Analysis of the Structure and Function of Phono-semantic Compounds Based on Complex Networks

Jianyu Li\(^1\), Jie Zhou\(^2\), and Xiaowen Mao\(^3\)

\(^1\)Engineering center of Digital Audio and Video, Communication university of China, Beijing, 100024 China
\(^2\)Department of Automation, Tsinghua University, Beijing, 100084 China
\(^3\)School of Computer Science, Communication university of China, Beijing, 100024 China

Abstract—Complex network has been used in phono-semantic compounds in this paper, especially the degree correlation, hierarchical structure, diversity of the function of nodes and their relationships has been studied. We found that high degree nodes tend to be connected to low degree nodes, displaying a disassortative mixing property in the whole; some nodes can act both as phonetic radical and semantic radical, leading to a more complicated structure; and the hierarchical structure makes the disassortative mixing property more obvious. This paper reveals that phono-semantic compounds are optimal combination of phonetic radicals with semantic radicals, and they bear the merits of easy remembering, easy spreading and easy understanding. The phono-semantic compounds follow the principle of least effort [1] and characteristics of human cognitive mechanism.

Keywords: Disassortative Mixing, Hierarchical Structure, Complex Network, Phono-semantic Compounds, Degree Correlation

1. Introduction

There are various complex systems, whose internal members interact with each other to reach a dynamical balance. This phenomenon is common in nature and human society, to name a few, brain structures, food chain networks, the Internet and WWW. The properties of those complex systems could be described by theories of complex network. A node is utilized to demonstrate an entity and an edge to depict their connection. Nodes and edges connectively form the entire network. Recently, quite a lot of work have been carried out making a more detailed analysis of the structure of complex networks[2][3][4]. People find that social interactions[5][6][7][8] and technology structures[9][10] characterize the assortative mixing, others such as brain networks[11], food chain networks, cell networks[12] and online social networks(OSN)[13] reflect disassortative mixing. Besides, some networks, such as metabolic network[14] and Chinese character network[15], even have hierarchical structure.

Chinese characters are carriers of Chinese culture and visualization of Chinese way of thinking. Study their formation and construction are very meaningful. Research of Chinese characters could disclose the behavior, lifestyles and ideology of ancient Chinese. For example, the existence of so many derogatory characters constructed with “女” (e.g., “奸”, “妓”, “妓”, etc.) in Chinese characters reflect the thousands-year-old traditional ideology of patriarchy. When ancient Chinese created characters, usually they constructed them based on the objects’ appearances intuitively. These objects are common in people’s daily life such as “氵” (water), “木” (grass), “木” (wood). That’s why a great deal of characters were constructed with the radicals “氵”, “木”, “木” etc.

Quite a lot of contribution has been made in analysis of the Chinese language using complex networks theory. Li[16], Yamamoto[17], and Wang[18] built phrase related networks which display scale free and small world features. In [16], the nodes are meaningful words which may be single Chinese character or multi-character combination. Jianyu Li[15][19] found that Chinese phrase networks display some important features: not only small world and the power-law distribution, but also hierarchical structure and disassortative mixing. But in the area of Chinese characters not many results are yet reported. More than 80 percents of Chinese characters is phono-semantic character. Phono-semantic compounds consist of phonetic radicals and semantic radicals, which could be treated as nodes in our network. This paper mainly investigates the degree correlation, diversity of the function of nodes, hierarchical structure and their relationship in phono-semantic compounds network.

2. Data preparation

2.1 Concepts

Complex network is different from regular network and random network, because it has some special features, such as small world and scale free, etc. Recently, people studied more properties of complex network, for example, degree distribution, average path length (APL for short) and clustering coefficient, etc.

The degree of a node in a network is the number of connections or edges the node has to other nodes. If one node has a high degree, and widely connected, it must be good at composing characters in Chinese characters structure network. The degree distribution is very important in studying both real world networks, such as the Internet.
and social networks, and theoretical networks. The degree distribution $P(k)$ of a network is then defined to be the fraction of nodes in the network with degree $k$. Thus if there are $N$ nodes in total in a network and $N_k$ of them have degree $k$, we have

$$P(k) = \frac{N_k}{N}. \quad (1)$$

Average path length is the average shortest distance of two nodes.

$$APL = \frac{1}{N(N-1)} \sum_{i \neq j \in V} d_{ij}(ij), \quad (2)$$

where $d_{ij}$ is the shortest distance of node $i$ and node $j$. The correlation is characterized by the

$$C_i = \frac{2e_i}{k_i(k_i - 1)} \quad (3)$$

and defined as the rate of edges $e_i$ actually connected with node $i$ and all links over $i$, where $k_i$ is the degree of $i$. And average clustering coefficient of the whole network is

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i. \quad (4)$$

### 2.2 Network construction

Semantic radicals and phonetic radicals are treated as nodes to construct a phono-semantic compounds network. If two nodes could construct a Chinese character, that means they are relevant. According to Xinhua Dictionary, the most widely used dictionary in China, we divided 4000 Chinese characters into 211 semantic radicals and 1084 phonetic radicals. We construct the networks in the following senses.

- Two nodes are connected if relevant.
- Self-connections are ignored.
- Variants (such as 匪, 匪) of a semantic radical are treated as two different nodes.
- Node which can be used as both semantic radical and phonetic radical (e.g. "木" is semantic radical in "木", and it is phonetic radical in "沐"), will be treated as one single node.

Based on above, we construct a phono-semantic compounds network:

![Fig. 1: Part of phono-semantic compounds network](image)

### 3. Analysis

#### 3.1 Degree Correlation

There are several correlation coefficients, often denoted $r$, measuring the degree of correlation. Pastor-Satorras[20], describe degree correlation using a straightforward method, which calculate degree of nodes connected to node that degree is $k$, and the result is a function of $k$. If high degree nodes tend to be connected to low degree nodes, we call the network to have a disassortative mixing pattern, and when high degree nodes tend to link to other high degree nodes, it appears assortative mixing feature. If the network has an assortative mixing pattern, the value of function is increase by degrees, vice versa. An uncorrelated network exhibits the neutral degree-mixing pattern whose function is constant. After that, Newman[21][5] simplified the method of calculating degree correlation. The correlation could be characterized by the assortativity $r(-1) \leq r \leq 1$ and defined as the Pearson correlation coefficient:

$$r = \frac{M^{-1} \sum_{i,j} j_i k_i - [M^{-1} \sum_i \frac{1}{2} (j_i + k_i)]^2}{M^{-1} \sum_i \frac{1}{2} (j_i^2 + k_i^2) - [M^{-1} \sum_i \frac{1}{2} (j_i + k_i)]^2}, \quad (5)$$

where $j_i$ and $k_i$ are the remaining degrees at the two ends of an edge and $M$ represents the number of all links. $r$ range from -1 to 1. When $r > 0$, we call the network to have an assortative mixing pattern, and when $r < 0$, disassortative mixing. An uncorrelated network exhibits the neutral degree-mixing pattern whose $r = 0$.

According to statistics, there are 1295 nodes in our phono-semantic compounds network, and the average number of links between semantic radicals is 18.8768, while that between phonetic radicals is only 3.6744. Obviously, semantic radicals could construct more Chinese characters than phonetic ones, which quite directly leads to the observation that semantic nodes have higher degree. The correlation coefficient is $r = -0.4155$, which is negative. Table 1 lists many concrete examples in real social network and TechNet, etc.
Table 1: Degree assortativity coefficients of various kinds of networks. $N$ indicates the number of nodes, $r$ is degree assortativity coefficient[15][21].

<table>
<thead>
<tr>
<th>Network</th>
<th>$N$</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>physics coauthorship</td>
<td>52909</td>
<td>0.363</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>1520251</td>
<td>0.127</td>
</tr>
<tr>
<td>mathematics coauthorship</td>
<td>253339</td>
<td>0.120</td>
</tr>
<tr>
<td>Film actor collaborations</td>
<td>449913</td>
<td>0.208</td>
</tr>
<tr>
<td>Chinese character network</td>
<td>4892</td>
<td>$-0.4097$</td>
</tr>
<tr>
<td>Chinese phrase network</td>
<td>4858</td>
<td>$-0.0645$</td>
</tr>
</tbody>
</table>

Compare with social network, phono-semantic compounds network is similar to Chinese phrase network [15]. It is disassortative mixing, which means the high degree nodes tend to be connected to low degree nodes. Generally speaking, semantic radicals are nodes of high degree and phonetic ones are low degree nodes. Disassortative mixing in this paper means high degree semantic radicals prefer to be linked to low degree phonetic radicals. The semantic radical suggests only a general category of meaning of the compound character and it does not provide a specific meaning or definition. Their degrees tend to be higher. But the phonetic couldn’t be higher, otherwise it could be a big challenge for people to remember so many characters with the same pronunciation. Therefore the disassortative mixing tendency in constructing Chinese characters follows the principle of least effort [1]. The constructed characters are easy to be recognized by their semantic radicals and their low degree phonetic radicals.

3.2 Diversity of nodes’ function

Table 2 lists the top ten of characters construction capability. Study reveals that the characters constructed by the same phonetic radicals may not pronounce uniquely (22 characters constructed by “肖” have 6 pronunciations).

Table 2: Top ten of the highest characters construction of nodes in phono-semantic compounds network. The number of characters these nodes can construct has been given in brackets.

<table>
<thead>
<tr>
<th>radicls</th>
<th>Top ten</th>
</tr>
</thead>
<tbody>
<tr>
<td>semantic radicals</td>
<td>$\hat{\dagger}$(270), $\hat{\top}$(202), $\hat{\dagger}$(191), $\hat{\top}$(168), $\hat{\top}$(162), $\hat{\dagger}$(151), $\hat{\top}$(134), $\hat{\top}$(108), $\hat{\top}$(102)</td>
</tr>
<tr>
<td>phonetic radicals</td>
<td>$\hat{\dagger}$(22), $\top$(21), $\hat{\top}$(21), $\hat{\top}$(20), $\hat{\top}$(20), $\hat{\dagger}$(19), $\hat{\top}$(18), $\hat{\top}$(18), $\hat{\top}$(17)</td>
</tr>
</tbody>
</table>

This phenomenon not only exists in high degree phonetic nodes, but also in low degree ones. We divided this case into four categories:

- Character with the same pronunciation as their phonetic radical (俏xiāo and 肖xiāo e.g.);
- Character with pronunciation differs only in tone from their phonetic radicals (such as 哼fēi and 喉fēi);
- Characters with pronunciation differs only in consonants or vowels from their phonetic radicals (俏xiāo, 悲bēi, 冷lèng e.g.);
- Characters with pronunciation differs both in consonants and vowels from their phonetic radicals (צרצציו, 排pái e.g.).

According to statistics, Type.1 accounts for about 38.5%, Type.2 accounts for 17.6%, Type.3 accounts for 26.3%, Type.4 accounts for 17.6%. Evidently, it is extremely common that characters with the same phonetic radical pronounce differently. In fact, such differentiation is good for understanding and memorization. For instance, people can hardly understand what are you talking about if 22 characters constructed by “肖” all pronounce xiāo. Beyond that, one semantic radical may have multiple meanings too. For example, “日” could construct a large number of characters as semantic radical. “明”, “晴”, “昼”, “晴”, “晴”, “明” and “照” all have meaning of bright, while “晚”, “眠”, “暗” and “昏” have meaning of dusk. However, they are all relevant to light. So instead of indicating the accurate meaning of characters, semantic radicals can only show a general concept.

In phono-semantic compounds network, we also found some nodes which can be used as both semantic radicals and phonetic radicals (such as “木” is a semantic radical in “枱”, but it is a phonetic radical in “枱”). There are 119 this kind of nodes approximately. It accounts for 56.4% of the semantic nodes, accounts for 11.0% of the phonetic nodes and accounts for 9.3% of the whole network nodes. Almost half of semantic radicals are this kind of nodes and they play a very important role in our network. Because the average number of characters which are connected to them is 21.9160, while semantic and phonetic nodes’ is only 18.8768.
3.6744. That’s why the semantic radicals can construct more characters than phonetic ones. This kind of feature not only makes phono-semantic compounds network more complex, but also makes nodes’ function more various.

3.3 Hierarchical structure

As we know, several concepts are proposed to measure the hierarchy in a network, such as the hierarchical path[22], the scaling law for the clustering coefficients of the nodes[23], etc. These measures can tell us the existence and the extent of hierarchy in a network.

In this paper our networks are multilevel. Like Fig.4, simple components are used to form complicated components (phono-semantic characters), then the complicated components are used to form more complicated ones. For example,  

"┣" and "┫" are used to construct "('{{}}" and "('{{}}" are used to construct "('{{}}. We call the phono-semantic characters such as "┣" and "┫", the "mid-components" provisionally. "Mid-components", 194 in network, are all phonetic radicals, which tend to be connected to semantic radicals. They generally have low power on phono-semantic compounds construction, the average number of characters they can construct is only $2.4536$. Compare with simple components, the mid-components have a tendency of stability. The more complicated the components are, the less characters they can construct. Because too complicated structure of Chinese characters is not convenient for writing.

<table>
<thead>
<tr>
<th>Table 3: Average degrees of different radicals.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td></td>
</tr>
<tr>
<td>Phonetic radicals(total)</td>
<td>3.6744</td>
</tr>
<tr>
<td>Phonetic radicals(except &quot;mid-components&quot;)</td>
<td>4.7072</td>
</tr>
<tr>
<td>Semantic radicals</td>
<td>18.8768</td>
</tr>
<tr>
<td>&quot;Mid-components&quot;</td>
<td>2.4536</td>
</tr>
</tbody>
</table>

Fig. 4: Illustrating how a “mid-component” is constructed and how a “mid-component” constructs a new phono-semantic character

The intrinsic hierarchy can be characterized in a quantitative manner using the recent finding of Drogovtsev, Goltsev, and Mendes[24]. If the clustering coefficient of a node with $k$ links follows the scaling law, like

$$C(k) \sim k^{-1},$$

so the network have hierarchical structure. We found our phono-semantic compounds network have following characteristics like Fig.5.

Fig. 5: Relationship between degree and clustering coefficient of nodes

We can see the phono-semantic compounds network has the tendency of hierarchy. The Chinese ancients reuse the preexisting simple characters to construct the new ones. It reduces the burden of the memory of people and improves the efficiency of creating characters.

4. Conclusion

We have already known that phono-semantic compounds network has diversity of the function of nodes. It is also hierarchical and disassortative mixing. Some nodes can be
used as both phonetic radicals and semantic radicals. Thus the network is more complicated. Moreover, some “simple components” are used to form “complicated components” and the “complicated components” are relatively harder to combine with other components than the simple ones. So their degree is relatively low. The emergence of these nodes with low degree makes the disassortative mixing of the network more obvious.

We use complex network to analyze phono-semantic characters. We can not only learn extensive and profound Chinese culture better, but also understand the unique and creative formation of Chinese characters from ancient Chinese. Moreover, it will help us devise more interesting Chinese character learning course. And based on the differentiation of different radicals’ degree, we can improve the layout of “Five-stroke Coding System”, a well known character input method.

References


