Character Relationship Analysis in *Journey to the West* by Principle Analysis and Multidimensional Scaling

Fu Xiaowen and Li Jiayi

April 2021

1 Abstract

Journey to the West is a great Chinese ancient vernacular novel,one of the four classic novels of Chinese literature. This report is inspired by the article written by Miss Wan Mengting[1] from Peking University focusing on the analysis of *Dream of Red Chamber*. Different from traditional methods, we use statistical methods including PCA and MDS method with the tool of R studio to analysis the correlation between events and characters in this novel. By using Principle components analysis of character-event data, we find that the novel is developed on different small stories of the whole narration, and in each single story, different new characters have been generated. Thus the novel *Journey to the West* has rather loose narration than *Dream of Red Chamber*. Thus when applying the PCA technique, rather small variation has been explained by the principle component. Besides,by using Multidimensional Scaling, we find that the relationship of the characters in the novel have two polarized types. A small amount of main characters are closely related and often appear in the same events. However, most of other characters have loose connection with others, appearing individually in event. The network of characters is more loose than that of *Dream of Red Chamber*, which is corresponding with the result of MDS and also consistent with the result of PCA.

2 Introduction

Journey to the West, written by Cheng'en Wu is one of the Four Great Novels in China, which narrates the story of 81 troubles encounterd by the

[1]Mengting Wan. Character-Event Analysis and Network Analysis in Dream of the Red Chamber.

master Tripitake (' 唐僧') and his three apprentices Monkey King (' 孙悟空'), Pigsy(' 猪八戒') and Sandy (' 沙僧') along the way to learn Buddhist scriptures. With the technology of machine learning, many research apply text extraction and analysis to novels to extract features they focus on. For example, Wan Mengting applied Principal Component Analysis and modularity classification to study the events and characters in *Dream of Red Chamber*.

Many technologies can be applied to text analysis. For instance, Principal Component Analysis (PCA) uses orthogonal transformation to linearly transform the observations of a series of potentially correlated variables, thereby projecting the values of a series of linearly uncorrelated variables. Multidimensional scaling (MDS) is a mean of visualizing the level of similarity of individual cases of a dataset. In our project, Li Jiayi applies Principal Component Analysis to study the events in *Journey to the West*, Fu Xiaowen applies Multidimensional Scaling to study the relationship of characters. Besides, we compare our results to those of *Dream of Red Chamber*, and conclude the differences which accord with the narration styles of the two novels.

3 Overview of Data

In this report, we use the event-characters data of *Journey to the West*. Firstly we want to have a rough overview of the whole data set. There are totally 302 characters and totally 408 events in the novel. By dividing number of events by total chapters (100 chapters), we find out that the expectation value of event in each chapter is around 4. So we just select characters among 302 characters who participate in events larger than the average number 4 to be our main characters. This gives us 76 characters ranking in a decreasing order of the number of events they have participate in. Namely, in the list, all the characters have showed up in at least one chapter,see Table1.

The figure 1 shows that at every level of events, how many characters have participate in. the largest number of event a single character has participated in is 329, by our main hero "Monkey King", (' 孙悟空'). And also by this figure, we see that after around event number 5, the line tend to be flat, which is consistent with the result of average events of 5.

4 Principle Components Analysis of *Journey to the West*

4.1 Basic algorithm of PCA

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar

character	events number	character	events number
孙悟空	329	牛魔王	7
猪八戒	212	地涌夫人	7
唐僧	198	天竺国小王子	7
沙僧	169	天竺国二王子	7
白龙马	81	天竺国三王子	7
观音菩萨	34	文殊菩萨	6
玉皇大帝	27	王母娘娘	6
哪吒	21	南海龙王敖钦	6
土地神	21	十殿阎王	6
木吒. 慧岸	20	黑风山熊罴怪	6
托塔李天王	18	黄袍怪. 奎星	6
如来佛	16	车迟国虎力大仙	6
唐太宗李世民	15	车迟国鹿力大仙	6
太白金星	13	车迟国羊力大仙	6
阿难	12	独角兕	6
伽叶	11	铁扇公主	6
太上老君	11	镇元大仙清风	6
狮驼岭老妖	11	镇元大仙明月	6
南方增长天王. 四大天王	10	普贤菩萨	5
西方广目天王. 四大天王	10	灵吉菩萨	5
北海龙王敖顺	10	南极寿星	5
许逊. 字敬之. 号许旌阳. 四大天师	9	巨灵神	5
邱弘济. 四大天师	9	二郎神	5
东方持国天王. 四大天王	8	镇元大仙	5
北方多闻天王. 四大天王	8	黄风大圣	5
张道陵. 四大天师	8	黄眉大王	5
葛洪. 四大天师	8	盘丝洞蜘蛛精	5
魏徵	8	雷公	5
金角大王	8	高老庄高太公	5
红孩儿	8	车迟国国王	5
狮驼岭二怪	8	雷公电母	5
狮驼岭三怪	8	西梁国女王	5
天竺国老王	8	朱紫国国王	5
西海龙王敖闰	7	朱紫国金圣娘娘	5
山神	7	郡侯上官	5
银角大王	7	豹头山狼头怪妖一	5
通天河鱼怪	7	豹头山狼头怪妖二	5
六耳猕猴	7	黄狮精	5

Table 1: Event number of characters

projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

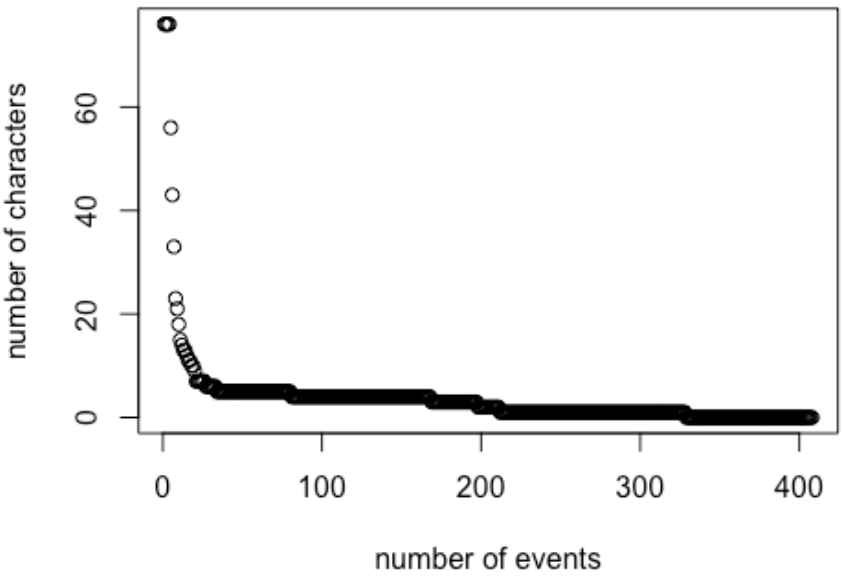

Figure 1: Distribution of events that characters participate in

This means to reduce the dimension of the data to the best k space approximation of data:

$$\text{Let } X = [X_1 \,|X_2| \cdots \mid X_n] \in \mathbb{R}^{p \times n}$$
$$\min_{\beta,\mu,U} I := \sum_{i=1}^{n} \|X_i - (\mu + U\beta_i)\|^2$$

where $U \in \mathbb{R}^{p \times k}, U^T U = I_p$, and $\sum_{i=1}^{n} \beta_i = 0$.

And PCA is given by the top k eigenvectors of covariance matrix $\widehat{\Sigma}_n = \frac{1}{n-1}\tilde{X} \cdot \tilde{X}^T$ $\tilde{X} = XH = X - \frac{1}{n}X \cdot 11^T = \tilde{U}\tilde{S}\tilde{V}^T, \quad H = I - \frac{1}{n}11^T, 1 = (1, \ldots, 1)^T \in \mathbb{R}^n$

In the novel *Journey to the West*, the develop of narration depends on characters' connection and interaction, to further analyze the relationship between story lines and characters we can apply Principle Component Analysis to deeply learn the potential relationships between characters and narration styles of the novel.

In the analysis of P C A, we used the events as observations and characters as variables. However, our experiments shows that every dimension has relatively small variation explained, with first principal component only explained 7.1%

variance and second principal component only explained 5.9% variance and third one with 5.7% variance. To see this result, from figure 2.

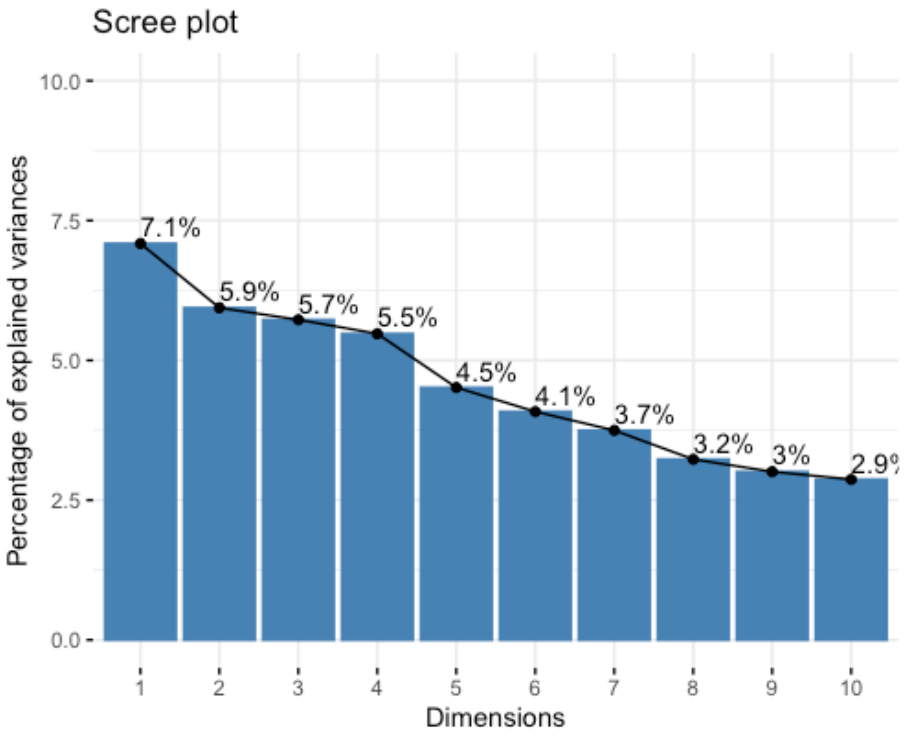

Figure 2: The variance explained by the first 10 principal component

This phenomenon can be explained by that *Journey to the West* is composed by many units of story, which was called as "ninety-nine eighty-one difficult",("九九八十一难"). In every 'difficult', there are new characters showing up and never showing up again in the following narration. So the whole novel is very hard to reduce to a small dimension k.

4.2 Close Analysis of the First Principal Component

According to the PCA analysis, we firstly focus on the first principal component of the novel. We choose the top 10 characters with largest coefficients and the 10 characters with smallest coefficient, see the table 2 below. The left hand side is the characters with largest coefficients and the right hand side is the characters with smallest coefficients.

From this table 2, we observe that the characters with largest coefficients are

	character	coefficient	character	coefficient
1	北方多闻天王. 四大天王	8.38	山神	0.01
2	东方持国天王. 四大天王	8.38	黑风山熊罴怪	0.00
3	南方增长天王. 四大天王	8.11	朱紫国金圣娘娘	0.00
4	西方广目天王. 四大天王	8.03	地涌夫人	0.00
5	伽叶	6.40	牛魔王	0.00
6	阿难	6.30	红孩儿	0.00
7	如来佛	5.35	唐太宗李世民	0.00
8	普贤菩萨	4.31	铁扇公主	0.00
9	文殊菩萨	3.73	黄风大圣	0.00
10	张道陵. 四大天师	2.65	魏徵	0.00

Table 2: coefficients of characters in first principal component

all 'Buddhas'(佛), they are all figures of Chinese Buddhist myth system. This related to the story that Monkey King loses a bet regarding whether he can leap out of the Buddha's hand in a single somersault and afterwards Buddha （' 如来佛祖'）seals Monkey King under a mountain called Five Elements Mountain. So this dimension can be viewed as a story that dominated by the "Buddhas". This component separates "Buddhas" involving in this unit of stories with other characters whose coefficients are even 0.

4.3 Close Analysis of the Second Principal Component

Similarly, we do the same analysis for the Second Principal Component, and present the 10 top characters and 10 tail characters. (See Table 3).

	character	coefficient	character	coefficient
1	天竺国小王子	10.32	高老庄高太公	0.00
2	天竺国二王子	10.32	黄袍怪. 奎星	0.00
3	天竺国三王子	10.32	黄眉大王	0.00
4	天竺国老王	9.31	十殿阎王	0.00
5	车迟国虎力大仙	5.73	白龙马	0.00
6	车迟国鹿力大仙	5.73	西梁国女王	0.00
7	车迟国羊力大仙	5.73	朱紫国国王	0.00
8	豹头山狼头怪妖一	5.51	黄风大圣	0.00
9	豹头山狼头怪妖二	5.51	孙悟空	0.00
10	南海龙王敖钦	4.20	木吒. 慧岸	0.00

Table 3: coefficients of characters in first principal component

From the Table 3, we can see that this component is mainly dominated by the characters related to two stories in the novel. The first one happens in 'Tian Zhu Country', (' 天竺国'), this is according to the chapter 36 and chapter 37.The characters " 天竺国老王" and " 天竺国王子们" are all main characters

in this unit of story. Besides, we can see another key word "Che Chi Country"(' 车迟国'), which indicates the story happens at "Che Chi Country", which is according to the chapter 45 and chapter 46. Although the two characters ' 豹头山狼头怪妖一' and ' 豹头山狼头怪妖二' haven't shown up until chapter 88 and chapter 89, which also happens at 'Tian Zhu Country'. Thus this component can be viewed as a separation of characters related to two locations - 'Tian Zhu Country' and 'Che Chi Country' from other characters.

4.4 Conclusion of the Principal Component Analysis

From the above analysis, we can see that the narration style of *Journey to the West* is developed by narrating a series of small stories, which are independent from each other. The supporting characters of the novel only show up in a single unit of story in the novel and have no contribution to other stories. This is very different from the narration style of *Dream of Red Chamber*, which mainly developed with two main characters "Jia Baoyu"(' 贾宝玉') and "Wang Xifeng' (' 王熙凤') through the whole novel.

5 Multidimensional scaling of *Journey to the West*

There are 76 main characters *Journey to the West* who participate in events larger than the average number 4. To visualize the relationship among the main characters, we consider the numbers of events where a pair of main characters present. There are 2850 pairs of 76 main characters, see Table 4.

character 1	character 2	events number
孙悟空	猪八戒	199
孙悟空	唐僧	177
孙悟空	沙僧	150
孙悟空	白龙马	159
孙悟空	观音菩萨	161

Table 4: Events number where pairs of main characters present

We consider the relationship between two characters is close if the number of events where they both appear is large. With this principal, we can draw connection network with package *igraph* in R, see Figure 3. We can observe that the community network of main characters with igraph seems a little vague, which result from the frequency of co-appearance of two characters. We observe that most of pairs hardly co-appear through the book, and the co-appearance

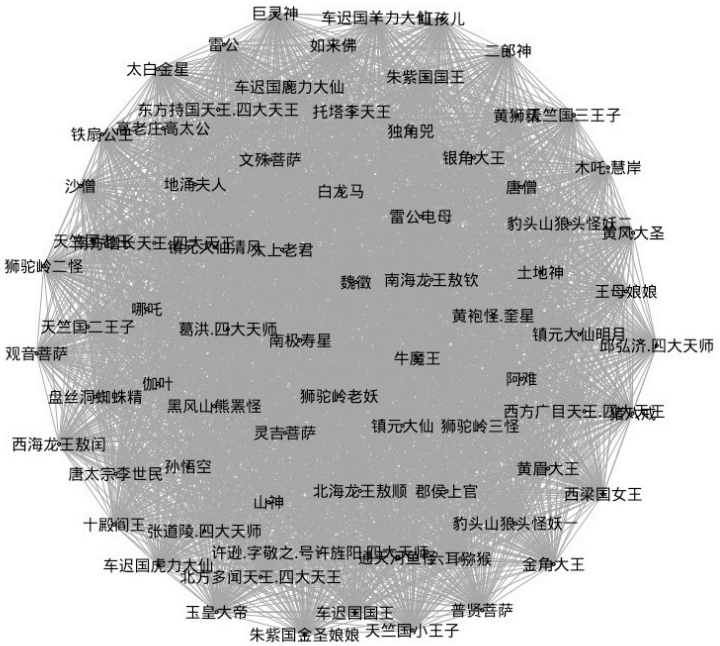

Figure 3: Community network of main characters using igraph

event numbers of some pairs like 孙悟空 and 唐僧 are very large. To solve this problem, we conduct MDS to determine the relationship among characters.

5.1 Basic algorithm of MDS

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. Give a distance matrix which is formed with the distance between respective pairs of groups, MDS places objects into m-dimensional space, which is a lower dimensional space. For example, with $m = 2$, MDS can visualize the relative distances of objects in plane.

MDS uses the fact that the coordinate matrix X can be derived by eigenvalue decomposition from $B = XX^T$, and the matrix B can be computed from proximity matrix D by using double centering. Thus, the MDS is conducted with the following steps.

1. Set up the squared proximity matrix $D^{(2)} = \left[d_{ij}^2\right]$

2. Conduct double centering: $B = -\frac{1}{2}CD^{(2)}C$, where C is the centering matrix $C = I - \frac{1}{n}J_n$, and n is the number of objects, I is the $n \times n$ identity matrix, and J_n is $n \times n$ matrix of all ones.

3. Determine the m largest eigenvalues $\lambda_1, \cdots, \lambda_m$ and corresponding eigen-

vectors $e_1, \cdots, e_m$ of B.

4. The coordinate matrix X is calculated as $X = E_m \Lambda_m^{\frac{1}{2}}$, where $E_m = [e_1, \cdots, e_m]$ is the the matrix of m eigenvectors and $\Lambda_m = diag(\lambda_1, \cdots, \lambda_m)$ is the diagonal matrix of m eigenvalues of B.

5.2 Relationship of characters using MDS

With the event numbers listed in Table 4, we can define a co-appearance event matrix P, see Table 5. As for the distance matrix D in MDS, considering that if the event number is larger, the relationship of respective characters are closer. Thus, we can consider an inverse relationship between co-appearance event matrix P and distance matrix D.

$$d_{ij} = \begin{cases} 0 & i = j \\ 1/0.1 & i \neq j \quad \& \quad p_{ij} = 0 \\ 1/p_{ij} & i \neq j \quad \& \quad p_{ij} > 0 \end{cases}$$

The distance between a character and himself considers to be 0. Since some pairs of characters never appear in the same event, then we consider their number of co-appearance events to be 0.1, and thus their distance to be 1/0.1. The distance matrix D is shown in Table 5.

	孙悟空	唐僧	沙僧	$\cdots$	如来佛	唐太宗李世民	$\cdots$
孙悟空	0	0.005	0.006	$\cdots$	0.09	0.1	$\cdots$
唐僧	0.005	0	0.007	$\cdots$	0.2	0.5	$\cdots$
沙僧	0.006	0.007	0	$\cdots$	0.5	0.33	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
如来佛	0.09	0.2	0.5	$\cdots$	0	0.33	$\cdots$
唐太宗李世民	0.1	0.5	0.33	$\cdots$	0.33	0	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 5: Distance matrix of characters

By MDS, we can get the final result as in Figure 4. We can observe that there is one main community including 孙悟空, 唐僧, 猪八戒 etc. And other characters scatter around the main community. To make the relationship of kernel characters more clear, we apply MDS to 18 characters who participate in events larger than the average number 10, and get the result in Figure 5.

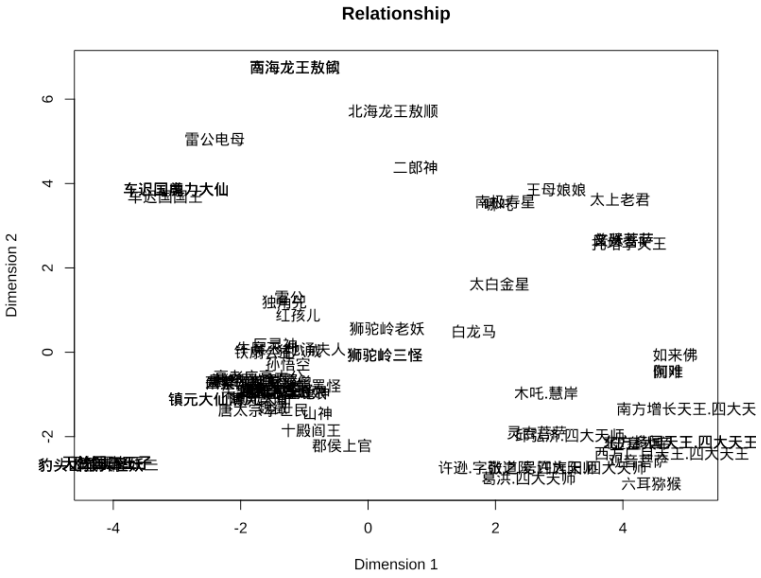

Figure 4: Relationship of main characters using MDS

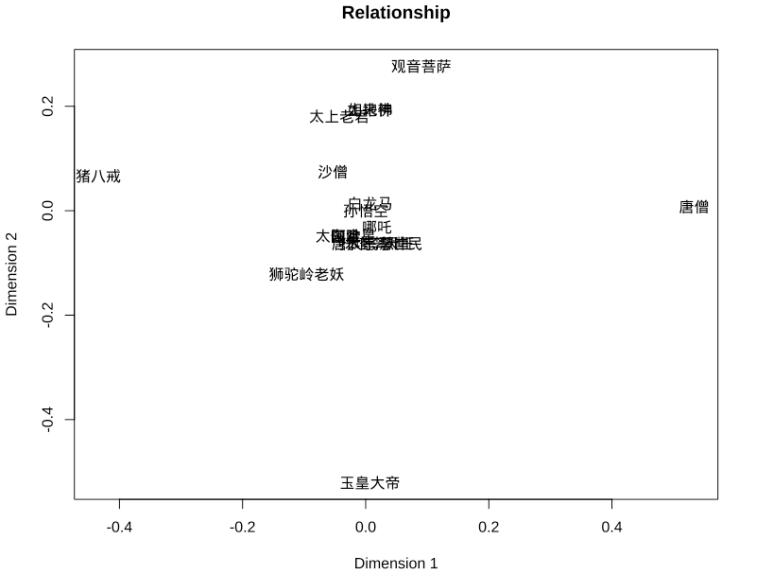

Figure 5: Community network of kernel characters using MDS

5.3 Conclusion of the Multidimensional Scaling

From the above analysis, we can see that the there are two type of characters in *Journey to the West*. A small number of characters are closely related, participating in events throughout the book. And other characters are loosely related to others, only appearing in several events separately. From this aspect, the characters' relationship in *Journey to the West* is also very different from *Dream of Red Chamber*, which have large quantities of kernel characters, and most characters have clear relationship with others.