Statistical Properties of Chinese Near-synonyms Network

Jinze Li
School of Computer Science
Communication university of China
Beijing, China
lijinze@cuc.edu.cn

Shuangwen Chen
Information Engineering School
Communication university of China
Beijing, China
csw@cuc.edu.cn

Xiaoyue Luo
Department of Mathematics
Linfield College
OR 97128, USA
xluo@linfield.edu.cn

Feng Xiao
Department of Automation
Tsinghua University
Beijing, China
xiaof99@mails.tsinghua.edu.cn

Jianyu Li
Engineering center of Digital Audio and Video
Communication university of China
Beijing, China
lijianyu@tsinghua.edu.cn

Hao Shen
School of Television and Journalism
Communication university of China
Beijing, China
shenhao@cuc.edu.cn

***Abstract*—We investigate Chinese near-synonyms (include synonyms and near-synonyms) based on complex network theory. In the study of near-synonyms network, scale-free effect and hierarchical structure features are found in this complex system. In particular, we find that n-node cliques exist widely in near-synonyms network. N-node clique structure is a powerful tool for understanding global structures of complex networks. Through the analysis of n-node cliques, near-synonyms are clustered in clique style. The structures properties of n-node cliques in a way may explain the associative mechanism and information storage in human brain.**

Keywords-complex network; near-synonym network; scale-free; hierachical; n-node clique

I. INTRODUCTION

Any language in the world, including Chinese, makes sentences and phrases by connecting basic units based on complex grammar, syntax and semantics [1]. Recently, with the rapid development of complex networks studies, language complex networks are actively investigated [2] [3] [4]. For instance, word co-occurrence networks [5], semantic networks [6] and synonyms network [7] display two important features: small-word effect and scale-free distribution. Furthermore, study of Wordnet lexicon [8] demonstrates that Wordnet has global properties common to many self-organized systems, and polysemous links have a profound impact in the organization of the semantic graph which may be crucial for metaphoric thinking, imagery, and generalization. Again, network of free word associations [9] represents a proxy of the way in which our mind stores and organizes all words and related meanings.

Characters are treated as the basic units of Chinese phrases. They are monosyllabic, square-shaped and primitive, having some relationship to iconicity and combination [10]. They combine into phrases and sentences based on meanings and syntactic rules.

In the current paper we investigate the statistics and organization of Chinese near-synonyms network. Near-synonyms are words with the same or similar meanings. We define the single characters correspond to nodes of near-synonyms network, and an undirected link exists between characters if they can form a phrase.

The paper is organized as follows. First, we focus on the graph properties of near-synonyms network based on complex network theory. Next we present an analysis aimed at statistical properties of n-node cliques existed in near-synonyms network. The n-node cliques are very valuable since they help us to understand the global structures [11]. Finally, we draw the conclusions in last section.

II. DATA AND NETWORK

Near-synonymy is a common semantic relationship in our daily life. Here we mainly focus on the Chinese two-character near-synonyms because most of Chinese near-synonyms are two-character. The experimental data we analyze are collected from [12], and consists of 3446 characters and 12621 two-character phrases. We construct Chinese near-synonyms network in the following way: (1) characters are served as nodes; (2) connections exist between two characters if they can form a phrase. All the characters are obtained from near-synonyms lexicon [12].

In the following we attempt to uncover the statistical properties of Chinese near-synonyms network based on

graph theory. Let us consider the undirected graph of near-synonyms, $G = (W, E)$, where $W = \{w_i\}$, $(i = 1, 2, \ldots, N_L)$ is the set of nodes and $E = \{\{w_i, w_j\}\}$ is the set of connections between characters. Here, $\xi_{ij} = \{w_i, w_j\}$ indicates that there is an edge between characters w_i and w_j.

III. SCALE-FREE AND HIERARCHICAL FEATURE

In this section, we will investigate properties of the resulting network. First, we research the degree distribution $P(k)$. The degree of a given character is the number of edges that connect the given character with other characters. Degree distribution is defined as the existence probability of nodes with degree k. Reflected in Chinese near-synonyms network, degree distribution means the word-formation ability of characters.

Fig. 1 shows that degree distribution of Chinese near-synonyms network follows power-law distribution, $P(k) \sim k^{-\gamma}$. The exponent γ is 1.92. Power-law distribution is often referred to as scale-free structures. Scale-free network indicates that the majority of the nodes have a small amount of links, but a few nodes called hubs, can link to the most of nodes in the network. For instance, in Chinese near-synonyms network, character "心" (heart), "人" (person), "不" (no) are highly connected nodes because they are familiar to us and they can form a great deal of Chinese phrases. On the contrary, characters like "鞠" (bow), "饪" (cook), "锄" (hoe) have low degrees, as they are less common in our life and their word-formation abilities are weak.

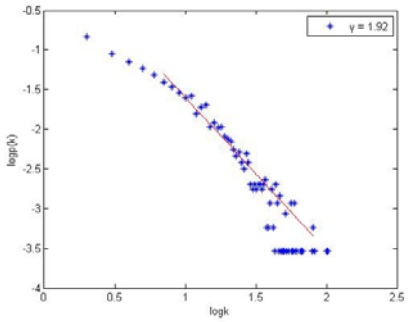

Figure 1. *Logk-logp(k)* degree distribution of near-synonyms network.

Another parameter is clustering coefficient. Clustering coefficient means the probability that two neighbors of a node are also neighbors to each other (node w_i and w_j are neighbors if there is a link between w_i and w_j). For a node w_i with k_i neighbors, the local clustering coefficient C_i is defined as the ratio between the number of links among the k_i neighbors and the maximum possible number of links among these neighbors. This can be expressed as (1)

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

where e_i is the number of existing links between the k_i neighbors.

Clustering spectrum, $C(k)$, is defined as an average clustering coefficient of nodes with degree k. As Fig. 2 shows, clustering coefficient decreases linearly with the degree. This implies that the small nodes are part of highly cohesive, densely interlinked clusters, while the hubs are not, as their neighbors have a small chance of linking to each other. The power-law clustering spectrum $C(k) \sim k^{-\alpha}$ suggests that the network has a hierarchical feature [13].

The hierarchical feature of Chinese near-synonyms network is consistent with hierarchical network models in lexical networks [14]. As shown in Fig. 3, the most common and important characters, such as "心" (heart), "放" (put), "关" (close), "居" (house), should be stored in higher level so that people can learn languages conveniently and efficiently.

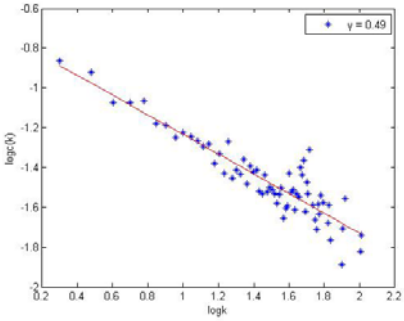

Figure 2. *Logk-logc(k) clustering distribution of near-synonyms network.*

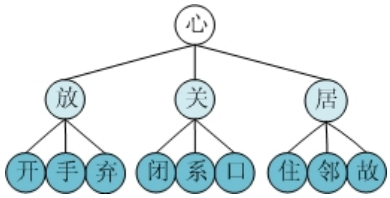

Figure 3. *Hierarchical structure of near-synonyms network.*

IV. N-NODE CLIQUE

In recent years the cliques are actively investigated because of provisions of important insights to information processing, hierarchical modularity, and community structures. Cliques are highly-interconnected subgraphs (complete graphs). A maximal clique is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique. N-node clique in an undirected graph is a maximal clique with n-nodes, and every two nodes in the cliques are connected by an edge, corresponding to n-node complete graph. Research shows that a lot of n-node cliques exist in Chinese near-synonyms network. The network includes 10 1-node cliques (reduplicated phrases such as "哒哒" (tick), "孜孜" (diligent)), 7230 2-node cliques, 2449 3-node cliques and 93 4-node cliques. We mainly focus on 3-node cliques and 4-node cliques due to the appropriate sizes.

A. 3-node clique

In Chinese near-synonyms network, the 3-node cliques consist of 1532 characters and 5402 phrases. The ratio between characters of 3-node cliques and the characters of whole network is 44%, meanwhile the ratio between phrases included in 3-node cliques and the total number is 43%. The statistics show that 3-node clique structures are common in Chinese near-synonyms network.

Fig. 4 shows the relationship between degrees and frequencies of characters in 3-node cliques. We can find that highly connected nodes can compose more 3-node cliques, since the characters with high degrees have strong power to form phrases.

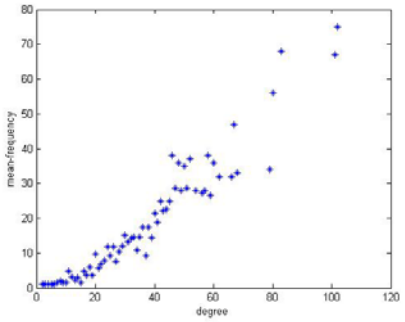

Figure 4. *Relationship between degrees and frequencies of characters in 3-node cliques.*

After research of semantic meanings of nodes we find that there exist 3 types of basic structures in 3-node cliques. Fig. 4 gives examples of every basic structure.

1) 2-near-synonymous-characters

In Fig. 5 (a), the clique consists of 3 characters "查" (check), "问" (ask), "询" (inquire), two of which ("问" and "询") share the similar meanings. There are three phrases formed related to the nodes, "查问", "查询" and "询问". Among them, "查问" and "查询" are near-synonyms.

2) 2-antonymous-characters

Fig. 5 (b) shows an interesting structure that the clique contains 2 antonymous characters "胜" (win) and "败" (defeat), and two antonymous phrases "战胜" and "战败". The similar examples are "多-少-最", "大-小-名", "长-短-处", etc. The result reveals that antonymy does exist in near-synonyms networks. Antonymy has a close relationship with synonymy.

3) 3-near-synonymous-characters

In addition to the above, the third type of 3-nodes cliques includes 3 near-synonymous characters. As shown in Fig. 5 (c), "称", "赞" and "誉" all mean praise, "称赞", "称誉" and "赞誉" are near-synonyms as well.

The analysis of 3-node cliques and their structures displays the clustering features of Chinese near-synonyms. The formations of Chinese near-synonyms network have great relationship with semantic clustering. A group of near-synonyms tend to share a character while the other characters are near-synonymous, showing that Chinese near-synonyms are organized in the light of the least effort principle [15]. On the other hand, the result suggests that Chinese characters are possible to construct new phrases if they have similar, opposite and same meanings [16].

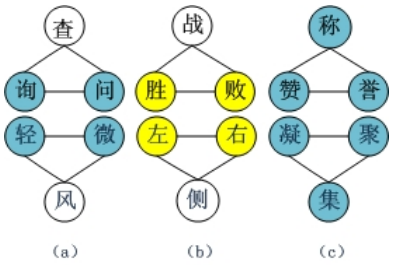

Figure 5. *Basic structures of 3-node cliques.*

B. 4-node Clique

As mentioned, the number of 4-node cliques (93) in Chinese near-synonyms network is far less than the number of 3-node cliques (2449). 4-node cliques contain 219 characters and 470 phrases. That is because cliques are complete graphs, the more nodes included, the more complex the structures are. Complex structure may be a burden for our brain and interferes with the memory storage. The excess of near-synonyms does not follow least effort principle either.

The structure features of 3-node cliques also exist in 4-node cliques. Fig. 6 shows the basic structures of 4-node cliques. Among all 4-node cliques, 52 4-node cliques contain near-synonymous characters and phrases, accounting for 58%. This illustrates that near-synonymy relationships are the base of clustering. Besides, seven 4-node cliques possess antonymy clusters as well as in 3-node cliques. This phenomenon does indicate that antonymy is closely related to synonymy.

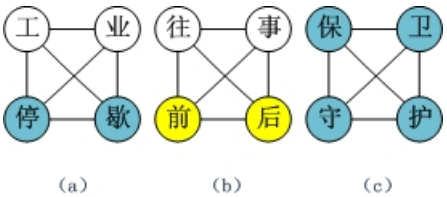

Figure 6. *Basic structures of 4-node clique.*

C. N-clique Community

Actually 4-node cliques are the combination of 3-node cliques. N-node cliques can aggregate into more complicated structures, such as communities. Here a k-clique community is define as the union of all cliques of size k that can be reached through adjacent (sharing k-1 nodes) k-cliques. Fig. 7 and Fig. 8 show the 3-clique communities and 4-clique communities. Small and simple cliques can form complicated structures. The composite pattern may explain the normal processes of human learning and memory storing.

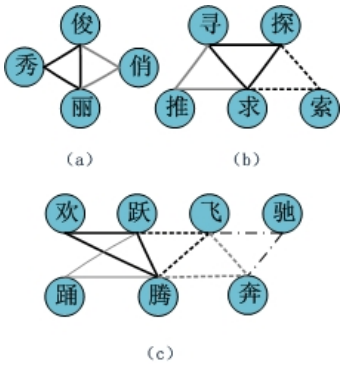

Figure 7. *3-clique communities*

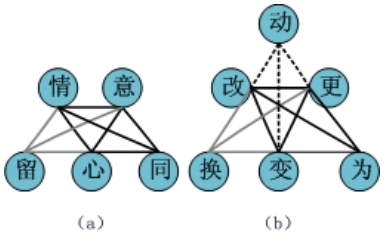

Figure 8. *4-clique communities*

V. DISCUSSION AND CONCLUSION

In this paper, we have presented the results of the analysis performed on Chinese near-synonyms network. As other Chinese phrases network [4] [17], Chinese near-synonyms network displays sale-free and hierarchical structure features, which are responsible for robustness. The phrases in Chinese do not exist in isolation, but put together through certain semantic relationships, such as synonym, near-synonym and antonym aggregation [18]. A group of Chinese near-synonyms tend to sharing same characters in order to communicate efficiently with least effort. The very interesting phenomenon that antonymous nodes emerge in cliques should be further investigated.

The appearance of near-synonyms not only enriched human language but also have important effects on word association. It is well known that information in our brain is associative and retrieved by connecting similar concepts. Our experiment has been brought to the attention of n-node cliques as a valuable tool to understand the basic cognitive mechanisms and information retrieval processes. Similar pieces of information are stored together, due to the high clustering, which makes searching by association possible and efficient. We guess that the storage of 3-node cliques in our brain is stable since triangles are stable structures in nature. N-node cliques can form more complicated structures, like communities. The composite pattern may explain the normal processes of human learning. The structure features of n-node cliques may be related to increasing our memory retention and recall, which is probably necessary for the brain to store information and associate.

ACKNOWLEDGMENTS

The authors would like to thank Prof. Zhou 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] Grimes, B. F., ed., Ethnologue, Languages of the World, Dallas: Summer Institute of Linguistics, 14th ed., 2000.

[2] Ke,Jinyun, “Complex networks and human language”, arXiv: cs/0701135, January 2007.

[3] Jianyu Li, Jie Zhou, “Chinese Character Structure Analysis Based on Complex Networks”, Physica A, vol.380, pp.629-638, 2007.

[4] Jianyu Li, Jie Zhou, Xiaoyue Luo, and Zhanxin Yang, “Chinese lexical networks: the structure, function and formation”, Physica A, vol 391, Issue 21, pp. 5254–5263, November 2012.

[5] Ramon Ferrer i Cancho, and Richard V. Solé, “The small world of human language”, Proc. R. Soc. Lond. B, 268(7), pp: 2261—2265, November 2001.

[6] Mark Steyvers and Joshua B.Tenenbaum, “The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth”, Cognitive Science, vol. 29(1), pp. 41-78, 2005.

[7] Motter, Adilson E., Alessandro P. S. de Moura, Ying-Cheng Lai and Partha Dasgupta, “Topology of the conceptual network of language”, Physical Review E, 65.065102:1-4, 2002.

[8] Mariano Sigman and Guillermo A. Cecchi, “Global organization of the Wordnet lexicon”, Proceedings of the National Academy of Sciences USA, vol.99, no.3:1742–1747, February 5, 2002.

[9] Pietro Gravino, Vito D.P. Servedio, Alain Barrat and Vittorio Loreto, “Complex structures and semantics in free word association”, Advances in Complex Systems 15, 1250054-1, 2012.

[10] Chuan Lu, Linguistics for Knowledge Engineering, Tsinghua University Press, 2010.6.

[11] Takemoto, Kazuhiro, Oosawa, Chikoo and Akutsu, Tatsuya, “Structure of n-clique networks embedded in a complex network”, Physica A, Volume 380, pp. 665-672, July 2007.

[12] Rong Cheng, Dictionary of synonyms, Shanghai Lexicographical Publishing House, October 2010.

[13] Erzsébet Ravasz and Albert-László Barabási, “Hierarchical organization in complex networks”, Physical Review E, vol. 67, Issue 2, id. 026112 (PhRvE Homepage), February 2003.

[14] Allan M. Collins and M. Ross Quillian, “Retrieval time from semantic memory”, Journal of Verbal Learing and Verbal Behavior, vol. 8, Issue 2, April 1969, Pages 240–247.

[15] Ramon Ferrer i Cancho and Ricard V. Sole´, “Least effort and the origins of scaling in human language”, Proceedings of the National Academy of Sciences USA, vol.100, pp.788–791, February 2003.

[16] Jinze Li, Jianyu Li, Zhanxin Yang and Feng Xiao, “Formation and simulation on Chinese lexical networks”, 2012 Fifth International Joint Conference on Computational Sciences and Optimization, pp. 614-618, June 2012.

[17] Yong Li, Luoxai Wei, Yi Niu and Junxun Yin, “Structural organization and scale-free properties in Chinese Phrase Networks”, Chinese Science Bulletin, Vol.50 No.13:1304-1308, July 2005.

[18] Benyi Ge, Modern Chinese Lexicology, Shandong People’s Publishing House, April 2001.