**Social Network analysis of *Romance of the Three Kingdoms***

Duan Bowen and Zhou Zhenman

# Introduction

*Romance of the Three Kingdoms*, written by Luo Guanzhong, is famous as one of the four greatest Chinese novels, which is devoted to the turbulent years in the end of the Han dynasty and the Three Kingdoms period in Chinese history. This book was created on the basis of the historical text *Records of the Three Kingdom*. Compared with its counterpart, *Records of the Three Kingdom* covers historical facts instead of imaginary novel plots, therefore is more objective. Many researches have been done for these two great works.

However, traditional literary research methods rely on intensive reading of texts, which requires much effort for analyzing a novel with more than 1000 characters. Besides, such a qualitative method based on subjective feeling leads to the lack of objectivity.

With the rise of computational linguistics, social network analysis which uses computer tools to analyze text characteristics has become an important approach to literary research. In a social network, the characters of a novel are regarded as nodes, and the relationships between them are described as edges. With the help of the great computing power of computers, the complex relationship between characters in a novel with thousands of words can be clearly presented through visual diagrams. And a lot of hidden information can be mined in an easier way. In addition, the quantitative methods such as calculating node centralities can be applied to analyze the importance of main characters, which makes the research results more objective and convincing.

This project primarily aims to reveal the social network structure of the Chinese novel *Romance of the Three Kingdoms*, and conduct quantitative analyses to verify the existing theories.

# Literature review

Social network analysis uses quantitative methods to study the relationships between nodes in a social network, and visualizes these relationships through computer tools. Many scholars and researchers have explored the ways to apply social network analysis to the field of literary studies.

The research of Alberich, Miro-Julia and Rossello (2002) showed that the social network in Marvel Comics is very similar to the networks in real world.

Stiller and Hudson (2005) revealed that the relationship networks of characters in the ten plays of Shakespeare are in line with the “small-world” model.

Elson, McKeown and Dames (2010) analyzed the social networks of 60 English literary works in nineteenth century. The networks were extracted by looking at the quotation marks which indicate dialogues. The techniques of natural language processing and machine learning were used to determine all possible names of each character and the characters involved in each dialogue. This research studied the properties of each social network, and showed that there are significant regional differences in social interactions in nineteenth century English novels.

Agarwal, Kotalwar and Rambow (2013) explored the technology for accurately extracting social events and social networks from novels. They used their corpus to train support vector machines, and put this method into practice by constructing the social network in the fiction *Alice in Wonderland*.

It is worth pointing out that Stanford University established the Literary Laboratory in 2010, aiming to use computer tools to analyze literary works. Researchers used social network analysis and graph theory to study the plots in the play "Hamlet" and the series of novels "Harry Potter", and conducted empirical researches for existing theories.

The researches mentioned above show that it is completely feasible to apply the approach of social network analysis to studying the structure and relationship between characters in a literary work.

# Methods

The procedure of this project can be divided into collecting data, preprocessing data, constructing social networks and analyzing social networks.

The data for this project are the text of the book *Romance of the three kingdoms*. The text is collected through python from website <http://www.guoxuemeng.com/guoxue/sanguoyanyi/yingwenban/>. These two texts contain over 300,000 words.

Before constructing the social networks, the plain texts should be preprocessed by natural language processing technology. The first step will be tokenization, which splits the sentences into single words. Then the names of each character will be extracted, using named entity recognition. This step is done with the help of the python library Spacy.

With such a dataset containing words and names of characters, we can start building the social networks. In social networks, people are usually viewed as nodes, and their interconnections are viewed as edges. In this project, the edges are determined by the co-occurrences of two characters in the same paragraph. And the weights of each edge are determined by the frequency of co-occurrences. Thus, by counting the frequency of co-occurrences of two characters, we construct weighted undirected social networks.

After the social networks are built, statistical analyses are performed. For example, protagonists can be analyzed via node centralities.

The steps mentioned above are realized with the help of Python, Gephi and other additional computer tools.

**Results**

By applying the methods discussed in the previous section, we build the social network of the novel *Romance of the three kingdoms are built* and then visualize it. Figure 1 shows the whole structure of the constructed network.

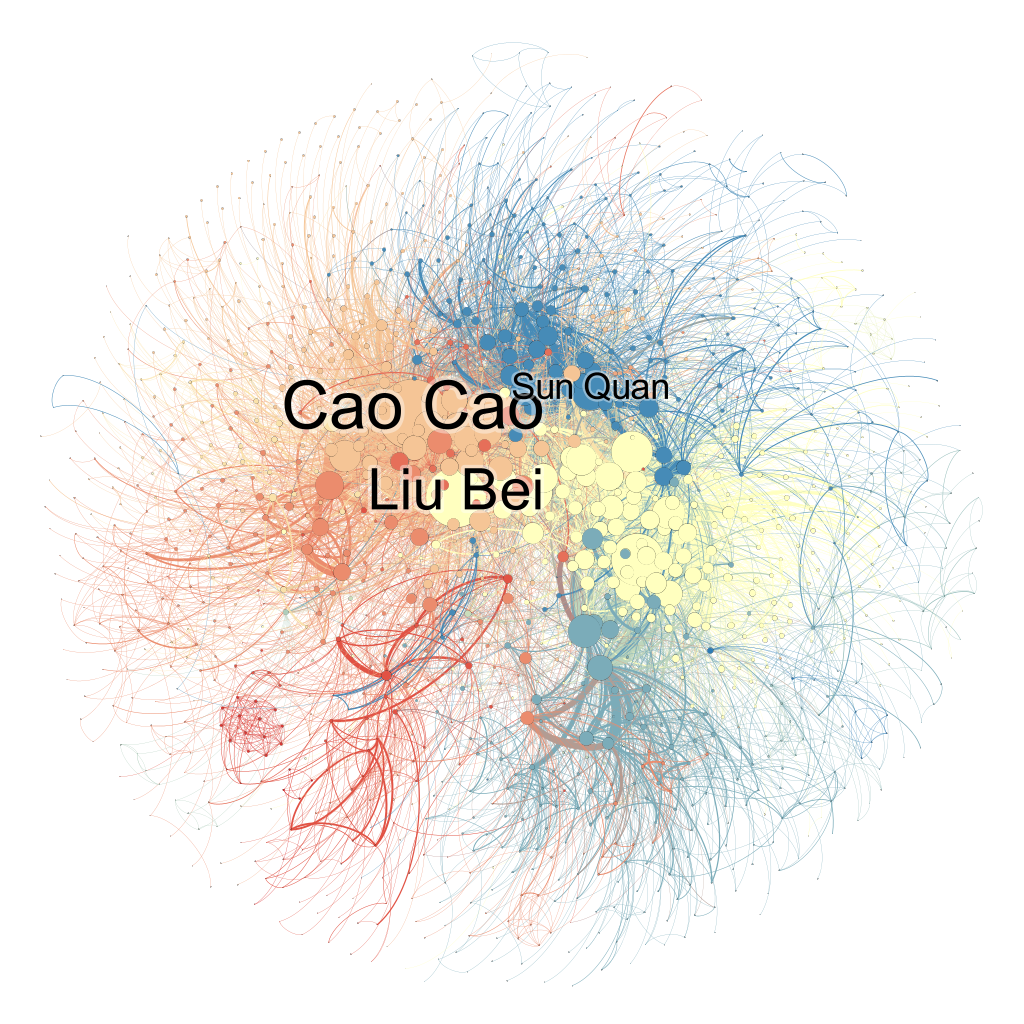


Figure Whole network of Romance of the three kingdoms

## 1 General information of the network

We construct a huge and complex network with 12 connected components. Then the giant connected component is picked up for later analysis. Table 1 shows the general information of the network of *Romance of the three kingdoms.*

|  |  |
| --- | --- |
| Number of nodes | 1358 |
| Number of edges | 8641 |
| Network density | 0.009 |
| Average shortest path length | 3 |
| Diameter | 8 |
| Radius | 4 |

Table 1. General information of the network of *Romance of the three kingdoms*

First of all, this is a rather huge network with 1388 nodes and 8677 edges. In addition, the network density is only 0.009, indicating that the network is very sparse. In fact, these two network features are unusual for a novel. A novel typically focuses on a few main characters, and all other roles are developed based on them to support them. Therefore, the network should be small and dense. However, *Romance of the Three Kingdoms* is a novel based on a certain historical fact. The detailed historical background is bound to expand the range of characters described in the novel, *Romance of the three kingdoms* includes many redundant historical events which are not strongly connected to the main story, which makes the network so big and sparse.

## 2 Comparison of empirical network features

All novels are written on the basis of real world, and therefore networks in novels should have the same features as in the real world. So does *Romance of the three kingdoms.* However, when the authors create novels, they may make some changes on the plots and facts to make the books more appealing for their readers. By comparing the network features of real world and novels, the modifications by the authors are revealed. An empirical network model typically has three features: power-law (heavy-tailed) degree distribution, small average distance and large clustering coefficient.

Figure 2 shows the degree distribution of the constructed network. As we can see, most of nodes have very low degree, and only a few nodes are with high degree. When the node distribution is plotted in a log-log scale, the distribution is arranged as a straight line, which shows that our network has a power-law degree distribution.

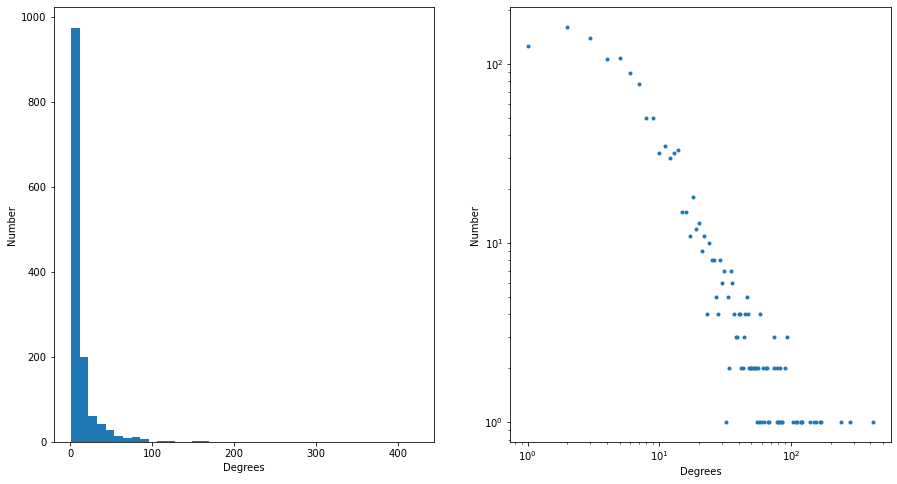


Figure Degree distribution

Figure 3 demonstrates the distance distribution of the constructed network. As shown by this histogram, the average distance lies between 3 and 4. That is a really small number, and therefore the network of *Romance of the three kingdoms* meets the feature of small average distances. In fact, a typical average distance in the real world is around 6. The smaller number of our network indicates the writing technique of the author. Specifically speaking, there are protagonists which connect all the characters. Another assumption is that one character is introduced immediately after another so that the frequency of co-occurrences is high, which causes the small average distance. These two writing techniques make the novel more consistent and coherent.

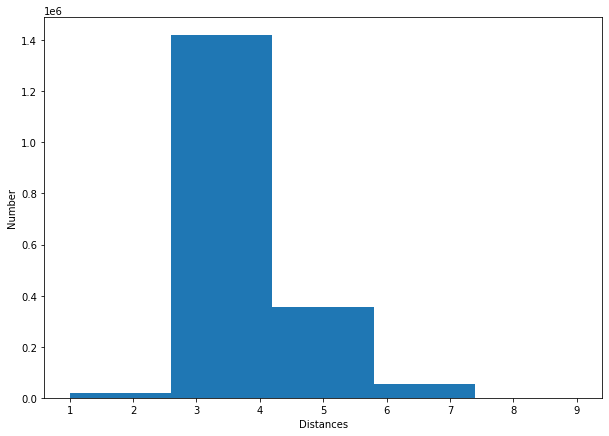


Figure Distance distribution

However, for the third feature, the global clustering coefficient of the network of *Romance of the three kingdoms* is only 0.215. It is obvious that the network does not meet the feature of large clustering coefficient. This fact once again confirms the conclusion discussed above, i.e., as a historical novel, *Romance of the three kingdoms* includes many redundant characters who may have no attribution to plot development. These unnecessary characters are not strongly clustered.

## 3 Analysis for main characters

There are over 1000 characters in the novel *Romance of the three kingdoms*. It is impossible to manually analyze all of them. However, with the great computing power of machines, we now can easily deal with them by social network analysis. The importance of each character can be revealed by the node centralities. The following figures demonstrate the top 20 characters by different node centralities.

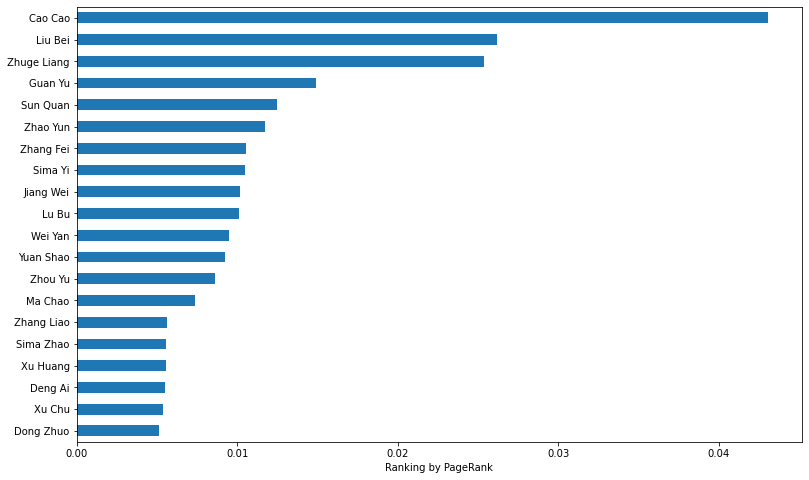


Figure Top 20 characters of ranking by PageRank

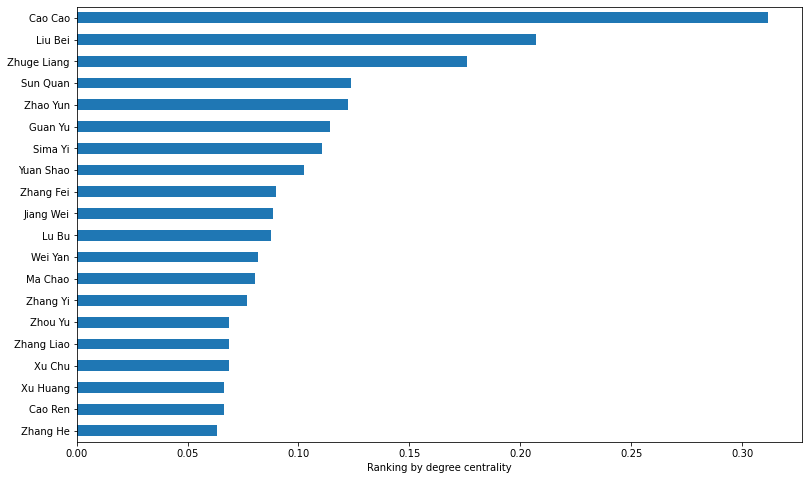


Figure Top 20 characters of ranking by degree centrality

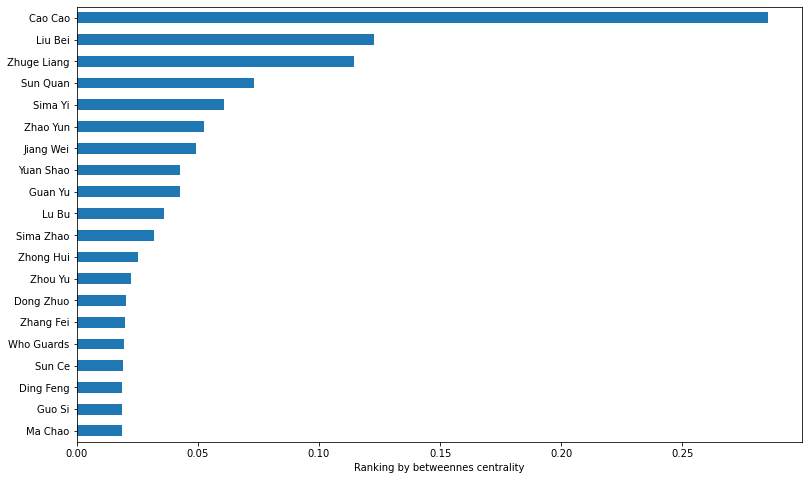


Figure Top 20 characters of ranking by betweenness centrality

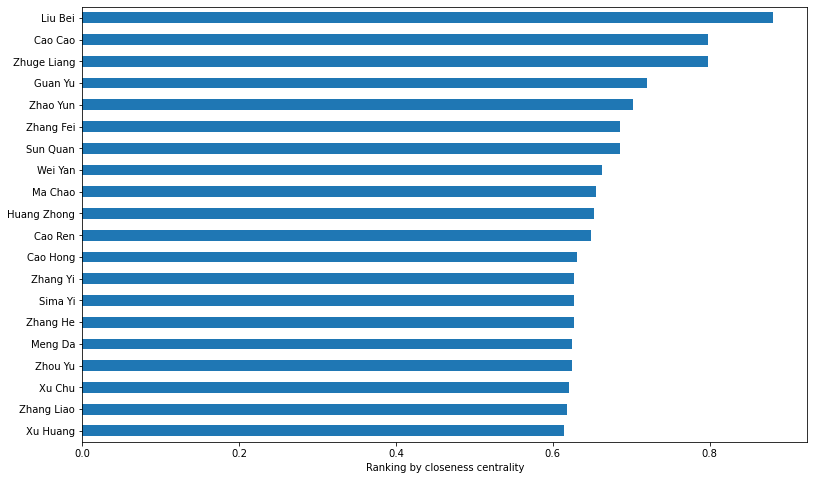


Figure Top 20 characters of ranking by closeness centrality

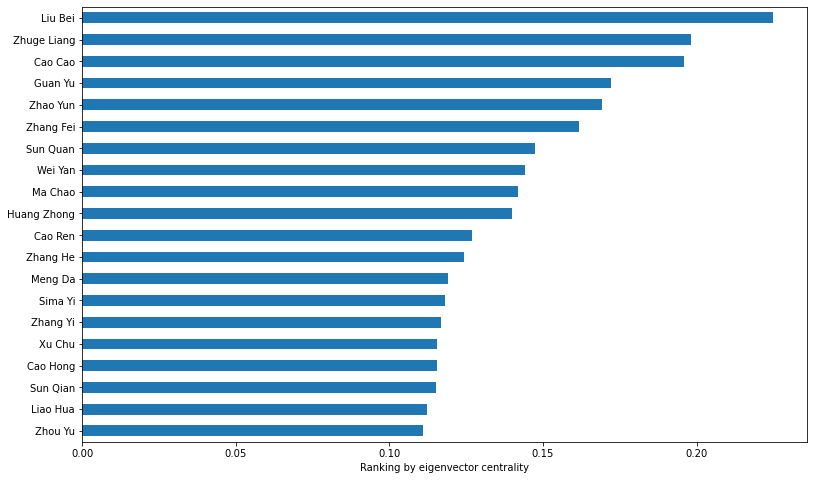


Figure Top 20 characters of ranking by eigenvector centrality

As suggested by the name, there are three countries in the novel *Romance of the three kingdoms*: Kingdom Wei governed by Cao Cao, Kingdom Shu governed by Liu Bei and Kingdom Wu governed by Sun Quan. As indicated by the above figures, Cao Cao is definitely the most important character in this novel, then followed by Liu Bei. Besides, two other interesting facts are also observed.

One is that these three kingdoms are not treated equally by the author, especially Kingdom Wu. Except the ranking by betweenness centrality, there are only 1 or 2 characters from Wu enter the lists of top 20 by each centrality. Moat position in the rankings are occupied by characters from Wei and Shu. And rating of the leader of Wu, Sun Quan, is never higher than that of Cao Cao and Liu Bei. In fact, Kingdom Wu was the second strongest from the three, and Shu was the weakest. It can be assumed that the author purposely paid less attention to Kingdom Wu. This may be related to the historical period of the author. The author of the novel, Luo Guanzhong, was from the late Yuan and early Ming Dynasty. It was the period when the regime changed and the people yearned for stability. Therefore, Luo Guanzhong may have been influenced by the popular preference of the Shu regime at that time. Because among the three regimes at that time, although the power of Shu was relatively weak, because its emperor claimed to be a relative of the emperor of the Han Dynasty, Shu had the inheritance and legitimacy that the other two regimes did not have.

Another interesting fact is that the leader of Kingdom Shu, Liu Bei and Zhuge Liang, has highest rating in the raking by eigenvector centrality, while in other rankings, Cao Cao is the first. In a social network, eigenvector centrality is a measure of the influence of a node. Node with high eigenvector centrality always cluster together. The higher rating of characters from Shu by eigenvector centrality indicates that there are actually more important characters in Kingdom Shu, which confirms that the author spent more words on the country. In history, because Wei was the actual power center of the last Han Dynasty, it absorbed the management team from the previous dynasty, and during the war it absorbed advisors and generals from other warlords, such as Xu You from Yuan Shao, Huang Quan from Shu and two generals of Liu Biao, the former chief executive of Jingzhou, also surrendered to Cao Cao. Therefore, according to historical facts, the important figures from Wei, such as Cao Cao, should be more important.

# Conclusion

In this research, we apply natural language processing and social network analysis to study the classical Chinese novel *Romance of the three kingdoms,* and have three findings.

First of all, as a historical novel, *Romance of the three kingdoms* includes many redundant plots and characters which make its network big and sparse.

Secondly, the author arranged the characters in such a way that they are strongly connected, and the average distance is small.

Finally, compared to the real history, the author paid more attention to Kingdom Shu but less attention to Kingdom Wu. This novel actually focuses on Kingdom Shu and Kingdom Wei.

**References**

Alberich, R., Miro-Julia, J., & Rossello, F. (2002). Marvel Universe looks almost like a real social network. *arXiv preprint cond-mat/0202174*.

Agarwal, A., Kotalwar, A., & Rambow, O. (2013, October). Automatic extraction of social networks from literary text: A case study on alice in wonderland. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 1202-1208).

Easley, D., & Kleinberg, J. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge university press.

Elson, D. K., McKeown, K., & Dames, N. J. (2010). Extracting social networks from literary fiction.

Stiller, J., & Hudson, M. (2005). Weak links and scene cliques within the small world of Shakespeare. *Journal of Cultural and Evolutionary Psychology*, *3*(1), 57-73.

刘海燕, & 尹晓虎. (2015). 文学作品中的 “小世界”——菲茨杰拉德小说人物关系网络的实证分析. *统计与信息论坛*, *30*(12), 102-107.