

# Structural Equivalence Between Co-occurrences of Characters and Words in the Chinese Language

Yuming Shi<sup>\*†</sup>, Wei Liang<sup>\*†</sup>, Jing Liu<sup>#</sup> and Chi K. Tse<sup>#</sup>

<sup>\*</sup>Department of Mathematics, Shandong University, Jinan, Shandong 250100, China

<sup>#</sup>Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China

<sup>†</sup>Also with Hong Kong Polytechnic University as visiting researchers during early development of this work.

Email: ymshi@sdu.edu.cn, encktse@polyu.edu.hk

**Abstract**—Complex networks are constructed for studying the co-occurrence of characters and words in the Chinese language. Two types of networks are investigated. In the first type, nodes correspond to Chinese characters, and in the second type, nodes correspond to Chinese words. Moreover, edges correspond to connections of characters and/or words that occur consecutively. Networks are built from a collection of Chinese texts of four different styles, namely, essays, novels, popular science articles, and news reports. Their statistical properties are studied in terms of some complex network parameters, including average degree, diameter, average path length, clustering coefficient, degree distribution, as well as connected subnetworks. It is found that although these two kinds of networks have different parameter values, they display qualitatively similar properties, such as exhibition of small-world and scale-free features. This qualitative equivalence between the network of Chinese characters and the network of Chinese words provides a valid basis on which either types of networks can be used for comparing different languages regardless of the incompatibility of the linguistic roles that words play in the Chinese language and in other languages.

## 1. Introduction

The study of complex networks has aroused much interest among researchers across a variety of scientific disciplines in recent years. Complex networks provide a powerful alternative approach for studying some problems in natural science, social science, and engineering. It has been found that many real-world networks possess small-world and scale-free features [1]–[3]. Recently, language networks were studied in terms of complex network theory at different levels such as co-occurrence, syntactic dependency, and semantic dependency [4]–[5]. Networks constructed from text samples at different levels reveal different characteristics of the language.

The Chinese language is spoken by the largest population in the world, and is one of the important languages. There are a number of differences between the Chinese language and other languages. In most languages, words are phonetic-based, but the Chinese language uses mostly picture-based characters. More than 85,000 Chinese characters

have been created, of which about 5000 are in common use. Recently, the Chinese language was studied from a complex network perspective, with emphasis on co-occurrence of characters, syntactic structure, and semantic relation of words [6]–[8]. The networks constructed for such studies exhibit small-world and scale-free features.

Apart from the above-mentioned unique picture-based characters of the Chinese language, the formation of words is also noticeably different. Most words in the English language have a meaning, but *in the Chinese language, a group of characters forms a Chinese word that has a semantic meaning in general*. For example, 中 and 国 are characters, and together the word 中国 means China. So, networks constructed by words from other languages are incompatible with those constructed from Chinese characters due to the fundamental difference in the role that a word plays in different languages. In this paper, two types of networks are considered. In the first type, nodes are defined by Chinese characters, and in the second type, nodes are defined by Chinese words. Moreover, edges are defined by co-occurrence of characters or words within a sentence. We construct networks from a collection of Chinese essays, novels, popular science articles, and news reports. Their basic network properties are studied, including average degree, diameter, average (shortest) path length, clustering coefficient, degree distribution and connected components. We will show that the two kinds of networks are qualitatively equivalent. This finding is important for future comparative studies of the Chinese language with other languages as it permits comparison of different languages in terms of networks constructed based on either characters or words.

## 2. Some basic concepts of complex network theory

A network is a set of nodes connected via edges (links). Average degree, average path length, diameter, clustering coefficient, and degree distribution are important statistical parameters of networks. It is interesting to note that many real-world networks possess either a small-world effect or a scale-free degree distribution, or both. For self-containedness of this paper, we will review these concepts briefly in this section.

Suppose that a network has  $N$  nodes and  $E$  edges. The degree of node  $i$  is the number of edges that the node has, denoted by  $k_i$ ; that is, node  $i$  has  $k_i$  connected neighbors. The average degree of the network is defined by  $K = (\sum_{i=1}^N k_i)/N$ . The clustering coefficient of node  $i$  is defined by  $C_i = 2E_i/k_i(k_i - 1)$ , where  $E_i$  is the number of the actual edges among the neighbors of node  $i$ . Thus, the clustering coefficient of the network is defined by  $C = (\sum_{i=1}^N C_i)/N$ . A network is called connected if for any two nodes in the network, there is at least a path connecting these two nodes. Given two nodes  $i$  and  $j$ , let  $d(i, j)$  be the minimum path length that connects these two nodes. Now further suppose that the network is connected, the average path length of the network is defined by  $d = 2(\sum_{i>j} d(i, j))/N(N - 1)$ , and the diameter of the network is defined by  $D = \max_{1 \leq i, j \leq N} d(i, j)$ .

If a network has a small average path length  $d$  and a large clustering coefficient  $C$ , then it is said to possess the small-world effect; in detail,  $d \sim d_r$  and  $C \gg C_r$ , while  $d_r$  and  $C_r$  are the average path length and the clustering coefficient of the corresponding random network with the same  $K$ .

Another important statistical concept for network is the *degree distribution*. The degree distribution of a network is defined by a probability function  $p(k)$ , which is the probability of a random-picked node that has degree  $k$ . A network is said to be a scale-free network if its degree distribution  $p(k)$  is a power-law distribution; that is,  $p(k) \sim k^{-\gamma}$ , where  $\gamma$  is a positive constant and is called the exponent of the power-law distribution.

### 3. Constructing networks based on co-occurrence

In this section two types of Chinese language networks are constructed. With nodes representing characters, we construct the *character networks*, or simply C-networks. With nodes denoting words, moreover, we have the *word networks*, or simply W-networks. These networks are built from 53 articles of four different styles: essays, novels, popular science articles, and news reports. All these articles are written by native speakers. The articles of the first two styles are written by some famous Chinese writers and are widely known. Since these different texts have differing functional purposes, their lengths vary to some extent. For instance, a novel may be very long, while a popular science article and a news report may be very short. These extreme situations are avoided in our choice. The collection of essays consists of 15 articles with length ranging from 891 to 6739, and an average length of 2747. Here, the length of an article is the number of all characters in the article, regardless of any repetition of characters. The collection of novels consists of 8 articles with length from 3731 to 13062, and an average length of 6640. The collection of popular science articles consists of 15 articles with length from 857 to 2937, and an average length of 1730. The collection of news reports consists of 16 articles with length from 584 to 3368, and an average length of 1276.

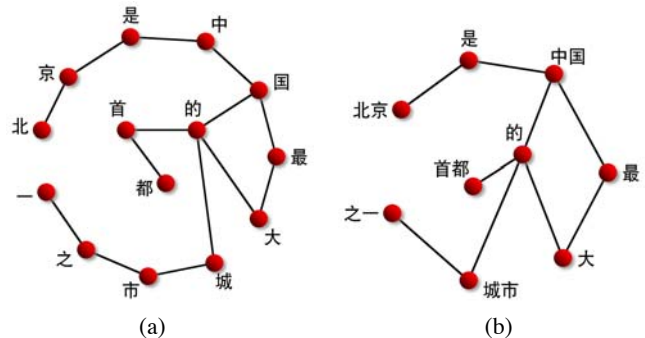


Figure 1: Co-occurrence networks of the sample Chinese sentence with nodes being (a) characters; and (b) words.

Table 1: Network parameters found from Chinese texts. The 3-letter acronym (e.g., ECW) represents the text style, network type and coverage. First letter: E for essay, N for novels, P for popular science articles, and R for news reports. Second letter: C for character network, and W for word network. Third letter: W for whole article and C for largest connected part of the article.

	$L$	$N$	$E$	$K$	$D$	$d/d_r$	$C/C_r$ (%)/(%)	$\gamma$
ECW	2747	622	1599	4.7	-	-	8.95/0.85	2.13
ECC		615	1556	4.8	9	3.53/4.10	9.05/0.87	2.14
EWV		703	1190	3.3	-	-	6.48/0.58	2.79
EWV		670	1180	3.4	12	3.87/5.27	6.67/0.62	2.74
NCW	6640	997	3331	6.5	-	-	11.93/0.65	1.90
NCC		992	3330	6.5	9	3.29/3.68	12.00/0.68	1.90
NWV		1318	2688	4.1	-	-	10.16/0.31	2.55
NWC		1265	2670	4.2	11	3.60/4.97	10.51/0.37	2.54
PCW	1730	387	921	4.7	-	-	7.72/1.66	1.90
PCC		383	919	4.7	10	3.61/3.84	7.81/1.30	1.89
PWV		394	683	3.4	-	-	6.81/0.95	2.38
PWC		379	679	3.5	12	3.93/4.71	7/06/1.04	2.38
RCW	1276	385	748	3.7	-	-	4.05/0.98	2.22
RLC		380	745	3.8	12	4.20/4.46	4.11/1.06	2.21
RWV		349	510	2.8	-	-	3.75/0.81	2.74
RWC		329	502	2.9	14	4.69/5.36	3.95/1.00	2.73

Characters (words) can interact in many ways. In this paper, we focus on co-occurrence of characters (words). Specifically, two characters (words) occurring consecutively within a sentence are connected by an edge. Thus, when characters are used as nodes, we obtain the character network, and when words are used as nodes, we get the word network. For simplicity, the directions of the connections are not considered. So, the networks studied here are undirected. For example, consider the following Chinese sentence:

北京是中国的首都，是中国最大的城市之一。

The corresponding C-network and W-network are shown in Fig. 1.

### 4. Empirical results with analysis

In order to make a qualitative comparison of the statistical properties of C-networks and W-networks, we compute

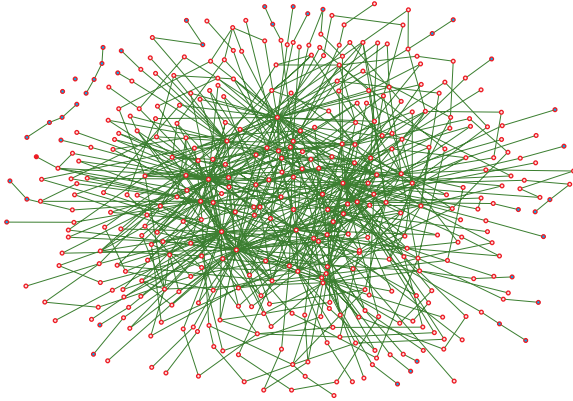


Figure 2: Character network from an essay with 393 nodes and 783 edges. This is represented by ECW in Table 1.

several important parameters, including average degree, diameter, average path length, clustering coefficient, degree distribution and the number of connected components (subnetworks), for the two kinds of networks constructed from the 53 selected articles and the combined (concatenated) article of each style. For simplicity and due to limited space, we present only their average values for the four text styles in Table 1. For each text style (i.e., essays (E), novels (N), popular science (P) or news reports (R)), we build both C-networks and W-networks, and consider the whole article (W) as well as the largest single connected part of each article (C).

#### 4.1. Connected subnetworks

Based on our empirical data, we observe that most of the C-networks and the W-networks are disconnected. For any of the networks, the number of nodes of the largest connected subnetwork is almost the same as that of the whole network, and the numbers of nodes of the other connected subnetworks are very small and less than 10, most of which only contain 1 node, 2 nodes, and 3 nodes. This result implies that the statistical network parameters of the whole network is the same as those of its largest connected subnetwork (see Table 1). A typical C-network for essay is shown in Fig. 2, where its connected subnetworks can be clearly seen.

For each of the 53 articles and the 4 concatenated articles, the number of connected subnetworks of its corresponding C-network is less than that of its corresponding W-network. This means that the connectivity of the C-network is better than that of the W-network.

#### 4.2. Scale-free degree distribution

For all the selected articles, both of the corresponding C-networks and the W-networks exhibit a power-law degree distribution. The power-law exponents,  $\gamma$ , have been found to fall in the range of 1.53 to 2.59 for the C-networks and in the range of 1.96 to 3.22 for the W-networks. The values

of  $\gamma$  have been computed using the least-square-error estimation. In addition, for the C-networks and W-networks from the concatenated articles, the power-law exponents  $\gamma$  are 1.33, 1.29, 1.46, 1.51, and 2.34, 2.20, 1.92, 2.12, respectively. They are smaller than those of the C-networks and W-networks constructed from the individual articles of each style. Furthermore, the power-law exponent of each C-network is less than that of its corresponding W-network.

Based on the above results, for each of the 53 articles and the 4 concatenated articles, *both the C-network and the W-network clearly exhibit a scale-free degree distribution*. Figure 4 shows the degree distributions plotted in a log-log scale for typical C-networks and W-networks from the four text styles. A scale-free degree distribution is characterized by the majority of nodes in the networks having a few connections to other nodes, and only a few nodes (hubs) having connections with many other nodes. These hubs are often some functional words and some words that are closely related to the topics of the articles. This scale-free feature shows that the Chinese language is a self-organizing system like other languages and many real-world networks.

#### 4.3. Small-world effect

By the definition given in Section 2, small-world effect is mainly characterized by average path length and clustering coefficient. Since this effect is restricted to connected networks, the average path lengths and the clustering coefficients are considered only for the largest subnetworks of all the networks discussed in this subsection although they are nearly the same as those of the whole networks, as discussed in Section 4.1.

For all the 53 articles, both the C-networks and the W-networks have small average path lengths  $d$ , ranging between 3.07 and 4.98 for the C-networks and between 3.23 and 5.55 for the W-networks, while the average path lengths of the corresponding random networks  $d_r$  lie between 3.44 and 5.07 for the C-networks and between 4.18 and 6.36 for the W-networks. In addition, for the C-networks and W-networks from the 4 concatenated articles, the average path lengths  $d$  are 2.89, 2.81, 2.88, 3.06 and 3.33, 3.25, 3.24, 3.72, respectively. They are smaller than those of the C-networks and W-networks from the individual articles of each text style in general. Furthermore, we observe that  $d < d_r$  and  $d \sim d_r$  for all the networks. A small average path length of a network implies that the distances between nodes in the network are short. This phenomenon is resulted from the existence of hubs, which play a bridging role between nodes of the networks.

For all the selected articles, both of the C-networks and the W-networks have small clustering coefficients  $C$ , ranging from 0.0195 to 0.1732 for the C-networks and from 0.0177 to 0.1855 for the W-networks, while the clustering coefficients of the corresponding random networks  $C_r$  range from 0.0050 to 0.0195 for the C-networks, and from 0.0018 to 0.0177 for the W-networks. In addition, for

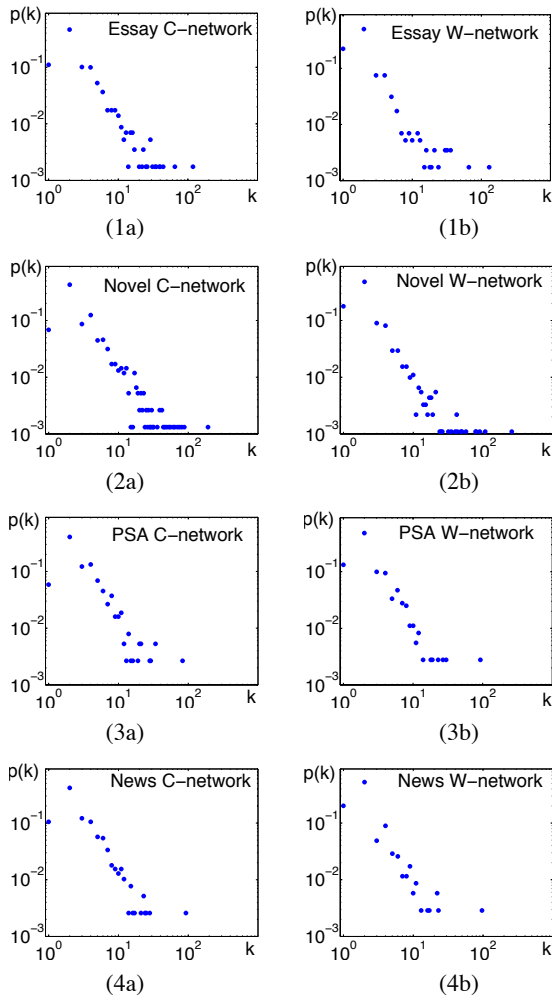


Figure 3: Degree distributions for C-networks and W-networks of articles of different styles.  $p(k)$  versus  $k$  in log-log scale. They are fitted by power law with exponents 2.34 (1a), 2.83 (1b), 2.03 (2a), 2.56 (2b), 1.83 (3a), 2.21 (3b), 2.01 (4a), and 2.62 (4b).

the C-networks and W-networks from the 4 concatenated articles, the clustering coefficients  $C$  are 0.2133, 0.2440, 0.1811, 0.1256 and 0.1555, 0.1902, 0.1564, 0.0822, respectively. They are larger than those of the C-networks and W-networks from the individual articles of each text style in general. We also observe that  $C \gg C_r$  for all networks.

Our empirical data also shows that the number of nodes in every C-network is less than that of its corresponding W-network, and that the number of edges in almost every C-network is larger than that of its corresponding W-network. Therefore, the average path length of a C-network is less than that of its corresponding W-network, and the clustering coefficient of the C-network is larger than that of the W-network in general.

From the above results, both the C-networks and the W-networks clearly manifest the small-world effect for all the 53 articles and the concatenated articles. It is also found

that these networks have a small diameter with average around 10. These results are consistent with the function of human languages as a rapid and accurate means for communication and transmission of information among members of a community.

## 5. Conclusions

Based on a large collection of selected Chinese articles, we compare *character networks*, which are formed by connecting co-occurring characters, with *word networks*, which are formed by connecting co-occurring words. It has been found that both the character networks and the word networks exhibit scale-free degree distribution and small-world effect. Thus, despite the difference in the actual parameter values, the two types of networks are qualitatively equivalent. This qualitative equivalence will facilitate the establishment of a common platform for comparing different languages, especially where the linguistic role of words in one language differs from that of the others. Furthermore, our study shows that the statistical parameters of the networks constructed from one text style are different from those of the networks from another style. These differences may help characterize different kinds of writing styles or text styles from a complex network perspective.

## Acknowledgments

This research was supported by Hong Kong Polytechnic University Research (Grant 1-BBZA) and the NNSF of Shandong Province (Grant Y2006A15).

## References

- [1] A. L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.
- [2] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small world' networks," *Nature*, vol. 393, pp. 440–442, 1998.
- [3] M. E. J. Newman and D. J. Watts, "Renormalization group analysis of the small-world network model," *Phys. Lett. A*, vol. 263, pp. 341–346, 1999.
- [4] R. F. Cancho and R. V. Solé, "The small world of human language," *Proc. R. Soc. Lond. B*, vol. 268, pp. 2261–2265, 2001.
- [5] A. E. Moter *et al.*, "Topology of the conceptual network of language," *Phys. Rev. E*, vol. 65, 065102, 2002.
- [6] Z. Liu and M. Sun, "Chinese word co-occurrence network: its small word effect and scale-free property," *J. Chinese Information Processing*, vol. 6, pp. 52–58, 2007.
- [7] S. Zhou, G. Hu, Z. Zhang, and J. Guan, "An empirical study of Chinese language networks," *Physica A*, vol. 387, pp. 3039–3047, 2008.
- [8] H. Liu, "The complexity of Chinese syntactic dependency networks," *Physica A*, vol. 387, pp. 3048–3058, 2008.