 edges. We also delete the ones unconnected with other words from the CWWLDN, among which including the words constructed by the same two Chinese characters, the number of which is only 87, such as "座座", "尊尊", and "醉醉". In the following, in order to investigate the structure characteristics of the CWWLDN, we will only consider the topology properties of the giant connected component of the CWWLDN ignoring the unconnected units.

## III. EMPIRICAL RESULTS

We focus on the topological characteristics of the CWWLDN, including degree distribution $P(k)$, APL, and clustering coefficient.

### A. *Degree Distribution*

The degree distribution is one of the most important statistical characteristics of a network. By definition, the degree $k_i$ of a node $i$ is the number of edges incident from $i$, and is defined in terms of the adjacency matrix A as:

$$k_i = \sum_{j \in N} a_{ij}$$

If the graph is directed, the degree of the node has two components: the number of outgoing links $k_i^{out}=\Sigma_j a_{ij}$ (referred to as the out-degree of the node), and the number of ingoing links $k_i^{in}=\Sigma_j a_{ji}$ (referred to as the in-degree of the node). The total degree is then defined as $k_i=k_i^{out}+k_i^{in}$.

Degree distribution $P(k)$ is the probability that a randomly selected node has exactly $k$ edges. For many real complex networks, $P(k)$ follows a power-law distribution: $P(k)\sim k^{\gamma}$. The networks with a power-law distribution are called scale free.

In order to understand the topology characteristics of Chinese word-word language directed networks, we investigate two distributions, $P(k^{in})$ and $P(k^{out})$ (see Fig.3 and Fig.4).

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

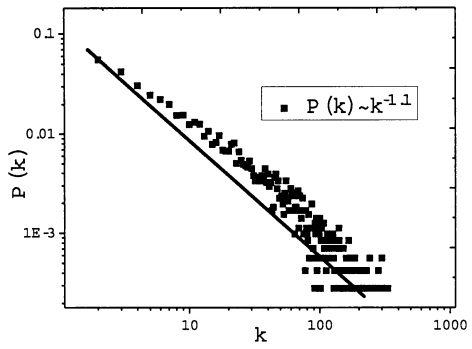


Figure 3. In-degree distribution of the CWWLDN.

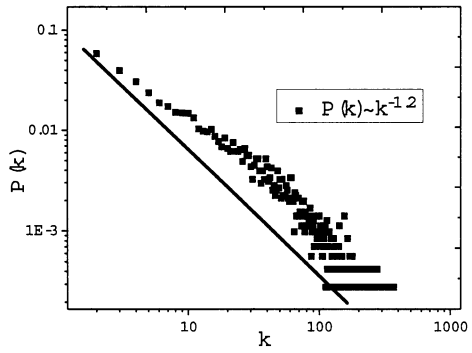


Figure 4. Out-degree distribution of the CWWLDN.

In Fig.3 and Fig.4, we report the in-degree and the out-degree distributions of the CWWLDN, which includes 7125 nodes and 178,770 edges. We get the values of the parameter $\gamma$ of two distributions respectively, which follows a power-law distribution: $P(k)\sim k^{-\gamma}$, based on the slope of the fitting lines, value $\gamma$ is about -1.1 in the in-degree distribution and about -1.2 in the out-degree one. In Fig.5, we further discuss the total degree distribution of the CWWLDN (see Fig.5), where value $\gamma$ is about -1.15.

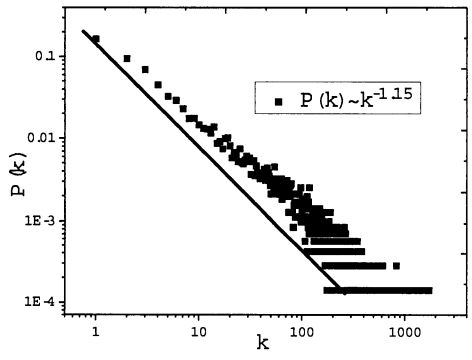


Figure 5. Degree distribution of the CWWLDN.

Based on the empirical results, the considered CWWLDN obviously exhibits a scale-free behavior. Moreover, one can easily see that the degree spectrum is continuous and has a fat tail for large degree values, in agreement with the analytical results to a relatively homogeneous topology similar to most scale-free networks.

As a matter of fact, because of language characteristic and convenient communication, some characters used frequently become the hubs that connect other characters. Frequency of occurrence of a character leads to the emergency of scale-free behavior and the phenomenon of "the rich get richer" of the CWWLDN.

### *B. Clustering Coefficient*

Clustering [14], also known as transitivity, is a typical property of acquaintance networks, where two individuals with a common friend are likely to know each other. Most real-life networks show a cluster structure which can be quantified by the clustering coefficient. The clustering coefficient of a node gives the relation of connections of the neighborhood nodes connected to it. By definition, clustering coefficient $C_i$ of a node $i$ is the ratio of the total number $e_i$ of existing edges between all $k_i$ its nearest neighbors and the number $k_i(k_i-1)/2$ of all possible edges between them, i.e. $C_i=2e_i/k_i(k_i-1)$. The clustering coefficient $C$ of the whole network is the average of all individual $C_i$'s:

$$C =< c >= \frac{1}{N}\sum_{i\in N} C_i .$$

By definition, $0\le C_i\le 1$, and $0\le C\le 1$.

We calculate the clustering coefficient of the CWWLDN and find that the networks have a relatively high clustering coefficient. Compared with the clustering coefficient $C_{rand}=<k>/N=0.003521$ in agreement with the same scale random network, the clustering coefficient $C=0.245169$ of the CWWLDN is about 70 times higher than that the random network.

The high clustering coefficient may be explained by the people who would like to use some familiar Chinese characters to create new words. Thus, the characters frequently used are connected to each other, such as "中" and "心", while the characters less used do not tend to form the clustering, such as "锳" and "锵", no connection with other characters, respectively. From the characteristic of the CWWLDN, we can see that the elements of the new words appear in the common characters, which will lead to the emergency of more clustering nodes.

### *C. Average Path Length*

Shortest paths play an important role both in the transport and communication within a network and in the characterization of the internal structure of the network, which is defined as the length of the geodesic from one node to the other. Average path length (APL) is defined based on the concept of shortest path, which is the average value of shortest path length of all node pairs in the networks. APL is one of the most important properties in measuring the efficiency of

communication networks. In a store-forward computer network, for example, the most useful measure characterizing the performance is the transmission delay in sending a message through the network from the source to the destination. The transmission delay is approximately proportional to the number of edges that a message must pass through. Therefore, APL plays a significant role in measuring the transmission delay. Moreover, it is well known that the most important property of a small-world network is a logarithmic APL with the number of nodes. Hence, its study has attracted much attention.

Since the original conception of small-world effect is defined based on undirected networks, hereinafter we only consider the undirected version of CWWLDN. If we use the value of $d_{ij}$ to represent shortest path length from node $i$ to node $j$, mathematical expression of APL, also known as characteristic path length, is defined as [15]:

$$L = \frac{1}{N(N-1)} \sum_{i,j \in N, i \neq j} d_{ij},$$

where $N$ represents the total number of all the nodes in the networks.

We get the value APL of the CWWLDN in the case of ignoring the direction of edges, which is about 2.73. Compared with the number $N=7125$ of nodes in the CWWLDN, 2.73 is very small. It is obvious that random chosen two characters from the CWWLDN exist in the correlations from each other by a few steps. The reason why the APL is so small may be related to the existence of hubs, which are bridges between different nodes of the network that would otherwise be separated by many links.

In the last few years, a series of empirical studies report that many real complex networks possess small-world property including biological and technological ones, which is the unifying concept constituting our basic understanding of the organization of real-life complex systems.

Small-world effect is measured by clustering coefficient and APL. The CWWDLN is considered as the small-world network, since which possesses high clustering coefficient and small APL. The emergency of this phenomenon can be better explained by the evolution of Chinese language network. As the network grows, since the existing some nodes that have more connections have opportunity to attach to the new node according to human's cognitive principle, these nodes become the "bridge" ones naturally, i.e., the "shortcuts" ones, which drastically reduces the APL, leading to a small-world behavior.

## IV. CONCLUSION

We have presented a comprehensive investigation on the statistical properties of the CWWLDN within the framework of complex network theory. Like other many complex networks from nature and society, the CWWLDN also possesses to show similar statistical features: scale-free behavior and small-world effect. We try to explain the empirical results from linguistics and the method to using characters, which may be useful to study the dynamics of Chinese language networks.

Although the results reported in this paper represent only the starting point towards the understanding of Chinese language networks, it could be relevant for a more realistic modeling of the Chinese language networks, and could find other implications.

Further studies planned to elucidate remaining issues which can also be seen as pointing to the potential problems and limitations of the current approach, are:

(1) There exist other methods to the construction of Chinese language networks, such as a syntactic network;

(2) We may further discuss the dynamics of the words' evolution and the effect of weights of edges in the CWWLDN. In summary, for better understanding of the language science, it is necessary that we explore in depth the correlations among the characters.

## ACKNOWLEDGMENT

Jianwei Wang would like to thank the database provided by Tsinghua Unisplendour.

## REFERENCES

[1] M.E.J. Newman, "Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality", Phys. Rev. E 64, 2001, 016132.

[2] Newman M E J, "Scientific collaboration networks. I. Network construction and fundamental results", Phys. Rev. E 64, 2001, 016131.

[3] Xu T, Chen R, He Y, et al., "Complex network properties of Chinese power grid", International Journal of Modern Physics B 18, 2004, pp. 2599-2603.

[4] Jeong H, Mason S, Barabási A-L. et al., "The largescale organization of metabolic networks", Nature 407, 2000, pp. 651-654.

[5] Albert R, Jeong H, Barabási A-L, "Diameter of the World Wide Web", Nature 401, 1999, pp. 130-131.

[6] Vázquez A, Pastor-Satorras R, Vespignani A, "Large-scale topological and dynamical properties of the Internet", Phys. Rev. E 65, 2002, 066130.

[7] R.F. Cancho, R. Solé, "The small world of human language", Proc. R. Soc. London B 268, 2001, pp. 2261-2265.

[8] A. E. Motter, P. S. de Moura, Lai Ying-Cheng, P. Dasgupta, "Topology of the conceptual network of language", Phys. Rev. E 65, 2002, 065102.

[9] M. Steyvers, J. B. Tenenbaum, "The large scale structure of semantic networks: statistical analyses and a model of semantic growth", Cogn. Sci. 29, 2005, 41.

[10] M. A. Nowak, D. C. Krakauer, "The evolution of language", Proc. Natl. Acad. Sci. USA 96, 1999, 8028.

[11] Jianyu Li, Jie Zhou, "Chinese character structure analysis based on complex networks", Physica A 380, 2007, pp. 629-638.

[12] Shuigeng Zhou, Guobiao Hu, Zhongzhi Zhang, Jihong Guan, "An empirical study of Chinese language networks", Physica A 387(2008) 3039-3047.

[13] Haitao Liu, "The complexity of Chinese syntactic dependency networks", Physica A 387, 2008, pp. 3048-3058.

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

[15] Watts D J, Strogatz S H, "Collective dynamics of 'small-world networks", Nature 393 (6684), 1998, pp. 440-442.