Available online at www.sciencedirect.com

ScienceDirect

Physica A 380 (2007) 629–638

PHYSICA A

www.elsevier.com/locate/physa

Chinese character structure analysis based on complex networks

Jianyu Li, Jie Zhou*

Department of Automation, Tsinghua University, Beijing 100084, China

Received 6 December 2006; received in revised form 10 February 2007
Available online 4 March 2007

Abstract

In this paper, Chinese character networks are modelled using complex networks theory. We analyze statistical properties of the networks and find that character networks also display two important features as other real networks, i.e., small-world feature and the non-Poisson distribution. These results indicate that the discovered features of Chinese character structure reflect the combinatorial nature of Chinese characters. We also simulate the formation of Chinese phono-semantic characters using bipartite graph theory. The bipartite graph model generates non-Poisson distributions and disassortative mixing as the empirical networks, which effectively explain the origin and formation of phono-semantic characters.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Complex networks; Bipartite graph; (dis)Assortative mixing

1. Introduction

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. Chinese characters are ideograms and hieroglyph. The language allows the construction of a virtually infinite range of combinations naturally from a limited set of basic units: semantic and phonetic radicals. Chinese character structure analysis based on radicals is a challenging, interesting, and very important problem. It will help to study fundamental unsolved puzzles in cognitive science, in particular, the origins, the evolution of language and the universal rule of Chinese character combination principles.

Recently, many important results on complex networks have been reported since the appearance of seminal papers by Watts and Strogatz [1,2] on the small-world characteristic and by Barabási and Albert [3–5] on scale-free features. Small-world network is often characterized according to two graph measurements of the network: clustering coefficient C and characteristic shortest-path length L [1]. 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 ($P(k)\sim k^{-\gamma}$) that is independent of the network scale [3–5]. Different from an exponential network, a scale-free network is inhomogeneous in

*Corresponding author. Tel.: +86 10 6278 2447; fax: +86 10 6278 6911.
E-mail addresses: lijianyu@tsinghua.edu.cn (J. Li), jzhou@tsinghua.edu.cn (J. Zhou).

0378-4371/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2007.02.059

nature: most of the nodes have very few link connections and yet a few nodes have extremely high number of connections.

Research results show that complex networks are pervasive in our real world, and many real complex networks share the above stated underlying organizing properties and universal patterns [2,5–8]. For example, brain structures [9], protein–protein interaction networks [10], social interactions [11–14], the Internet and WWW [15–17], and language networks [18–24,30].

Languages are clearly complex adaptive systems. They can be modelled by complex networks at all levels: phonetic, lexical, syntactic, and semantic. Work based on complex networks has aroused more attention in the field of language modelling. Language networks mainly include five networks: thesaurus networks [5,8,20,21,25,26], WordNet [20], word association networks [20,27], word co-occurrence networks [18,28,29] and syntactic dependency networks [31,32]. Sole [24] reviewed some early efforts to build up language networks about western languages, to characterize their properties, and to show in which direction models are being developed to explain them. Changizi and Shimojo [33] studied character complexity of non-logographic writing systems from five major taxa: Ancient Near-Eastern, European, Middle Eastern, South Asian, Southeast Asian and they found that some commonalities of character length and redundancy might help to explain how subcharacter-level parts combine to form characters. Logographic-based Chinese characters have its combination principles different from the non-logographic writing systems. Li [34] studied the properties of Chinese phrases using small-world networks theory. But in the area of Chinese character networks no results are yet reported. Chinese characters can also be regarded as a network in the following sense: (1) the radicals correspond to nodes of the network; and (2) a link exists between two radicals if they can form a character or a part of it; (3) components (radicals) build up new characters, which in turn act as components in more complex characters. Therefore the complex networks of Chinese radicals come into existence and the features of Chinese characters are explored based on the networks. Just like the word co-occurrence networks, our Chinese character networks are radical co-occurrence, i.e., two radicals are linked if they appear together within at least one character. In this paper, we study the network structure of Chinese characters. We construct a radical network with the commonly used and simplified characters of Modern Chinese. We argue that this network exhibits the small-world property and special property different from scale-free networks. We believe and shall argue that these findings are important not only for linguistics, but also for cognitive science because of the semantic and phonetic structure.

The paper is organized as follows. Section 2 focuses on finding simplified Chinese character construction principles. In Section 3, results of empirical measurements are presented. Section 4 introduces a new model to simulate Chinese characters' formation. In Section 5, the new model is compared to the empirical networks. Finally, Section 6 ends the paper with discussion and conclusions.

2. Chinese character construction principles

Chinese has many hundreds of thousands of words, most of which are created by combining just a few thousand characters or radicals. In the terms of their composition, there are six types of characters: pictographic, indicatives, ideographs, phonetic compounds, mutual explanatory, and phonetic loans. Strictly speaking, only the first four refer to the ways of composing Chinese characters, the last two are concerned with the ways to use them. The first two types are single-body, meaning that the character was created independently from other Chinese characters. More productive for the Chinese script were the next two types, i.e., the character was created from assembling different characters. The final two types are rarer.

Although more than 85,000 Chinese characters have been created in history, the characters in common use are about 5000. The Chinese characters studied in this paper are from a dictionary (the 10th edition of *xīnhuācídiǎn*) which contains 7000 characters, among which 6652 characters can be decomposed by our rules.

According to the characters structure, there are about 214 traditional radicals, which are mainly used to index and look up characters in the dictionary. In many cases, the decomposed radicals do not exist in the traditional radicals. It will be incomplete and unreasonable to study the characters structure without such radicals (characters), so the radicals must be expanded and other new radicals should be added. For instance, the phono-semantic characters “萌”, “蹴”, and “湖” should be decomposed into “艹” and “明”, “足” and “就” and “氵” and “胡” according to their phono-semantic structures and cannot be decomposed further. In fact, the characters “明”, “就” and “胡” are not traditional radicals, but in fact they have been treated as radicals in this work, because of their phono-semantic structure.

Many Chinese characters display hierarchical structure. They look like branching trees of characters based on their structure. Components (radicals) build up new characters, which in turn act as components in more complex characters. For example, the character “按” should be divided into two parts: “扌” and “安”. But for the character “安”, it can be divided further into “宀” and “女” (see Table 1 and Fig. 1). According to the tree structure, the character like “按” should be decomposed into “扌” and “安”, but not “扌”, “宀”, and “女”.

Many radicals are distorted or change in form in order to fit into a block with other components. In some cases, these written forms may have several variants. Some of the most important variant written forms are as follows: 刀 → (knife), 人 → 亻 (man), 心 → 忄 (heart), and 犬 → 犭 (dog), etc. Although the radicals and their variants have the same meanings, they are treated as different radicals due to their different shapes.

As described above, the character networks are constructed in the following ways: (1) radicals are served as nodes of the networks; (2) connections exist between two radicals if they can form a character. Let us consider the networks of Chinese radicals, $G_L = (W_L, E_L)$, where $W_L = \{w_i\}$, $(i = 1, \ldots, N_L)$ is the set of N_L nodes (radicals) and $E_L = \{w_i, w_j\}$ is the set of edges or connections (formed characters) between radicals. Here, $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge between radicals w_i and w_j. Two connected radicals are adjacent and the degree of a given radical is the number of edges that connect the given radical with other radicals. Fig. 2 shows what the network looks like.

3. Empirical results

In this section, 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 $k_i(k_i - 1)/2$ edges can exist among them. The clustering coefficient C_i of the node i is defined as $2e_i/k_i(k_i - 1)$, 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 have degree k.

Table 1
The Chinese characters 按 (press) and 安 (safe), and their corresponding decompositions into radicals

Radical	Chinese pronunciation	English meaning
宀	mián	Roof
女	*nǚ*	Woman
安	*ān*	Safe
扌	*shǒu*	Hand
按	*àn*	Press

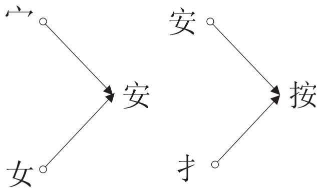

Fig. 1. The illustration of Chinese character decomposition and organizing rule.

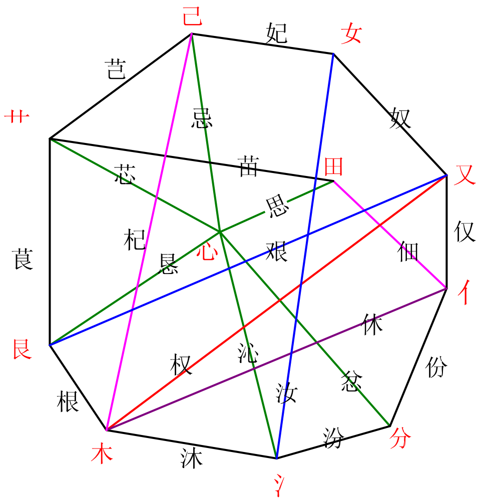

Fig. 2. The illustration of a Chinese character network.

The Chinese character networks we study here contain 1624 nodes (radicals) and 6652 edges (characters). Our analysis of the networks focuses on five properties [20]: sparsity, short path-lengths, high neighborhood clustering, degree distributions, and disassortative mixing.

Sparsity: The networks have 1624 nodes, and the average degree or average number of connections $\langle k\rangle$ is about 8.19. Given the size of the networks and the number of connections, it can be observed that the networks are sparse: on average, a node is connected to only a very small percentage of other nodes. The total edges of the networks, i.e., Chinese characters are 6652. The total edges of the complete graph with 1624 nodes are $C_{1624}^2 = (1624 \times 1623)/(1 \times 2)$. The ratio between them is 0.5%. From the ratio 0.5% and the average degree $\langle k\rangle = 8.19$, we can say that the combination of Chinese characters is sparse.

Short path-lengths: The networks display very short average path-length, i.e., 3.1651 and diameter $= 8$ relative to the sizes of the networks. For instance, the average path length (L) is about 3 while the maximum path length (D) is only 8. That is, at most eight associative steps (independent of direction) separate any two radicals in the 6652 characters. These short path-lengths and small diameters are well-described by random graphs of equivalent size and density, consistent with Watts and Strogatz's findings for their small-world networks [1].

Neighborhood clustering: In addition to a short average path length, the networks have a relatively high clustering coefficient. Compared with the clustering coefficient $C_{\text{rand}} = \langle k\rangle/N = 0.0050$ of a corresponding random graph, the clustering coefficient $C = 0.5311$ of the network is about 106 times higher than that of the random graph. For further comparison, we use the estimate $C' = (1/\langle k\rangle N)(\frac{\langle k^2\rangle}{\langle k\rangle} - 1)^2 = 0.4557$ (see [14, Eq. (3)]) for the upper bound of the average clustering coefficient of a network with the same degree distribution as the original data set, but randomly assigned links. As the values C and C' show, our network exhibits a higher average clustering coefficient than the network with links distributed randomly according to the same degree distribution.

Degree distribution: A vast majority Chinese characters are phono-semantic compounds and the characters are composed of a meaningful radical (semantic) and a radical used to indicate its pronunciation (phonetic). Both of them play a different role in the characters' construction. Generally speaking, the semantic radicals have high degree and the phonetic radicals have low degree. Moreover, some radicals can serve as semantic and phonetic radicals in different words. These are the factors which lead to unexpected degree distributions features and complexity of the net.

Fig. 3 plots the degree distributions for the nodes of the network in three coordinates. The plots show that the degree distribution for the networks is unlike Poisson, exponential and power-law distributions. Fig. 3(c) displays that there exist more high degree radicals at the tail than that of the power-law distribution. These

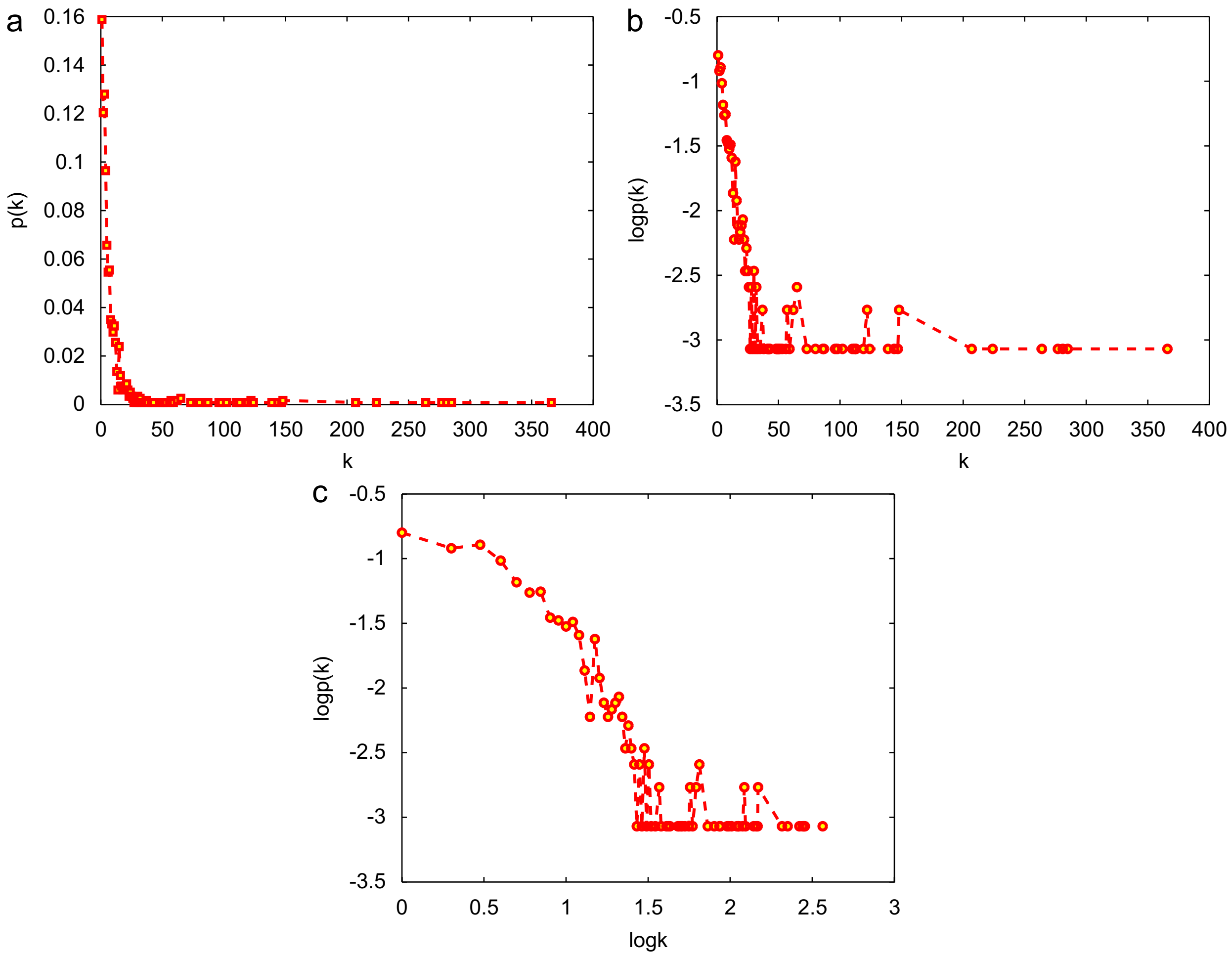

Fig. 3. The illustration of $k - p(k)$, $k - \log p(k)$ and $\log - \log$ degree distributions.

radicals can be thought of as the hubs of the network. These hubs typically correspond to important radicals, such as “亻” (people), “氵” (water), “木” (wood), “火” (fire), “土” (soil), and “钅” (metal).

Disassortative mixing: A network is said to show assortative mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections, otherwise it is called disassortative mixing. Social networks display specific features that put them apart from biological and technological ones. One of these features is assortative mixing [12,35–38] if the networks' assortative coefficient $r \geqslant 0$. r (see [35, Eq. (4)]) is defined as

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

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

In the Chinese character networks, we calculate the assortative coefficient $r = -0.4097$. Based on this value, we can say that unlike other social networks, the Chinese character networks display disassortative feature. In Section 5, we will explain why it is disassortative mixing ($r < 0$) and how it is advantageous for remembering.

We develop a model that reproduces this feature based on bipartite theory in Sections 4 and 5. The statistical properties arising from the simulations are in good agreement with those of the empirical Chinese character networks. The model also gives us an insight into the formation and evolution of Chinese characters.

4. Simulation of the formation of Chinese characters

More than 90% of Chinese characters takes a compound form, often one is semantic and one phonetic. In this section, 4952 phono-semantic characters are studied and their formation is simulated. The experiment results are compared with the empirical phono-semantic results.

The origin of phono-semantic characters can be explained in the following aspects: Firstly, people know the things from different aspects and at different levels. There may exist close relations among different things. In Chinese, the things with similar properties are represented by the same semantic radicals. The combinations of Chinese characters help to remember the characters and corresponding things easily. For example, “山” represents mountain, many characters in relation to mountain are created by “山”: “峰” (top of a mountain), “岳” (mountain), and “岭” (hill). Water is another example, which is represented by radical “氵”. Many characters related to water are created by “氵”, such as “河” (river), “汁” (juice), “汗” (sweat), and “泪” (tear).

Secondly, the generation of Chinese characters has close relations to our everyday life. The more frequently the things appear in everyday life, the greater possibility the corresponding radicals used to construct new characters have. For example, everyday we contact people, water, wood, fire, soil, and metal, therefore the corresponding radicals “亻” (people), “氵” (water), “木” (wood), “火” (fire), “土” (soil), and “钅” (metal) are used more frequently to construct characters than many other radicals because of their intimate relations to our life.

Thirdly, the generation of Chinese characters also depends on our culture, religious belief, and cognition. For example, the radical “礻” is related to religious belief. Many characters are constructed by it such as “神” (God), “社” (God of land), “祠” (temple), “祝” (pray), and “祈” (pray).

As most Chinese characters are phono-semantic, they are usually constructed with a phonetic and a semantic radical (character). It is natural to decompose the phono-semantic characters into two radicals (parts). Motivated by the empirical observations of the Chinese characters' formation and their phono-semantic structure, the following model of a bipartite network is proposed to study the assembly of Chinese characters. Some nodes in the model represent semantic radicals and the other nodes represent phonetic radicals. Two different type nodes are linked if they form a character.

A bipartite graph is a triple $G = (S, P, E)$, where S and P are two disjoint sets of vertices, and $E \subseteq S \times P$ is the set of edges between them. The networks can be viewed this way with S being the set of semantic radicals and P the set of phonetic radicals, each semantic radical being linked to the phonetic radical if they can form a character or part of it (see Fig. 4). Here 4952 phono-semantic characters are studied. S and P consist of 217 semantic radicals and 1293 phonetic radicals, respectively.

From the empirical networks we can obtain the distribution functions of the semantic and phonetic radicals, respectively. Here we denote them as $p(x_i)$ and $q(y_j)$. The semantic and phonetic radical sets are represented by $S = \{x_1, x_2, \ldots, x_n\}$, $n = 217$ and $P = \{y_1, y_2, \ldots, y_m\}$, $m = 1293$. Their corresponding degrees are denoted as $\{k_1, k_2, \ldots, k_n\}$ and $\{v_1, v_2, \ldots, v_m\}$. The distribution functions are $p(x_i) = k_i / \sum_{j=1}^{n} k_j$ and $q(y_i) = v_i / \sum_{j=1}^{m} v_j$.

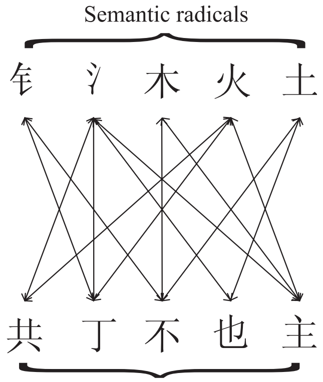

Fig. 4. Defining a bipartite graph of phono-semantic association. A set of semantic radicals S is linked to a set of phonetic radicals P. The constructed characters consist of “钉”, “钚”, “洪”, “汀”, “池”, “注”, “杯”, “柱”, “烘”, “灯”, “炷”, “坏” and “地”.

The procedure of the presented algorithm based on bipartite graph is as follows:

1. Select the iteration step N, where N is the number of the studied Chinese phono-semantic characters ($N = 4952$).
2. Randomly select a semantic radical from S, according to the following probability:
$$p(x_i) = \frac{k_i}{\sum_{o=1}^{n} k_o}, \quad i = 1, \ldots, n. \tag{2}$$
The selected semantic radical is denoted as x_i.
3. Randomly select a phonetic radical from P, according to the following probability:
$$q(y_i) = \frac{v_j}{\sum_{e=1}^{m} v_e}, \quad j = 1, \ldots, m. \tag{3}$$
The selected phonetic radical is denoted as y_j.
4. Connect x_i and y_j, if there exists a link between them, return to step 3.

5. Numerical results

We have carried out a thorough comparison with the empirical networks and the results are given as below. Because the models are stochastic, results vary from simulation to simulation. The described assortative coefficient r_{mean} of the simulating networks is averaged over 50 simulations.

Fig. 5 shows the degree distributions of our model and the empirical character networks. Both of the networks display a unique degree distribution unlike Poisson, exponential and power-law distributions. The degree distribution of the simulating networks mimics empirically measured distributions reasonably well. Although our model based on bipartite graph is different from Barabási's preferential attachment model [5], it still shows preferential attachment feature, i.e., high degree semantic (phonetic) radicals have high probability to be connected to phonetic (semantic) radicals.

The assortative coefficient from empirical data is $r = -0.4024$, while the coefficient of simulation result is $r_{mean} = -0.4291$ and its standard deviation is $std_r = 0.012$. These values show that r_{mean} does not vary too much, and it is similar to the empirical assortative coefficient. Unlike many social networks, the networks display disassortative mixing feature. Since the semantic radicals related to meaning have higher degree and the phonetic radicals have lower degree, the combination of a semantic radical and a phonetic radical

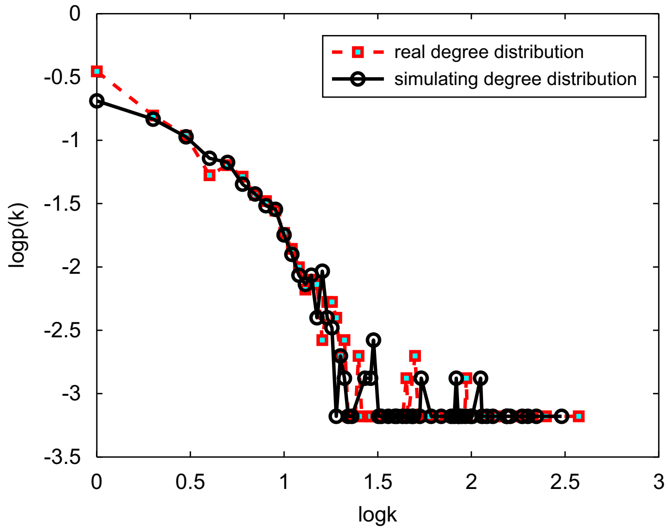

Fig. 5. The comparison of degree distributions of simulating networks and empirical networks in log–log coordinates.

shows disassortive feature naturally and reasonably. In addition, the degree distributions of the semantic and phonetic radicals also influence the disassortative mixing feature. This feature of Chinese characters' combination agrees with human's cognitive principle: the semantic radicals related to its meaning are easy to remember and the phonetic radicals with low degree reduce the burden and complexity of using them.

By the analysis, we have been able to provide a reasonable explanation for Chinese character structure, i.e., they have a non-Poisson distribution and disassortative mixing feature. In addition, the proposed model explains the origin and formation of Chinese phono-semantic characters. As described above, the high degree semantic radicals are closely related to our lives and they have high probability to be firstly combined into new characters. For instance, "氵" represents water or behavior and feature related to water; "足" represents foot or behavior and feature related to foot; "木" represents wood or behavior and feature related to wood; "扌" represents hand or behavior and feature related to hand. The phono-semantic characters like "河" (river), "汗" (sweat), "跑" (run), "柱" (column or pole) and "打" (hit) which are related to their relevant high degree radicals will be likely to be first created. Hence, it implies that the first created characters are more necessary and more useful for people's life than later created characters. These characters help people to express their thought and to communicate easily. Therefore the proposed model can provide a possible way to explain the origin and formation of the phono-semantic characters.

In this paper, we use the bipartite graph to model the structures of phono-semantic characters. In fact, any complex network can be viewed as a bipartite graph with some specific characteristics, and its main properties can be viewed as consequences of this underlying structure, therefore, using bipartite graph to study the structure properties of all Chinese characters will be our future work.

6. Conclusion and research in the future

Like other complex networks from nature and society, Chinese characters' networks also show a very short average distance between nodes, high local clustering, and a non-Poisson distribution. From the networks analysis, there are a few facts to be pointed out.

(1) Chinese characters have its special organizing principles. First, the radicals based on semantics show that they have intimate relations to nature, Chinese culture and our life. For example, the radical "氵" means water which is absolutely necessary in our life and many things are related to water, therefore the semantic-based radicals have high degree. Second, the combination of Chinese radicals has hierarchical structure: simple characters are created with strokes and radicals while complex characters are created with simple characters and simple radicals. Most of the generated radicals are phonetic-based and their degree is not high. Third, the empirical networks show small-world characteristics and a property of being different from scale free. These features ensure the language to be remembered and used efficiently. Fourth, the networks also show disassortative mixing feature: the high degree radicals (semantic) trend to connect themselves to the low degree radicals (phonetic).

(2) The construction of Chinese characters has its advantages. First, although there are about 1600 radicals in the networks, the basic radicals are less than 200 and most other radicals are created by combining them. That means the construction of a virtually infinite range of character combinations naturally from a limited set of basic units: semantic and phonetic radicals; Second, the character networks are sparse, and the averaged degree is small. That means it is not a heavy burden at all for our memory and our brain structure. Third, statistical properties including short path-length and big clustering coefficient also show that the construction and the organization of Chinese characters are coherent and integrative. Fourth, most semantic radicals have higher degree than the phonetic radicals. Because the characters, which they construct, represent related meanings to them, they are easy to remember semantically. Comparing with the semantic radicals, the phonetic radicals which have low degree facilitate to remember the same pronunciations of so many characters constructed by them.

The above features agree with the principle of least effort [39]. Frequent and close interactions between human and nature strengthen human memory, reduce the burden of storage in the brain, and thus make characters to be remembered reasonably and efficiently. In some sense, the principle of least effort might shape the combination and evolution direction of Chinese characters.

Although the resulting character networks display statistical features very different from random networks, it does not mean to rule out the random factors. As a complex adaptive system, there do exist random factors for combination of Chinese characters. Even in scale-free networks, random attachment still plays an important role as well as preferential attachment. The problem will be considered in the future.

Due to their special features, the formation and evolution of Chinese characters will be very helpful to study human's cognitive structure and mechanism. In the future we will focus on the research in that aspect. In addition, more complex and more realistic model should be considered.

Acknowledgments

We thank the reviewers for their valuable comments. We would like to thank Professor Frode Eika Sandnes and Siwei Luo for useful discussions and helpful suggestions. We are grateful to Ying Han, Wei Zhao and Xianhong Cheng for their efforts to prepare all the data. This work was supported by the National Natural Science Foundation of China under Grant nos. 60573023, 60673106 and the Specialized Research Fund for the Doctoral Program of Higher Education.

References

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

[2] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268.

[3] A.L. Barabási, R. Albert, H. Jeong, Scale-free characteristics of random networks, The topology of the World Wide Web, Physica A 281 (2000) 69.

[4] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509.

[5] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. 74 (2002) 47.

[6] N. Mathias, V. Gopal, Small worlds: how and why, Phys. Rev. E 63 (2) (2001) 1.

[7] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of networks, Adv. Phys. 51 (2002) 1079.

[8] M.E.J. Newman, The structure and function of complex networks, SIAM Rev. 45 (2003) 167.

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

[10] H. Jeong, S. Mason, A.-L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41.

[11] M.E.J. Newman, The structure of scientific collaboration networks, Proc. Natl. Acad. Sci. USA 98 (2001) 404.

[12] M.E.J. Newman, Why social networks are different from other types of networks, Phys. Rev. E 68 (2003) 036122.

[13] M.E.J. Newman, D.J. Watts, S.H. Strogatz, Random graph models of social networks, Proc. Natl. Acad. Sci. USA 99 (2002) 2566.

[14] J. Davidsen, H. Ebel, S. Bornholdt, Emergence of a small world from local interactions: modeling acquaintance networks, Phys. Rev. Lett. 88 (12) (2002) 128701.

[15] A. Broder, Graph structure in the web, Comput. Networks 33 (2000) 309.

[16] X. Li, G. Chen, A local-world evolving network model, Physica A 328 (1–2) (2003) 274.

[17] G.W. Flake, S.R. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization and identification of Web communities, IEEE Comput. 35 (2002) 66.

[18] R.F. I Cancho, R.V. Sole, The small-world of human language, Proc. R. Soc. London Ser. B 268 (2001) 2261.

[19] M.D. Hauser, N. Chomsky, W. Tecumseh Fitch, The faculty of language: what is it, who has it, and how did it evolve?, Science 298 (2002) 1569.

[20] M. Steyvers, J.B. Tenenbaum, The large scale structure of semantic networks: statistical analyses and a model of semantic growth, Cogn. Sci. 29 (1) (2005) 41.

[21] A.E. Motter, A.P.S. de Moura, Y.-C. Lai, P. Dasgupta, Topology of the conceptual network of language, Phys. Rev. E65 (2002) 065102.

[22] R.V. Sole, Syntax for free?, Nature 434 (2005) 289.

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

[24] R.V. Sole, B. Corominas, S. Valverde, L. Steels, Language networks: their structure, function and evolution, Trends Cogn. Sci. (2005).

[25] O. Kinouchi, A.S. Martinez, G.F. Lima, G.M. Lourenco, S. Risau-Gusman, Deterministic walks in random networks: an application to thesaurus graphs, Physica A 315 (2002) 665.

[26] A.J. Holanda, I. Torres Pisa, O. Kinouchi, A. Souto Martinez, E. Seron Ruiz, Thesaurus as a complex network, Physica A 344 (2004) 530.

[27] A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Detecting communities in large networks, Physica A 352 (2005) 669.

[28] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proc. Roy. Soc. London Ser. B 268 (2001) 2603.

[29] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, et al., Superfamilies of evolved and designed networks, Science 303 (2004) 1538.

[30] P. Allegrini, P. Grigolini, L. Palatella, Intermittency and scale-free networks: a dynamical model for human language complexity, Chaos Soliton Fract. 20 (1) (2004) 95.

[31] R.F. I Cancho, R.V. Sole, Patterns in syntactic dependency networks, Phys. Rev. E 69 (2004) 32767.

[32] R.F. I Cancho, The Euclidean distance between syntactically linked words, Phys. Rev. E 70 (2004) 056135.

[33] M.A. Changizi, S. Shimojo, Character complexity and redundancy in writing systems over human history, Proc. Roy. Soc. London Ser. B 272 (2005) 267.

[34] Y. Li, L.X. Wei, W. Li, Y. Niu, S.Y. Luo, Small-world patterns in Chinese phrase networks, Chin. Sci. Bull. 50 (3) (2005) 286.

[35] M.E.J. Newman, Assortative mixing in networks, Phys. Rev. Lett. 89 (2002) 208701.

[36] M.E.J. Newman, Mixing pattern in Networks, Phys. Rev. E 67 (2003) 026126.

[37] M. Catanzaroa, G. Caldarellib, L. Pietronero, Social network growth with assortative mixing, Physica A (2004) 338119.

[38] A.P. Quayle, A.S. Siddiqui, S.J.M. Jones, Modeling network growth with assortative mixing, Eur. Phys. J. B 50 (2006) 617.

[39] R.F. I Cancho, R.V. Sole, Least effort and the origins of scaling in human language, Proc. Natl. Acad. Sci. USA 100 (2003) 788.