Physica A 391 (2012) 5254–5263

Contents lists available at SciVerse ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# Chinese lexical networks: The structure, function and formation

Jianyu Li [a,*], Jie Zhou [b], Xiaoyue Luo [c], Zhanxin Yang [a]

[a] *Engineering Center of Digital Audio and Video, Communication of China, Beijing 100024, China*

[b] *Department of Automation, Tsinghua University, Beijing 100084, China*

[c] *Department of Mathematics, Linfield College, OR 97128, USA*

ARTICLE INFO

*Article history:*
Received 4 July 2011
Received in revised form 5 March 2012
Available online 27 May 2012

*Keywords:*
Complex networks
Scale free
Assortative mixing
Hierarchical structure

ABSTRACT

In this paper Chinese phrases are modeled using complex networks theory. We analyze statistical properties of the networks and find that phrase networks display some important features: not only small world and the power-law distribution, but also hierarchical structure and disassortative mixing. These statistical traits display the global organization of Chinese phrases. The origin and formation of such traits are analyzed from a macroscopic Chinese culture and philosophy perspective. It is interesting to find that Chinese culture and philosophy may shape the formation and structure of Chinese phrases.

To uncover the structural design principles of networks, network motif patterns are studied. It is shown that they serve as basic building blocks to form the whole phrase networks, especially triad 38 (feed forward loop) plays a more important role in forming most of the phrases and other motifs. The distinct structure may not only keep the networks stable and robust, but also be helpful for information processing.

The results of the paper can give some insight into Chinese language learning and language acquisition. It strengthens the idea that learning the phrases helps to understand Chinese culture. On the other side, understanding Chinese culture and philosophy does help to learn Chinese phrases. The hub nodes in the networks show the close relationship with Chinese culture and philosophy. Learning or teaching the hub characters, hub-linking phrases and phrases which are meaning related based on motif feature should be very useful and important for Chinese learning and acquisition.

## 1. Introduction

Network analysis approach has been successful in different fields including brain structures [1], protein–protein interaction networks [2], social interactions [3–6], the Internet and WWW [7–9], and language networks [10–16]. Two main features (small-world effect [17,18] and scale free features [19–21]) seem to be shared by most complex networks, both natural and artificial. The small-world network is often characterized according to two graph measurements of the network: clustering coefficient $C$ and characteristic shortest path length $L$ [18]. 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 [19–21]. Different from an exponential network, a scale-free network is inhomogeneous in nature: most of the nodes have very few link connections and yet a few nodes have an extremely high number of connections.

[*] Corresponding author. Tel.: +86 13 621212753; fax: +86 10 65779081.
*E-mail address:* lijianyu@tsinghua.edu.cn (J. Li).

0378-4371/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2012.05.058

Work based on complex networks has aroused more attention in the field of language modeling. Language networks mainly include a few networks: thesaurus networks [12,13,22,23], WordNet [24,12], Chinese character networks [25], word association networks [12,26], word co-occurrence networks [10,27,53,28–30], phonemic networks [31] and syntactic networks [32–36]. Solé [16,37] 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.

Chinese characters are not as many as the words of western languages. Generally used characters number about 3000 and these characters can form at least 100,000 phrases. Therefore the frequently used Chinese characters have a very powerful expression and the research of formation and organization of Chinese phrases is very important for language learning, acquisition and cognitive mechanism. But in the area of Chinese phrase or word networks not many results are yet reported. Li [38], Yamamoto [39] and Wang [40] built phrase related networks which display scale free and small-world features. In [38], the nodes are meaningful words which may be single Chinese characters or multi-character combinations. Chinese phrases are formed mainly based on single characters and two-character phrases should be the fundamental issue to study how Chinese phrases form and evolve.

The nodes of the building networks should be single characters, hence the network [38] is not ideal to explore the formation and organizing rule of phrases. We construct the networks in the following senses like [39,40]: (1) the single characters correspond to nodes of the network, and (2) an undirected or directed link exists between characters if they can form a phrase. Different from WordNet, HowNet and FrameNet, the networks are not designed by experts according to relationships such as meaning, and syntax, so they can reveal more reasonable and convincing results on how language self-organizes and evolves.

Except for the similar results such as small world and scale free as in Ref. [39,40], some more interesting results including disassortative mixing, hierarchical structure, motif structure of the two-character directed network, and Chinese culture related hub nodes are reported. These features play an important role in formation, organization and information processing of Chinese phrases. Moreover, some insight is obtained about Chinese acquisition, teaching and learning. The paper is organized as follows. Section 2 introduces the data source and acquisition. Section 3 focuses on computing the statistical properties of the undirected networks. In Section 4, analysis of motif structure of the directed networks is presented. Finally, Sections 5 and 6 end the paper with discussions and conclusions.

## 2. Data acquisition and construction of the networks

The purpose of the research is to study the formation and organization of Chinese phrases, hence the more regular phrases are included. The data were mainly collected from a few popular, frequently used and middle-sized Chinese dictionaries [54–56] such as the Contemporary Chinese Dictionary [54] which contains over 60,000 entries including characters, words, phrases, colloquialisms and idioms. The bigger dictionaries are not considered, because they contain a great many proper nouns and loanwords which are unrelated to the feature of phrases' formation and organization. The middle-sized database is better to explore the feature of Chinese phrase formation. In addition, the phrases which consist of over four characters are not considered. The data contain 69,417 two-character phrases, 11,581 three-character phrases and 28,533 four-character phrases. Since the data are gathered from dictionaries, every entry is considered as a phrase.

In Chinese the basic unit is the character which is a word or part of it. Two or more characters that form a syntactic and fixed unit are called a phrase. Most Chinese phrases consist of two characters, which are called two-character phrases. Based on two-character phrases, three-character phrases or four-character phrases which consist of three or four characters are generated.

The undirected phrase networks are constructed in the following ways (see Fig. 1, direction is not considered): (1) characters serve as nodes of the networks; (2) connections are established between two characters if they are next to each other in a phrase. For example, if *ABC* is a phrase which consists of three characters *A*, *B* and *C* in the order from *A*, to *B*, then to *C*, two edges are established. One edge is between *A* and *B*, and the other is between *B* and *C*. Let us consider the networks of Chinese characters, $G_L = (W_L, E_L)$, where $W_L = \{w_i\}$, $(i = 1, \ldots, N_L)$ is the set of $N_L$ nodes (characters) and $E_L = \{w_i, w_j\}$ is the set of edges or connections (phrases) between characters. Here, $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge between characters $w_i$ and $w_j$ which are next to each other in a phrase. Two connected characters are adjacent and the degree of a given character is the number of edges that connect the given character with other characters.

As shown in Table 2, there are three (Net1–Net3) kinds of undirected networks to be studied. Net1 is constructed with two-character phrases. Net2 is constructed with two- and three-character phrases. Net3 is constructed with two-, three- and four-character phrases.

In Section 4, the directed two-character phrase networks are studied. The directed networks are established like Net 1 (see Fig. 1). For example, if *AB* is a phrase which consists of character *A* and character *B* in the order from *A* to *B*, then the character *A* and character *B* serve as nodes, and an edge between *A* and *B* will be directed from *A* to *B*.

## 3. Statistical properties of the undirected networks

In this section, we calculate the statistical properties of the two-character phrase networks, and those of the networks with two-, three- and four-character phrases. 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.

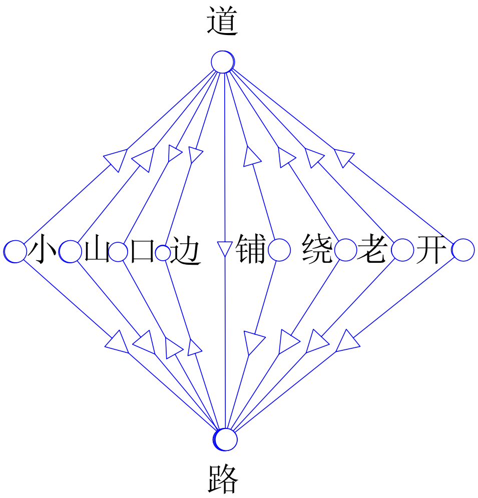
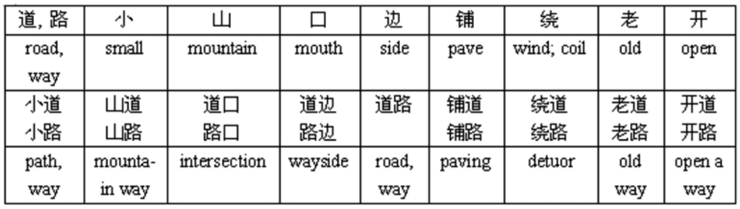


**Fig. 1.** Illustration of Chinese phrase networks. In the figure “ 道 ” and “ 路 ” are synonymous. Many related phrases are created by them. The meanings of the nodes and edges are listed in Table 1.

**Table 1**
Meaning of characters and two-character phrases which are nodes and edges of Fig. 1 are listed.



| 道, 路 | 小 | 山 | 口 | 边 | 铺 | 绕 | 老 | 开 |
|---|---|---|---|---|---|---|---|---|
| road, way | small | mountain | mouth | side | pave | wind; coil | old | open |
| 小道<br>小路 | 山道<br>山路 | 道口<br>路口 | 道边<br>路边 | 道路 | 铺道<br>铺路 | 绕道<br>绕路 | 老道<br>老路 | 开道<br>开路 |
| path, way | mounta-in way | intersection | wayside | road, way | paving | detuor | old way | open a way |

**Table 2**
Statistics properties for the phrase networks. ($n_1$: the number of nodes, $n_2$: the node number of the maximal component of the networks, $m$: the number of edges of the networks, $ra$: the ratio between the edge number of the networks and the complete graph with the same nodes. $\langle k\rangle$: the average number of connections, $L$: the average shortest path length, $D$: the diameter of the network, $L_{random}$: the average shortest path length with random graph of same size and density. Net1: two-character undirected phrase net, Net2: two- and three-character undirected net, Net3: two-, three- and four-character undirected net).

| | $n_1$ | $n_2$ | $m$ | $ra$ (%) | $\langle k\rangle$ | $L$ | $D$ | $L_{rand}$ |
|---|---|---|---|---|---|---|---|---|
| Net1 | 4858 | 4660 | 69,417 | 0.59 | 28 | 3.0441 | 9 | 3.2164 |
| Net2 | 4892 | 4718 | 80,998 | 0.68 | 34 | 2.9493 | 9 | 2.9985 |
| Net3 | 5322 | 5244 | 109,531 | 0.77 | 42 | 2.8459 | 9 | 2.8180 |

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 $\frac{2e_i}{k_i(k_i-1)}$, where $e_i$ is the 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 $k$.

In order to make the paper clear and easy to understand, Net1 which contains 4858 nodes (characters) and 69,417 edges (phrases) will be mainly discussed. Properties of Net2 and Net3 will be found in Tables 2 and 3. Our analysis focuses on seven properties: sparsity, short path-lengths, high neighborhood clustering, degree distributions, disassortative mixing, hierarchical structure and entropy.

### 3.1. Sparsity

The networks have 4858 nodes, and the average degree or average number of connections $\langle k\rangle$ is about 28 (see Table 2). Given the size of the networks and the number of connections, it can be observed that the networks are sparse: on average,

**Table 3**
Statistics properties for the phrase networks. ($C$: clustering coefficient, $C'$: the upper bound of the average clustering coefficient of a random network with the same degree distribution, $C_{random}$: the clustering coefficient for a random graph of same size, $\gamma$: power law exponent for the degree distribution of the networks and $H$ is Shannon's entropy. Net1: two-character undirected phrase networks, Net2: two- and three-character undirected net, Net3: two-, three- and four-character undirected net).

| | $C$ | $C'$ | $C_{rand}$ | $r$ | $\gamma$ | $H$ |
|---|---|---|---|---|---|---|
| Net1 | 0.4548 | 0.0857 | 0.0060 | −0.0645 | 3.32 | 3.9558 |
| Net2 | 0.4569 | 0.1105 | 0.0072 | −0.0786 | 3.172 | 4.0840 |
| Net3 | 0.4505 | 0.1432 | 0.0080 | −0.0938 | 3.054 | 4.2926 |

a node is connected to only a very small percentage of other nodes. The total edges of the networks, i.e. Chinese phrases are 69,417. The total edges of the complete graph with 4858 nodes are $C^2_{4858} = \frac{4858\times 4857}{1\times 2}$. The ratio between them is 0.59%. From the ratio 0.59% and the average degree $\langle k\rangle = 28$, we can say that the combination of Chinese phrases is sparse.

### *3.2. Short path-lengths*

The networks display very short average path-length, i.e. 3.0441 and diameter = 9 (see Table 2). For instance, the average path length ($L$) is about 3 while the maximum path length ($D$) is only 9. That is, at most nine associative steps (independent of direction) separate any two radicals in the 4858 characters. These short path lengths and small diameter are well described by random graphs of equivalent size and density, consistent with Watts and Strogatz's findings for their small-world networks [18]. The short path length means it takes just about three steps for any information in the net to 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 a long path, then without doubt, the shorter path can save time, energy and room to store, remember and process information. That agrees with the least effort principle [41]. Finally, if phrases and characters in the network can be treated as concepts, short average path implies that concepts reveal close relationships among them in a compact way which is helpful for information processing.

We also calculate the number of direct, second, and third neighbors of the highest degree node "不": 3747. The ratio 80.4% between 3747 and 4660 (see Table 2) shows the networks have small lengths. The hub node is linked to most of the characters in the networks through not more than three steps, which implies that the hub nodes have strong capability to construct a great many phrases. Human memory is not an isolated behavior. Due to the hub's powerful connecting ability, to remember more than 3000 related characters (as connected nodes) and more phrases (as edges) is not a big burden.

### *3.3. Neighborhood clustering*

In addition to a short average path length, the networks have a relatively high clustering coefficient. Compared with the clustering coefficient $C_{rand} = \langle k\rangle/N = 0.0058$ of a corresponding random graph, the clustering coefficient $C = 0.4548$ of the network is about 78 times higher than that of the random graph. For further comparison, we use the estimate $C' = \frac{1}{\langle k\rangle N}(\frac{\langle k^2\rangle}{\langle k\rangle} - 1)^2 = 0.0857$ (see equation 3 in Ref. [6]) 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.

From the definition of clustering coefficient, the neighbors of a node tend to interact among themselves more frequently than random networks and higher clustering means the net gathers similar pieces of information. Therefore many shorter meaning-related phrases are created in a local range. This helps to find the necessary and related information quickly and accurately.

### *3.4. Degree distribution*

Fig. 2 plots the degree distributions for the word 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 three undirected networks, power functions fit the degree distributions almost perfectly. The exponents $\gamma$ of the best-fitting power laws (corresponding to the slopes of the lines in Fig. 2) are quite similar in all three cases, varying between 3.054 and 3.32 (see Table 3). The high-connectivity characters at the tail of the power-law distribution can be considered as the hubs of the semantic network. In the networks, these hubs typically correspond to important general characters, such as "人", "不" and "一".

### *3.5. Disassortative mixing*

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

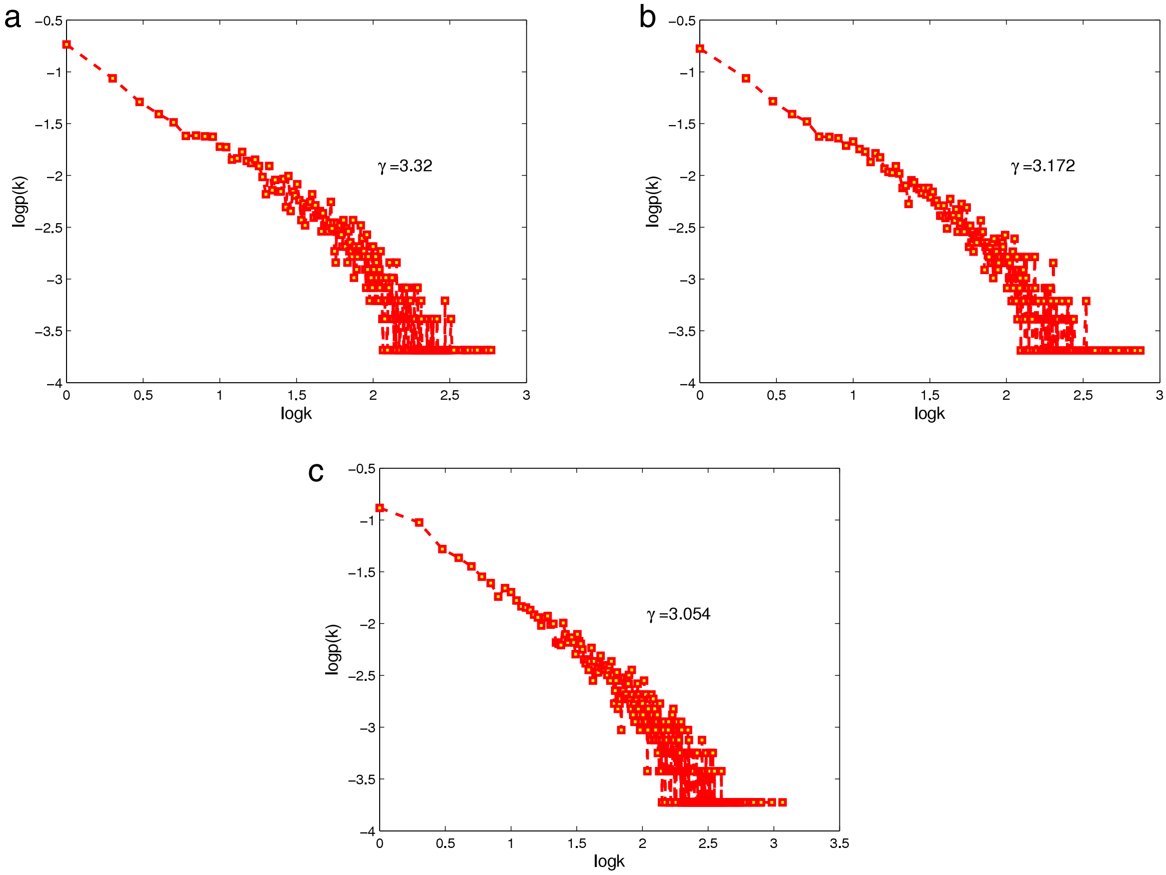


**Fig. 2.** Illustration of log–log degree distributions of Net1, Net2 and Net3.

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 [4,42–45] if the networks' assortative coefficient $r \geq 0$. $r$ (see equation 4 in Ref. [42]) is defined as

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

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

The assortative coefficient $r = -0.0645$ is unlike other social networks, the Chinese phrase networks display a disassortative feature: higher degree nodes tend to be connected to the low degree nodes. Comparing with the assortative coefficient $r = -0.4097$ of Chinese character networks [25], the trend of higher degree nodes linked to lower degree nodes is not so obvious. This interesting difference means there exist strict constraints of constructing Chinese characters, but creation of Chinese phrases is relatively free.

### *3.6. Hierarchical structure*

In literature, several concepts are proposed to measure the hierarchy in a network, such as the hierarchical path [46], the scaling law for the clustering coefficients of the nodes [47], the hierarchical components/degree [48], etc. These measures can tell us the existence and the extent of hierarchy in a network. This intrinsic hierarchy can be characterized in a quantitative manner using the recent finding of Dorogovtsev, Goltsev, and Mendes [49] that in the deterministic scale-free networks, the clustering coefficient of a node with $k$ links follows the scaling law

$$c(k) \longrightarrow k^{-1}. \tag{2}$$

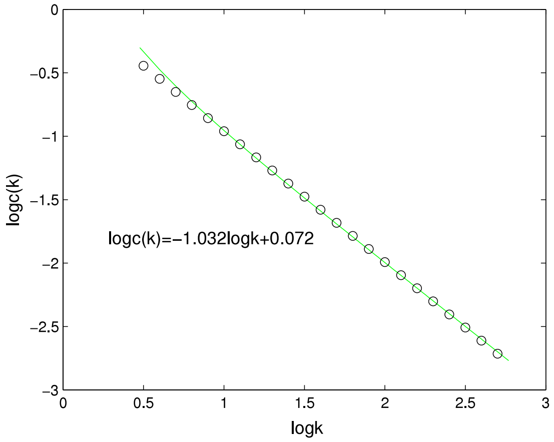


**Fig. 3.** Illustration of and $k - \log c(k)$ clustering distribution.

To investigate if such hierarchical organization is present in real networks, Ravasz et al. [47] measured the $C(k)$ function for several networks for which large topological maps are available. It includes actor network, language network, the World Wide Web, and power grid. And they find the scaling law (Eq. (2)) of this model indeed characterizes those real networks [50]. Similarly, the curve $C(k)$ in Fig. 3 follows the same scale law (Eq. (2)) very well, displaying that the phrase networks have a hierarchical feature. The level of clustering is not equivalent for all nodes, and appears to be a function of the degree of a given node. While low degree nodes belong to highly cohesive, densely interlinked clusters, hubs do not; their neighbors have a smaller chance of linking to each other.

### 3.7. Entropy

Given Shannon's definition of entropy $H = -\sum_{i=1}^{N} p_i \log p_i$, $H$ in Table 3 gets bigger from Net1 to Net3. This means the net shows greater uncertainty as the values are getting bigger, and the higher the uncertainty and complexity become, the more powerful and efficient the language is. From the viewpoint of entropy, three- and four-character phrases enrich the networks and improve the efficiency of two-character networks further.

## 4. Motif structure of the directed networks

### 4.1. Motif detection

In this section, the two-character phrase networks discussed above are considered and every edge is directed (see Fig. 1). Fig. 1 shows that there might exist bigger patterns such as network motifs which play an important role to form Chinese phrases. These patterns have been studied, especially in Biology. The concept of network motifs was first proposed by Uri Alon's group [51,52]. Network motifs are defined as patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks. The motif study in the paper not only displays how phrases are organized, but also may give us insight into how to construct semantical networks.

Motifs are to be found in the following way: by breaking a network down into all possible $n$-node subgraph patterns and counting them, it is possible to compare those counts against randomly generated networks with the same characteristics. Then, where certain $n$-node subgraph patterns are significantly more prevalent than in the random case, these are considered motifs of the network. These small subgraphs can be considered as simple building blocks from which the network is composed.

We detected the network motifs using the $Z_{score}$ [53] which is a certain variable to weigh the statistical significance in real networks with a comparison to the corresponding randomized networks. Generally speaking, the higher $Z_{score}$ a motif has, the more significantly it can present typical characteristics in a network.

### 4.2. Motif types and function

In this paper, three-node motifs (also called triads) are studied. Fig. 4 describes the triads with higher $Z_{score}$ that is larger than 15.0 while Table 4 lists all the triads. Comparing with other types of networks [53], different motifs have been detected in the Chinese phrase network. Ron Milo et al. [53] studied word-adjacency networks in which each node represents a word and a directed connection occurs when one word directly follows the other in the text. In their paper, the motifs of Fig. 4 are underrepresented with lower $Z_{score}$. Their explanation about their higher $Z_{score}$ motifs is that words belong to categories and a word from one category tends to be followed by one from a different category. Because their data are from sentences, a

**Table 4**
The motif IDs (triads) and their $Z_{score}$ are listed.

| *Triad* | 6 | 12 | 14 | 36 | 38 | 46 | 74 | 78 | 98 | 102 | 108 | 110 | 238 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $Z_{score}$ | −48 | −34 | −38 | −48 | 45 | 34 | −37 | −29 | 1.8 | 18 | 32 | 27 | 15 |

**Table 5**
The motif IDs (triads), their $Z_{score}$, the number of the characters which the motifs cover, and the number of the phrases which the motifs form are listed.

| *Triad* | 38 | 46 | 102 | 108 | 110 | 238 | All motifs |
|---|---|---|---|---|---|---|---|
| $Z_{score}$ | 45 | 34 | 18 | 32 | 27 | 15 | |
| *Characters* | 3,307 | 1,877 | 2,073 | 2,067 | 1,047 | 263 | 3,361 |
| *Phrases* | 59,678 | 21,704 | 22,806 | 23,162 | 8,110 | 1,108 | 65,933 |

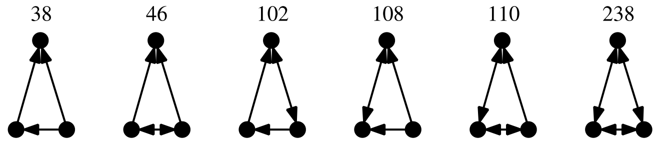


**Fig. 4.** Higher $Z_{score}$ (larger than 15.0) motifs (triads).

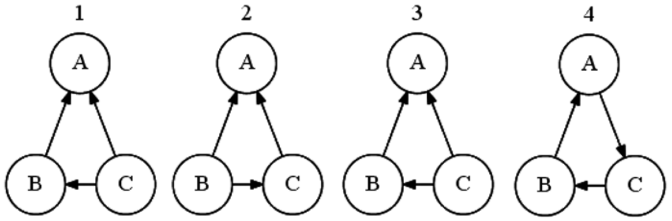


**Fig. 5.** Decomposition of triads 46, 108 and 102.

grammar rule could shape the motif structure. In our paper, the networks are constructed with phrases, not sentences. They don't follow the grammar rule in Ref. [53] and different motifs have been identified.

Table 5 shows the motifs have a powerful capability to form a great many phrases. They cover 3361 characters and 65,933 phrases. 95% (65,933/69,417) of phrases are constructed by the six motifs of Fig. 4. They play the role of basic building blocks and the networks are mainly composed of the motifs. This shows Chinese phrase networks are formed following a similar self-organizing rule in local structure and the recurring motifs serve as the basic building block to weave the whole networks. Especially, Triad 38 (feed forward loop–FFL) plays a more important role in phrase formation, and it forms 86 percent (59,678/69,417) of the phrases.

FFL plays a very important role in the formation of other motifs. Analyzing Fig. 4, it's not difficult to find the relationship between FFL and the other five motifs. Triads 46 and 108 are composed of two FFLs which share two edges and only a pair of edges is different with opposite direction Fig. 5 (see Figs. 5.1 and 5.2). Triad 102 is composed of an FFL and a triad 98 (see Figs. 5.3 and 5.4). Triad 110 is composed of three FFLs and a triad 98. Triad 238 is composed of six FFLs and two triads 98. Therefore FFL is not only an important pattern to form phrases, but also shows a close relationship with other motifs. The result is a very interesting finding and is reported for the first time as far as we know.

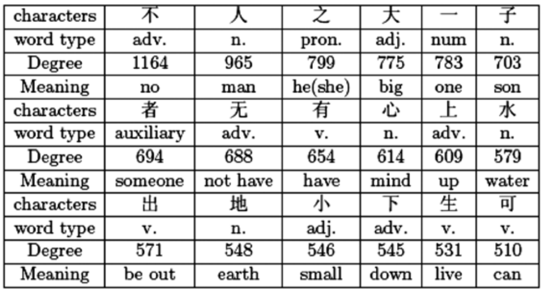
Why does FFL play such an important role in the networks? That may be explained as follows. Given three characters *A*, *B* and *C*, *A* connects to *B*, *B* connects to *C*. Generally speaking, the word *AB* and word *BC* have related meaning, and *ABC* has related or similar meaning to *AB* and *BC*. In most cases, *ABC* can be represented by *AC* and maintains the same meaning, but not *CA* which may lead to complication for human memory. Also *A*, *B* may have close meaning and if *A* connects to *B*, *B* connects to *C*, then *A* may have high chances to connect to *C*, and *AC* and *BC* share similar or opposite meanings. Therefore, FFL has great capability to form a great many phrases. The formation and organization of Chinese phrases based on FFL displays a few advantages. First, phrase *ABC* can be replaced by *AC*, and one character is saved, but meaning is maintained, which improves the efficiency for language communication; Second, the FFL structure is the easiest and natural way to form new phrases; Third, the loop structure of FFL is stable and information or concepts related to it are not easily lost. Even if some edge (phrase) is broken, it can be recovered by the FFL structure. Finally the FFL structure may reveal the basic structure and mechanism of the brain which processes and extracts related or similar information.

In order to further study the relationship among the motifs in Fig. 4, the character set which each motif in the networks covers and the intersection of each pair of sets are listed in Table 6. A few interesting results can be obtained from Table 6. First, the number of the character sets in triad 238 is the smallest, with 263 characters most of which appear in every other

**Table 6**
The motif IDs (triads), the number of the characters shared between any two motifs and the number of the characters which the motifs cover are listed.

| *Triad* | 38 | 46 | 102 | 108 | 110 | 238 |
|---|---|---|---|---|---|---|
| 38 | – | 1860 | 1926 | 2047 | 1044 | 263 |
| 46 | 1860 | – | 1585 | 1574 | 1007 | 263 |
| 102 | 1926 | 1585 | – | 1657 | 1011 | 262 |
| 108 | 2047 | 1574 | 1657 | – | 1017 | 262 |
| 110 | 1044 | 1007 | 1011 | 1017 | – | 258 |
| 238 | 263 | 263 | 262 | 262 | 258 | – |
| *Characters* | 3307 | 1877 | 2073 | 2067 | 1047 | 263 |

**Table 7**
Degree, word type and meaning of top 16 highest-degree characters. Most of the characters are polysemic, but in the table one type is listed.



| characters | 不 | 人 | 之 | 大 | 一 | 子 |
|---|---|---|---|---|---|---|
| word type | adv. | n. | pron. | adj. | num | n. |
| Degree | 1164 | 965 | 799 | 775 | 783 | 703 |
| Meaning | no | man | he(she) | big | one | son |
| characters | 者 | 无 | 有 | 心 | 上 | 水 |
| word type | auxiliary | adv. | v. | n. | adv. | n. |
| Degree | 694 | 688 | 654 | 614 | 609 | 579 |
| Meaning | someone | not have | have | mind | up | water |
| characters | 出 | 地 | 小 | 下 | 生 | 可 |
| word type | v. | n. | adj. | adv. | v. | v. |
| Degree | 571 | 548 | 546 | 545 | 531 | 510 |
| Meaning | be out | earth | small | down | live | can |

motif. Second, the characters (3307) in FFL almost cover the characters of the other five motifs. 2529 characters are used in the five motifs, but 2475 characters appear in FFL, which means about 97.9(2475/2529) percent are included in FFL. This further shows FFL plays a very important role in the formation of Chinese phrases, and is a basic pattern to form the whole networks. Third, the characters which appear in the six motifs are commonly used characters. These phenomena display that in Chinese many characters have a high tendency to be reused to form new phrases. Because the characters of different motifs are commonly used Chinese characters, these characters from every motif provide an alternative method for choosing commonly used Chinese characters.

Network motifs have been identified in a wide range of networks across many scientific disciplines and are suggested to be the basic building blocks of most complex networks [51–53,57,58]. The result in our paper has strengthened the conclusion that motifs act as basic building blocks. The research on language motifs is therefore helpful in uncovering the basic building blocks of complex networks and also offers a useful example to compare with other types of complex networks.

## 5. Discussion

The features of the networks are affected by Chinese culture and philosophy. In Table 6, the top 18 highest degree characters are listed. It's not difficult to see why these characters are selected as hub nodes. Human beings are the center of all societies (see Table 7) hence it is natural to select “人” (man) and “者” (man) as hubs to form a great many phrases in Chinese. “地” (earth) and “水” (water) are the most important resources for man to survive. They are treated as hubs. There is a deeper reason why these four characters become hubs. The cornerstone of traditional Chinese culture is to keep the harmonious and integrative relationships between man and nature “天人合一”, while “人” (man), “者” (man) “天” (heaven), “地” (earth) and “水” (water) are the key elements in this organic whole. Hence they are strengthened to become hubs. Meanwhile, Doctrine of Mean is the principle of Chinese Philosophy. Many characters highly related to the rule are viewed as hubs. For example, characters “不” (no) and “一” (one, unity) are used to form plenty of phrases which mean to maintain balance. They are treated as hubs. Also, there are some interesting antonyms such as “大” (big), “小” (small), “有” (have), “无” (not have), “上” (up) and “下” (down). Each of the antonym pairs can form many phrases. These phrases usually mean to keep balance between big and small, have and do not have, and up and down. Finally, “心” (heart, mind) which plays an important role in the spiritual life is also selected as a hub.

Disassortative linking means hubs tend to connect the low degree nodes. The low degree nodes tend to interconnect to form small blocks (high clustering features). Hence it is natural to conclude that disassortative linking and the higher clustering in a local range lead to the emergence of hierarchical structure.

The hierarchical structure suggests that Chinese phrases have an organizational structure that repeats itself at various scales. At the local level, there are many small, densely interconnected clusters. The small groups share close meaning, they are interconnected to create a lot of meaning-related edges (phrases). These groups combine to form larger, less interconnected groups, and these groups again combine to form larger and even less cohesive groups. But when the groups get bigger, the nodes in them become more heterogeneous. Therefore the ability to connect to other nodes in the group gets weaker and lower clustering.

It's unnecessary and also a burden to create a lot new characters to be remembered, while using the limited existing characters (frequently used) to form lots of phrases in motif style is better for human memory. The high $Z_{score}$ motifs in Fig. 4 share a closed path (loop) feature. This may reveal the loop style is a stable structure for memory or information processing (related patterns). The motif structure may strengthen and help to extract the demanded information. It is more reasonable to suppose that they reflect, at least in part, some abstract features of semantic organization in the human being's brain. This may help to develop semantic networks (entities, relationships) and optimize the design of semantic networks.

## 6. Conclusion

Like other complex networks from nature and society, Chinese phrase networks also show very short average distance between nodes, high local clustering, disassortative mixing, hierarchical structure and a power-law distribution.

We identified network motifs that reflect the underlying Chinese network, they are basic building blocks to form the whole networks, especially the role of FFL in forming most of the phrases and other motifs has been explored. Much information is stored because of the motif structure.

The results of the paper can give some insight into Chinese language learning and language acquisition. The high degree nodes (characters) and important relationships (edges) between hubs should be taught first to the learners and then expanded to other phrases which they link to. Because of the role of the hub, it should be helpful to learn other characters and phrases in an efficient, tight and integrated way. Since Chinese culture and philosophy strengthen the hubs, the hub-related learning could be very helpful to understand the culture better. The loop style of the motifs may imply that teaching phrases is a better way to remember and recall the related phrases in the motifs.

The future study should be in the following fields:

1. We will consider using the ant colony algorithm to find basic and general vocabulary and to detect the community structure in the networks.

2. Larger motif patterns will be tested and semantic information based on motifs will be mined.

3. The character frequency will be combined as edge weights into networks. The related statistical properties will be analyzed.

## Acknowledgments

The authors would like to thank Ms Shu Chen for helpful suggestions and comments. This work was supported by the National Natural Science Foundation of China under Grant No 61020106004, 61021063 and the Important National Science and Technology Specific Projects of China (2010ZX03005-001).

## References

[1] T.B. Achacoso, W.S. Yamamoto, AY's Neuroanatomy of C. elegans for Computation, CRC Press, Boca Raton, FL, 1992.
[2] H. Jeong, S.P. Mason, A.L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[3] M.E.J. Newman, The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences of the Unite States of America 98 (2001) 404–409.
[4] M.E.J. Newman, J. Park, Why social networks are different from other types of networks, Physical Review E 68 (2003) 036122.
[5] M.E.J. Newman, D.J. Watts, S.H. Strogatz, Random graph models of social networks, Proceedings of the National Academy of Sciences of the Unite States of America 99 (2002) 2566–2572.
[6] J. Davidsen, H. Ebel, S. Bornholdt, Emergence of a small world from local interactions: modeling acquaintance networks, Physical Review Letters 88 (12) (2002) 128701.
[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, et al., Graph structure in the web, Computer Networks 33 (2000) 309–320.
[8] X. Li, G. Chen, A local-world evolving network model, Physica A 328 (1-2) (2003) 274–286.
[9] G.W. Flake, S. Lawrence, C.L. Giles, F.M. Coetzee, Self-organization and indentification of web communities, IEEE Computer 35 (2002) 66–71.
[10] R.F.I Cancho, R.V. Solé, The small- world of human language, Proceedings of The Royal Society of London. Series B 268 (2001) 2261–2265.
[11] M.D. Hauser, N. Chomsky, W.T. Fitch, The faculty of language: What is it, who has it, and how did it evolve? Science 298 (2002) 1569–1579.
[12] M. Steyvers, J.B. Tenenbaum, The largescale structure of semantic networks: statistical analyses and a model of semantic growth, Cognitive Science 29 (1) (2005) 41–78.
[13] A.E. Motter, A.P.S. de Moura, Y.C. Lai, P. Dasgupta, Topology of the conceptual network of language, Physical Review E 65 (2002) 065102.
[14] R.V. Solé, Syntax for free? Nature 434 (2005) 289.
[15] M.A. Nowak, D.C. Krakauer, The evolution of language, Proceedings of the National Academy of Sciences of the Unite States of America 96 (1999) 8028–8033.
[16] R.V. Solé, B.C. Murtra, S. Valverde, L. Steels, Language Networks: their structure, function and evolution, Trends in Cognitive Sciences (2005).
[17] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[18] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[19] A.L. Barabási, R. Albert, H. Jeong, Scale-free characteristics of random networks, the topology of the World Wide Web, Physica A 281 (2000) 69–77.
[20] A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[21] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Reviews of Modern Physics 74 (2002) 47–97.
[22] O. Kinouchi, A.S. Martinez, G.F. Lima, et al., Deterministic walks in random networks: an application to thesaurus graphs, Physica A 315 (2002) 665–676.
[23] A.J. Holanda, I.T. Pisa, O. Kinouchi, A.S. Martinez, E.E.S. Ruiz, Thesaurus as a complex network, Physica A 344 (2004) 530–536.
[24] M. Sigman, G.A. Cecchi, Global organization of the Wordnet lexicon, Proceedings of the National Academy of Sciences of the Unite States of America 99 (3) (2002) 1742–1747.
[25] J. Li, J. Zhou, Chinese character structure analysis based on complex networks, Physica A 380 (2007) 629–638.

[26] A. Capocci, V.D.P. Servedio, G. Caldarelli, F. Colaiori, Detecting communities in large networks, Physica A 352 (2005) 669–676.
[27] S.N. Dorogovtsev, J.F.F. Mendes, Language as an evolving word web, Proceedings of the Royal Society of London B 268 (2001) 2603–2606.
[28] S. Zhou, G. Hu, Z. Zhang, J. Guan, An empirical study of Chinese language networks, Physica A 387 (12) (2008) 3039–3047.
[29] Y. Shi, W. Liang, J. Liu, C. Tse, Structural equivalence between co-occurrences of characters and words in the chinese language, in: International Symposium on Nonlinear Theory and its Applications, 2008, pp. 94–97.
[30] L. Sheng, C. Li, English and chinese languages as weighted complex networks, Physica A: Statistical Mechanics and its Applications 388 (12) (2009) 2561–2570.
[31] S. Yu, H. Liu, C. Xu, Statistical properties of chinese phonemic networks, Physica A: Statistical Mechanics and its Applications 390 (7) (2010) 1370–1380.
[32] R.F.I. Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Physical Review E 69 (2004) 32767.
[33] R.F.I. Cancho, The Euclidean distance between syntactically linked words, Physical Review E 70 (2004) 056135.
[34] B.C. Murtra, S. Valverde, R.V. Solé, The ontogeny of scale-free syntax networks: phase transitions in early language acquisition, Advances in Complex Systems 12 (3) (2009) 371–392.
[35] H. Liu, The complexity of chinese syntactic dependency networks, Physica A: Statistical Mechanics and its Applications 387 (12) (2008) 3048–3058.
[36] H. Liu, F. Hu, What role does syntax play in a language network? Europhysics Letters 83 (2008) 18002.
[37] R.V. Solé, B.C. Murtra, S. Valverde, L. Steels, Language networks: their structure, function, and evolution, Complexity 15 (2010) 20–26.
[38] Y. Li, L. Wei, Y. Niu, J. Yin, Structural organization and scale-free properties in chinese phrase networks, Chinese Science Bulletin 50 (13) (2005) 1305–1309.
[39] K. Yamamoto, Y. Yamazaki, A network of two-chinese-character compound words in the japanese language, Physica A: Statistical Mechanics and its Applications 388 (12) (2009) 2555–2560.
[40] J. Wang, L. Rong, An empirical study on chinese word–word network based on complex network theory, Journal of Dalian Maritime University 4 (2008) 15–18.
[41] R.F.I Cancho, R.V. Solé, Least effort and the origins of scaling in human language, Proceedings of the National Academy of Sciences of the Unite States of America 100 (2003) 788–791.
[42] M.E.J. Newman, Assortative mixing in networks, Physical Review Letters 89 (2002) 208701.
[43] M.E.J Newman, Mixing pattern in networks, Physical Review E 67 (2003) 026126.
[44] M. Catanzaro, G. Caldarelli, L. Pietronero, Social network growth with assortative mixing, Physica A 338 (2004) 119–124.
[45] A.P. Quayle, A.S. Siddiqui, S.J.M. Jones, Modeling network growth with assortative mixing, European Physical Journal B 50 (2006) 617–630.
[46] A. Trusina, S. Maslov, P. Minnhagen, K. Sneppen, Hierarchy and anti-hierarchy in real and model networks, Physical Review Letters 92 (2004) 178702.
[47] E. Ravasz, A.L. Barabási, Hierarchical organization in complex networks, Physical Review E 67 (2003) 026112.
[48] L.D.F. Costa, The hierarchical backbone of complex networks, Physical Review Letters 93 (2004) 098702.
[49] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Pseudofractal scale-free Web, Physical Review E 65 (1–4) (2002) 066122.
[50] D.H. Kim, G.J. Rodgers, B. Kahng, D. Kim, Modelling hierarchical and modular complex networks: division and independence, Physica A 351 (2005) 671–679.
[51] U. Alon, Network motifs: theory and experimental approaches, Nat. Rev. Genet. 8 (6) (2007) 450–461.
[52] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824–827.
[53] R. Milo, S. Itzkovitz, N. Kashtan, et al., Superfamilies of evolved and designed networks, Science 303 (5663) (2004) 1538–1542.
[54] S. Lv, S. Ding, Contemporary Chinese Dictionary, fifth ed., The Commercial Press, 2009.
[55] X. Su, Mordern Chinese Word Formation Dictionary, Shanghai Cishu Press, 2009.
[56] S. Kang, H. Liu, et al., Dictionary of Contemporary Chinese New Words, Shanghai Cishu Press, 2009.
[57] L. Krumov, C. Fretter, M.M. Hannemann, K. Weihe, M.T. Hütt, Motifs in co-authorship networks and their relation to the impact of scientific publications, The European Physical Journal B-Condensed Matter and Complex Systems (2011) 1–6.
[58] J.S. Baras, P. Hovareshti, H. Chen, Motif-based Communication Network Formation for Task Specific Collaboration in Complex Environments, in: Proceedings of the 2011 American Control Conference, 2011, pp. 1051–1056.