



本科生毕业论文

题 目 《红楼梦》中的人物事件及人物关系网络分析

英 文 Character-Event Analysis and Network

Analysis in *Dream of the Red Chamber*

姓 名: 万梦婷

学 号: 00946190

院 系: 元培学院

专 业: 统计学

导 师: 姚远

二〇一三年六月

版权声明

任何收存和保管本论文各种版本的单位和个人，未经本论文作者同意，不得将本论文转借他人，亦不得随意复制、抄录、拍照或以任何方式传播。否则，引起有碍作者著作权之问题，将可能承担法律责任。

摘要

《红楼梦》是一部伟大的中国古代白话小说，两百年来人们从未停止对它的研究热情。本文不同于传统的红学研究方法，运用了主成分分析、复杂网络中的社区发现等统计学方法和计算机工具讨论了《红楼梦》的人物事件关系及人物关系网络。本文通过对人物事件矩阵的主成分分析，提出以贾宝玉为中心的情感线索和以王熙凤为中心的事务线索并行发展的观点。并通过人物关系网络的模块化分类印证了上述观点。同时通过对比《红楼梦》和《西游记》的人物关系网络伴随情节发展的变化，指出《红楼梦》是多线索交叉结构。另一方面，其 Q 值伴随情节发展在后40回发生阶梯型减少，也在一定程度上与后40回作者并非曹雪芹的研究结果相符。

关键词：红楼梦，主成分分析，社区发现，模块化

Abstract

Dream of the Red Chamber is a great Chinese ancient vernacular novel, which people never stop studying in last two hundred years. Different from traditional methods, we use statistical methods including PCA and community discovery method in complex networks with the tool of computer to evaluate the correlation between characters and events and detect the character networks. By principal components analysis of character-event data, we find that the novel is developed with the emotional line (Baoyu Jia) and the political line (Xifeng Wang). Meanwhile, we verified this result by detecting communities in character network. Moreover, we point out that *Dream of the Red Chamber* is developed with multi threads compare with *Journey to the West*. On the other hand, the Q value step reduces in the last 40 chapters of the novel, which is consistent with the fact that the last 40 chapters are not written by Xueqing Cao, the original author.

Keywords: Dream of the Red Chamber, PCA, Community Discovery, Modularity

目录

1 介绍	7
2 数据来源与基本符号说明	8
2.1 数据来源	8
2.2 符号说明	8
3 数据基本特征	10
4 利用主成分分析《红楼梦》中的情节线	11
4.1 基本方法	11
4.2 第一主成分与其方向上的情节线	12
4.3 第二主成分与其方向上的情节线	13
4.4 其他主成分与其方向上的情节线	15
4.5 情节线总结	16
5 《红楼梦》中的人物关系网络	19
5.1 基本方法	19
5.2 无权重网络上的社区挖掘	20
5.3 有权重网络上的社区挖掘	24
6 《红楼梦》中的人物关系发展	27
6.1 基本方法	27
6.2 人物数量伴随情节发展的变化	27
6.3 人物关系网络伴随情节发展的变化	28
7 结论	31
参考文献	32
原创性声明和使用授权说明	33

插图

1	人物参与事件数量分布图	10
2	$CharEve_{main}$ 的不同主成分对于方差贡献比例图	13
3	$CharEve_{main}$ 的第三主成分和第四主成分系数图	17
4	《红楼梦》人物情节结构	19
5	《红楼梦》前80回主要人物关系无权重网络社区	21
6	《红楼梦》前80回全部人物关系无权重网络社区	23
7	《红楼梦》前80回主要人物关系有权重网络社区	25
8	《红楼梦》前80回全部人物关系有权重网络社区	26
9	《红楼梦》、《西游记》、随机数据随情节发展的参与事件人 数趋势图	28
10	《红楼梦》、《西游记》、随机数据随情节发展的 Q 值趋势图 .	30
11	《红楼梦》随情节发展的 Q 值增量趋势图	31

表格

1	主要人物参与事件数量表	11
2	第一主成分变量系数表	14
3	第一主成分样本得分表	15
4	第二主成分变量系数表	16
5	第二主成分样本得分表	18
6	《红楼梦》前80回主要人物无权重网络节点分类	22
7	《红楼梦》前80回主要人物有权重网络节点分类	24

1 介绍

曹雪芹先生所著的《红楼梦》，又名《石头记》，是中国古代四大名著之一。由于社会环境所迫，《红楼梦》原稿的部分回目在手抄流传的过程中被遗失，红学界普遍认为现在人们看到的《红楼梦》通行版本仅保存了曹雪芹原稿的80回，而后40回的作者则说法不一。尽管如此，《红楼梦》以其极高的艺术价值和历史价值受到了人们的喜爱。两百年来，人们一直保持着对作者、版本、脂砚斋评语等《红楼梦》相关话题进行研究的热忱，从而形成了今天的“红学”研究。大多数的红学流派，如评点派、题咏派、索隐派、考证派等，都是采用对《红楼梦》的文本解释或者对作者生平进行考证这一类传统的研究方法进行探索。伴随着计算机技术的发展，许多自然科学工作者也将目光投向这部鸿篇巨著，越来越多的人采用统计学的方法，利用计算机技术对《红楼梦》进行分析研究。这类研究的奠基工作由李贤平教授 [1]完成，至今为止的大多数工作仍然是通过对文本中的语言特征（字、词、句、语法特征等）进行收集，利用这些统计特征检验前80回和后40回的作者是否一致 [2]。鉴于此类从文本语言中提取特征判别作者的工作已趋于成熟，而《红楼梦》小说依托精巧的情节设计已构建起了一个庞大的人物关系网络，本文希望通过人物和事件的角度，来解释小说情节的铺成与发展。

近年来，伴随着互联网和新媒体的发展，人们越来越关注社交网络相关的问题。其中一个非常重要的问题便是网络节点的分类问题，对应在本文所探讨的《红楼梦》语境下，就是小说人物的分类问题。在社交网络中，人们把关系较为密切的一群节点称为一个社区或团体(community)。Michele Coscia等人于2011年给出了社交网络中社区发现与分类的方法的综述 [3]，其中的模块化(modularity)方法 [4]得到了非常广泛的应用，并且Newman等人直接将其应用到了Knuth等人给出的《悲惨世界》人物-事件数据集上 [5] [6]。这些工作对于本文都有很大的启发意义。

2 数据来源与基本符号说明

2.1 数据来源

本文的主要数据来源是北京大学2013年秋季的统计学习课程¹。这一数据集是由人工整理出《红楼梦》通行本中第一回到第一百二十回中的所有事件及人物，并且建立起了人物与事件对应矩阵，同时给出人物身份的补充说明。需要说明的是该数据集给出的人物-事件对应矩阵是一个 376×475 的矩阵，但是其第136行和第365行均对应“宝蟾”这一人物，可以合并；第71行和第218行均对应“胡氏”这一人物，可以合并。合并后得到一个 374×475 的人物事件矩阵。

2.2 符号说明

基本符号说明：

- $Ch = \{c_1, c_2, \dots, c_n\}$ 表示所有人物的集合；
- $Ch_{main} = \{c_{i1}, c_{i2}, \dots, c_{ik}\}$ 表示主要人物的集合；
- $Eve = \{e_1, e_2, \dots, e_m\}$ 表示所有事件的集合；
- $Eve_{org} = \{eo_1, eo_2, \dots, eo_{m_0}\}$ 表示前80回所有事件的集合；
- $CharEve_{n \times m} = \{ce_{i,j}\}, i \in Ch, j \in Eve$
表示所有人物事件对应矩阵，其中当第*i*位人物参与第*j*个事件时 $ce_{i,j} = 1$ ，否则 $ce_{i,j} = 0$ ；
- $CharEve_{main} = \{ce_{i,j}^{main}\}, i \in Ch_{main}, j \in Eve$
表示主要人物事件对应矩阵；
- $CharEve_{main,org} = \{ce_{i,j}^{main,org}\}, i \in Ch_{main}, j \in Eve_{org}$
表示主要人物前80回事件对应矩阵。

¹<http://www.math.pku.edu.cn/teachers/yaoy/Spring2013/>

第四节符号说明:

- 记 $\mathbf{X}' = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, 则 \mathbf{X} 表示一个 $n \times p$ 的样本矩阵;
- 记 $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$, 则 \mathbf{A} 表示 \mathbf{X} 的主成分系数矩阵;
- 记 $\mathbf{Z}' = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_3)$, 则 \mathbf{Z} 表示 \mathbf{X} 的主成分得分阵.

第五节符号说明:

- G 表示无权重无向图;
- G^* 表示有权重无向图;
- $Adj = \{A_{ij}\}$ 表示图 G 的连接矩阵, 当 i, j 至少一次参与同一事件时 $A_{ij} = 1$; 否则 $A_{ij} = 0$;
- $Adj^* = \{A_{ij}^*\}$ 表示图 G^* 的连接矩阵, A_{ij}^* 为 i, j 参与同一事件的个数;
- c_i 表示顶点 i 所属社区(community), 若图 G 可以被分为 k 个社区, 则 $c_i \in \{1, 2, \dots, k\}$;
- k_i 表示顶点 i 的度数, 即 $k_i = \sum_j A_{ij}$;
- $\delta(c_i, c_j)$, 当 $c_i = c_j$ 时 $\delta(c_i, c_j) = 1$, 否则 $\delta(c_i, c_j) = 0$
- $Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$ 表示社区模块化指标.

第六节符号说明:

- $\{s_i, i = 1, 2, \dots, n\}$ 表示参与指标为 $1, 2, \dots, i$ 的事件集的人物数量;
- $\{Q_i\}$ 表示由指标为 $1, 2, \dots, i$ 的事件集和参与人物构成的人物关系网络得到最优分割后的社区模块化指标;
- $\{\Delta Q_i\}$ 表示 $\{Q_i\}$ 随 i 的增大的变化情况, $\Delta Q_i = Q_i - Q_{i-1}, i > 1$.

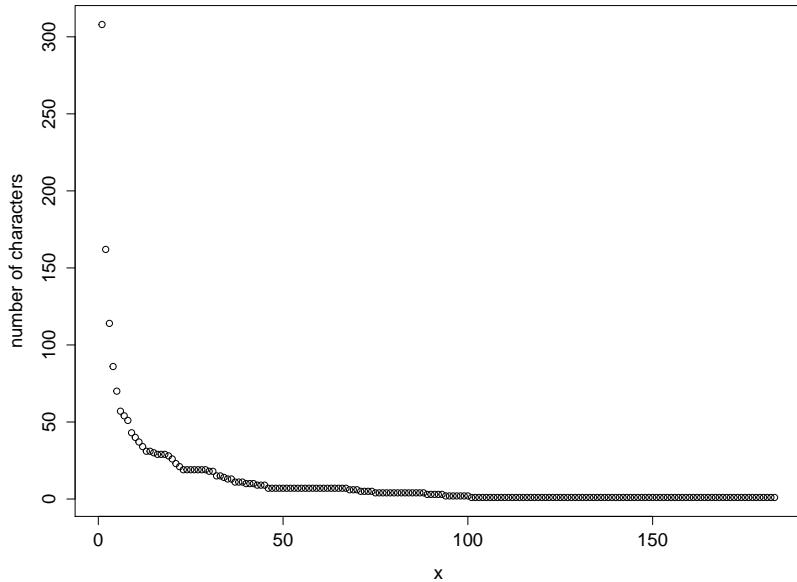


图 1: 人物参与事件数量分布图

横轴表示参与事件数 x , 纵轴表示参与事件数 $> x$ 的人物个数。

3 数据基本特征

原始数据包含374位人物和475个事件, 然而这374位人物中有55位人物没有参与任何事件, 即只有319位人物有意义, 出现这种情况的原因可能是数据收集过程中出现了错误。为了保证主成分分析的合理性, 本节分析承认了前80回与后40回出自不同作者的假设, 因此在后续的分析中只是用前80回的事件数据 Eve_{org} , 即前350个事件。

统计这319位人物参与前80回中事件的数量作为人物对于小说前80回的贡献, 结果如图1所示。根据图1可知, 对于参与事件数大于 x 的人物个数, 当 $x > 4$ 时趋于平滑。另一方面, 对于前80回小说中的350个事件, 可以估计出每一回目包含的事件数期望为 $350/80 \approx 4$ 。对于参与事件数大于4, 即平均参与回目数大于1的人物, 我们选取为主要人物, 其按照参与事件数从多到少的顺序排列如表1所示。

人物	参与事件数量	人物	参与事件数量	人物	参与事件数量
贾宝玉	183	林黛玉	100	王熙凤	93
薛宝钗	88	史太君	74	袭人	70
王夫人	67	贾探春	45	平儿	45
李纨	42	史湘云	39	贾珍	36
贾琏	36	薛姨妈	34	邢夫人	33
贾政	31	贾蓉	31	尤氏	31
晴雯	29	麝月	22	鸳鸯	22
贾惜春	21	薛蟠	21	贾赦	20
贾环	20	贾迎春	20	香菱	19
赵姨娘	19	薛宝琴	18	紫鹃	15
刘姥姥	14	秦氏	12	芳官	12
尤二姐	12	秋纹	11	周瑞家的	11
秦钟	11	贾兰	10	贾芸	10
小红	10	贾元春	9	茗烟	9
林之孝家的	9	贾蔷	8	琥珀	8
莺儿	8	玉钏	8	彩云	8
李贵	8				

表 1: 主要人物参与事件数量表

4 利用主成分分析《红楼梦》中的情节线

4.1 基本方法

对于p维随机向量 $X = (X_1, X_2, \dots, X_p)'$, 称 $Z_i = a'_i X$ 为 X 的第i主成分,
 $i = 1, 2, \dots, p$, 如果:

- (1) $a'_i a_i = 1$;
- (2) $a'_i a_j = 0, i \neq j$;
- (3) $Var(Z_i) = \max_{\alpha' \alpha=1, \alpha' \Sigma a_j=0, j \neq i} Var(\alpha' X)$.

若 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 是从均值为 μ , 协方差矩阵为 Σ 的 p 维总体中抽取的 n 个

独立的样本。其样本均值为 $\bar{\mathbf{x}}$, 样本协方差阵为 \mathbf{S} , 样本相关阵为 \mathbf{R} 。
记 $\mathbf{X}' = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, 假设 \mathbf{X} 已经得到标准化, 考虑线性变换:

$$\mathbf{Z}_i = \mathbf{a}'_i \mathbf{X} \quad i \in 1, 2, \dots, p,$$

记 $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_p)$, 则 \mathbf{R} 的第*i*主成分得分向量 \mathbf{Z}_i 应满足:

- (1) $\mathbf{a}'_i \mathbf{a}_i = 1$;
- (2) $\mathbf{a}'_i \mathbf{a}_j = 0, i \neq j$;
- (3) $\mathbf{Z}'_i \mathbf{Z}_i = (n - 1)\lambda_i$.

令 $\mathbf{Z}' = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_3)$, 则 $\mathbf{Z} = \mathbf{X}\mathbf{A}$ 为 \mathbf{X} 的主成分得分阵。

选取表1中的70个主要人物, 考察其在前80回所有事件下构成的矩阵 $CharEve_{main,org}$, 将前80回所有的事件 Eve_{org} 视为样本, 所有的主要人物 Ch_{main} 视为变量。假设主要人物可以分为若干组团体, 而小说中的事件可以归为若干条情节线, 为了使得小说的内容更丰富更精彩, 这些人物团体需要进行频繁的互动, 而这些情节线既能够独自发展, 又可以进行必要的交汇。为了挖掘出这些人物团体和情节线索, 本文采用了主成分分析的方法。

以下对于样本矩阵 $CharEve_{main}$ 进行主成分分析, 得到的不同的主成分对于方差的贡献比如图2所示。根据图2可以看出, 第一主成分和第二主成分对于方差的贡献较大, 我们在后续研究中会主要讨论这两个主成分。

4.2 第一主成分与其方向上的情节线

考察上一节中得到的第一主成分上各个变量的系数, 其系数最大的10个变量和系数最小的10个变量如表2所示。注意到贾宝玉以及荣国府的女眷在第一主成分中占据非常大的比重, 其影响远远大于有着相反方向系数的贾蓉、秦氏等宁国府众人和刘姥姥。因此第一主成分可以看作贾宝玉和荣国府的主要女眷与其他人物的分离(尤其是与宁国府人物的分离), 而根据阅读经验显示贾宝玉和荣国府的女眷恰是《红楼梦》一书中的主要人物。而对于除此之外的其他主要人物, 以贾蓉、秦氏、贾珍等为代表的

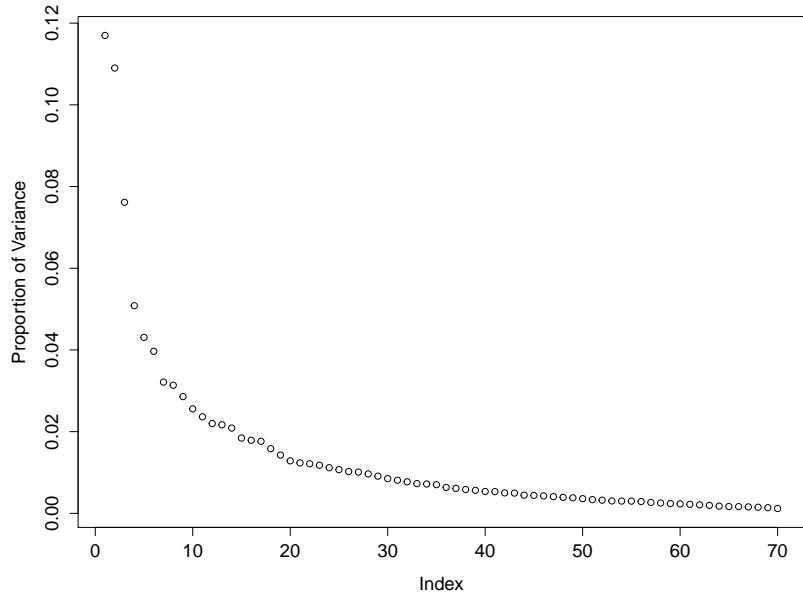


图 2: $CharEve_{main}$ 的不同主成分对于方差贡献比例图

宁国府贵族及线索人物刘姥姥也在小说拥有非常重要的地位，这一点在第一主成分也有所体现。

观察第一主成分单个方向上各个样本（即事件）的得分，选取得分最高的10个事件和得分最低的10个事件如表3所示。可以看出第一主成分方向上得分最低的事件大多是荣国府的群戏场景，例如聚餐、联诗、看戏等，而得分较高的事件多为宁国府及刘姥姥相关。刘姥姥本身就是一位非常重要的线索人物，与她相关的事件大多是贯穿两府人物的重要事件，宁国府众人如贾蓉、秦氏得以正面登场，因此第一主成分能够成功区分出荣国府和宁国府两条情节线索。

4.3 第二主成分与其方向上的情节线

与上一节类似，考察上一节中得到的第一主成分第二主成分上各个变量的系数，其系数最大的10个变量和系数最小的10个变量如表4所示。根据表4可以看出，第二主成分基本反映了《红楼梦》主要人物的年龄分布，

人物	系数	人物	系数
林黛玉	-0.5058839	贾蓉	0.11693321
贾宝玉	-0.4505834	贾琏	0.08943793
薛宝钗	-0.4407956	贾珍	0.08524105
史湘云	-0.2471643	尤氏	0.05100580
贾探春	-0.2354265	平儿	0.04618061
袭人	-0.1965056	尤二姐	0.04521055
李纨	-0.1753417	刘姥姥	0.03583708
史太君	-0.1625484	秦氏	0.03256178
贾迎春	-0.1336803	邢夫人	0.03105103
贾惜春	-0.1265797	贾蔷	0.02871870

表 2: 第一主成分变量系数表

表格左侧为系数最小的10个变量及其对应系数，右侧为系数最大的10个变量及其对应系数。

以宝、黛、钗为代表的年轻贵族与家人形成一个团体，另一方面以王熙凤、贾母等人为代表的年龄较长的贵族与家人则形成另一团体。需要说明的是此处的“年轻”与“年长”是相对的，而王熙凤、秦可卿的实际年纪并不大，此处的“家人”指的是贾府地位较高的仆人。

同样，观察第二主成分单个方向上各个样本的得分，选取得分最高的10个事件和得分最低的10个事件如表5所示。可以看出第二主成分方向上得分最低的事件大多是荣府青年男女间开展的天真烂漫的活动，例如作诗结社、情感互动等；而得分较高的事件多为较为年长的贵族之间开展的较为世故繁琐的家族事务性活动。并且这两类活动带有明确的情感色彩。因此第二主成分可以清晰的区分出由宝、黛、钗为代表的未婚青年男女间开展的积极、明快的情节线，以及由王熙凤和贾母为代表的已婚年长贵族发展出的家族利益相关情节线。

事件	得分	事件	得分
藕香榭聚餐	-2.095548	贾母瞧病	1.0088825
搬进大观园	-2.017665	尤家姐妹	1.0051883
宝玉用功练字	-1.997529	讨银	0.9925885
诗社作诗	-1.958459	宁府	0.9291682
林黛玉追打史湘云	-1.947453	开丧破孝	0.9291682
庆生辰雅座行令	-1.870498	给贾母准备礼物	0.9085889
芦雪庵联诗	-1.761347	贾蔷接事回贾琏	0.9020898
贾政辨凶兆	-1.722533	刘姥姥一进荣国府	0.8949761
众亲戚相认入住大观园	-1.627265	下恩旨	0.8781624
放风筝放晦气	-1.567481	打官司	0.8379200

表 3: 第一主成分样本得分表

表格左侧为得分最大的10个变量及其对应系数，右侧为得分最小的10个变量及其对应系数。

4.4 其他主成分与其方向上的情节线

尽管其他主成分对于方差的贡献远小于第一主成分和第二主成分，但是通过观察第三主成分和第四主成分，仍然能够得到一些非常有趣的发现。做出第三主成分和第四主成分各变量对应系数图如图3所示。

根据第三主成分可以区分出由以贾宝玉(-0.5841)和袭人(-0.4004)为中心开展的请安贾赦(-1.4406)、贾母携贾氏子侄赏月(-1.2361)等事件形成情节线，和以林黛玉(0.1669)、薛宝钗(0.3203)、史湘云(0.1127)为代表的贾府年轻女眷开展的祭饯花神(1.4567)、怡红院内金钗偶遇(1.4143)等年轻女性活动情节线。

而第四主成分则可以提取出以王熙凤(0.3620)、平儿(0.4173)、袭人(0.4060)为中心开展的惑奸谗抄检大观园(1.6608)、凤姐张罗袭人回家(1.2684)、袭人看望凤姐(1.0557)等事件形成的情节线。其中平儿作为王熙凤的贴身丫头与王熙凤在相同的方向上占据相似的比重是合理的，但是袭人作为贾宝玉的贴身丫头，与王熙凤和平儿在相同的方向上出现则显得

人物	系数	人物	系数
王熙凤	-0.4692509	贾宝玉	0.15963493
史太君	-0.4456338	晴雯	0.06729805
王夫人	-0.4130041	林黛玉	0.06171979
邢夫人	-0.2713506	袭人	0.05978446
贾珍	-0.2127553	麝月	0.03474533
尤氏	-0.1986364	香菱	0.03255482
贾琏	-0.1968891	茗烟	0.02359801
贾蓉	-0.1947474	莺儿	0.02221933
薛姨妈	-0.1465684	芳官	0.01922470
鸳鸯	-0.1411202	甄士隐	0.01346875

表 4: 第二主成分变量系数表

表格左侧为系数最小的10个变量及其对应系数，右侧为系数最大的10个变量及其对应系数。

比较异常。一种可能解释是，袭人作为王夫人内定的儿媳，而王熙凤与王夫人属于同一利益集团，因此王夫人与袭人之间有着几次非常重要的互动，王熙凤张罗袭人回家便是其中之一。²

结合上一节和本节的分析结果，可以推测《红楼梦》一书大体是由贾宝玉和王熙凤两位中心人物展开，由贾宝玉为中心展开以年轻贵族与家人活动为主的情感性情节线，同时由王熙凤为中心展开以年纪较长的贵族与家人活动为主的事务性情节线。

4.5 情节线总结

通过主成分分析，总结上述结果，《红楼梦》的整体人物情节结构可以归纳为图4所示。小说的人物情节可以划分为荣国府和宁国府两部分，其重心落在了荣国府这一部分之上。而对于荣国府部分的展开又可以分为

²本节人物名称后括号中的内容为变量在主成分中对应的系数，事件名称后括号中的内容为样本在该主成分方向上的得分。

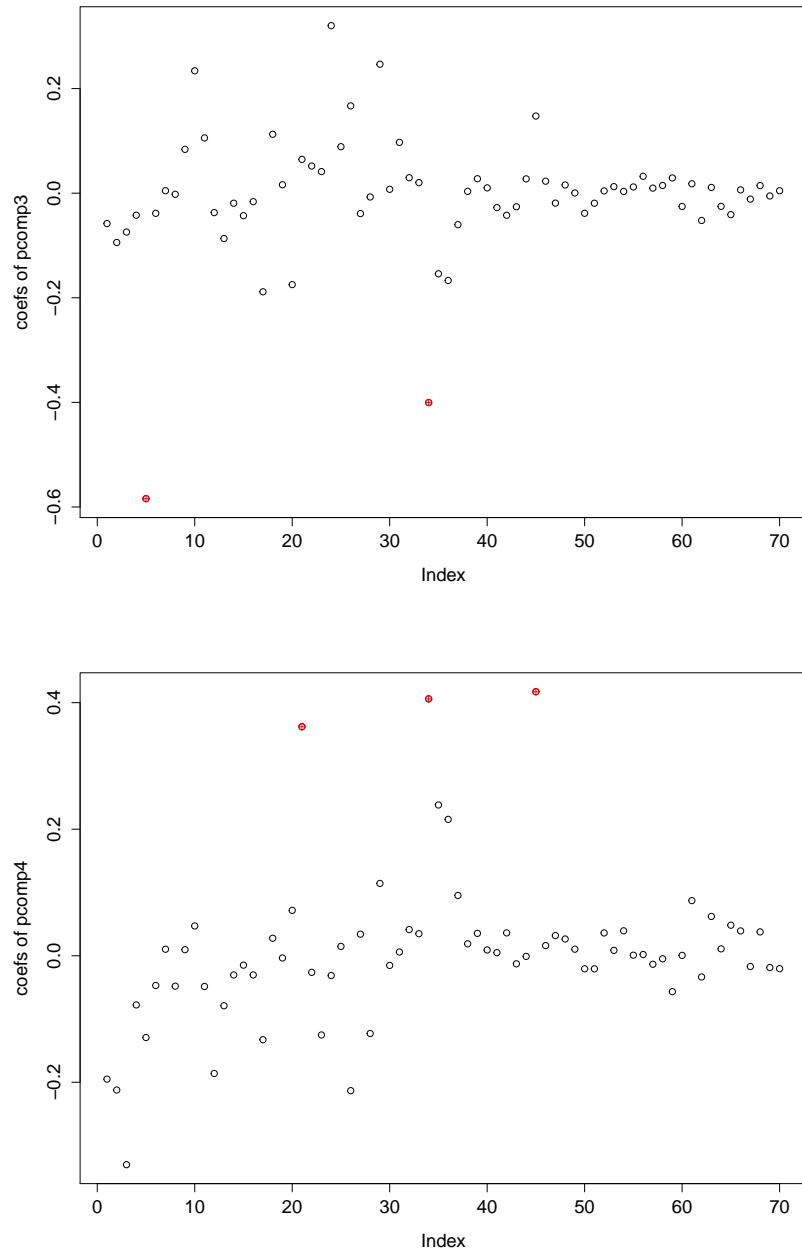


图 3: $CharEve_{main}$ 的第三主成分和第四主成分系数图

上图为第三主成分系数，被红色标出的点从左向右对应人物为贾宝玉和袭人；下图为第四主成分系数，被红色标出的点从左向右对应为王熙凤、袭人和平儿。

事件	得分	事件	得分
贾琏纳秋桐为妾	-2.352788	晴雯与贾宝玉吵架	0.7245726
春祭恩赏	-2.259318	葬花读会真记听牡丹亭	0.7134274
挑唆张华	-2.115266	袭人房中	0.7095033
秦氏后事	-2.086782	宝玉转话晴雯怒	0.6975981
凤姐生日凑钱	-1.917496	宝玉打闹	0.6975981
贾母携贾氏子侄赏月	-1.882197	梦游太虚幻境	0.6962702
众人斗牌哄贾母高兴	-1.844913	芳官独食	0.6938610
黛玉入府	-1.665450	看字	0.6647881
尤二姐吞金	-1.656763	宝玉命晴雯送手帕给林黛玉	0.6647881
多事之秋众女怨	-1.644008	查夜	0.6628528

表 5: 第二主成分样本得分表

表格左侧为得分最大的10个变量及其对应系数，右侧为得分最小的10个变量及其对应系数。

两条线索，一条是以贾宝玉为中心的年轻贵族与家人展开的情感性线索，最为突出的是宝、黛、钗的爱情线，另一条则是以王熙凤为中心的年长贵族与家人展开的事务性线索，其中的抄检大观园等情节也让人印象深刻。

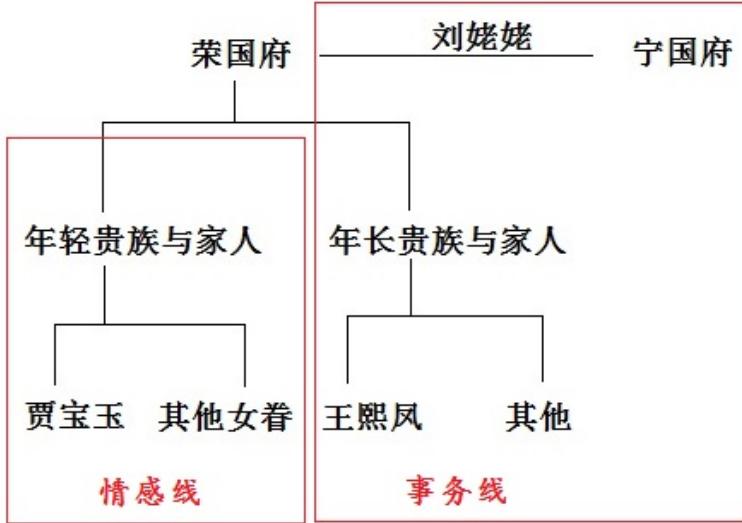


图 4: 《红楼梦》人物情节结构

5 《红楼梦》中的人物关系网络

5.1 基本方法

上一节是从人物事件矩阵 $CharEve_{main,org}$ 出发, 通过主成分分析得出了《红楼梦》小说的人物情节线索。而本节则以人物关系为基本的出发点, 通过对人物关系网络的分类来挖掘《红楼梦》中的人物团体。

构造无向图 G , 其顶点为人物集合, 对于两个顶点 i, j , 若 i, j 至少一次参与同一事件, 则 i, j 之间有一条边。定义连接矩阵 $Adj = \{A_{ij}\}$ 。当 i, j 至少一次参与同一事件时 $A_{ij} = 1$; 否则 $A_{ij} = 0$ 。

类似的, 可以构造有权重的无向图 G^* , 顶点仍然为人物集合, i, j 之间边的权重为 i, j 参与同一事件的个数。则 G^* 的连接矩阵为 $Adj^* = \{A_{ij}^*\}$ 。 A_{ij}^* 为 i, j 参与同一事件的个数。为简化叙述, 以下方法只讨论无权重的社交网络, 对有权重的无向图的处理方式与其非常类似。

问题转化为对图 G 顶点的分类问题, 本文采用基于模块化(modularity)的

分类方法 [4]。模块化是基于节点社区给定网络的一种分割后，该网络具有一种性质。当网络得到了一个好的分割，那么它大多数的边应当落在社区内部，而只有一小部分在社区之间。因此定义模块化指标如下：

$$Q = \frac{1}{2m} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j).$$

其中 $i, j \in Ch_{main}$, A_{ij} 是连接矩阵 Adj 中的元素； k_i 是顶点 i 的度数，即 $k_i = \sum_j A_{ij}$; c_i 是顶点 i 属于的社区类别，若图 G 可以被分为 k 个社区，则 $c_i \in \{1, 2, \dots, k\}$; 当 $c_i = c_j$ 时 $\delta(c_i, c_j) = 1$, 否则 $\delta(c_i, c_j) = 0$ 。

如果落在社区内的边数与一个随机网络的期望没有差别，那么 Q 应当为 0; 否则，这个网络与随机网络有区别。经验表明，当 $Q > 0.3$ 时，该网络便拥有非常显著的社区结构 [4]。

回到本文的语境之下，本节的任务是构建前 80 回的人物关系网络，并且找出一个好的分割，使得上述定义 Q 得到最大化。

优化 Q 的算法有很多，但是寻找一个分割使得 Q 达到最优解几乎是一个 NP-hard 问题，因此可以寻求次优解取而代之，故而本节采用 Walktrap 算法 [7] 寻求次优解。在本文的研究过程中，涉及求解社区发现问题以及社交网络可视化的部分均使用了 R 语言中的 igraph 包 [8]。

5.2 无权重网络上的社区挖掘

对《红楼梦》前 80 回文本的主要人物构建无权重人物关系网络 G_{main} ，这一网络只反映主要人物之间的关系，而不反映人物之间的亲密程度。因此这一网络可以较多地体现出人物之间的亲缘关系（亲戚）和从属关系（主仆）本身。通过上述模型求解，得到人物社区如图 5 所示，计算所得 $Q = 0.05791062$ 。同社区的规模及代表人物如表 6 所示。模型解得的分割方式，可以近似地将主要人物分为三个社区，以贾蓉、秦氏为代表的宁国府众人，以贾宝玉、王熙凤、贾母为代表的荣国府核心人物，以林黛玉、薛宝钗、史湘云为代表的荣国府年轻女眷。从结果来看，这一模型能够一定程度上反映出人物的亲缘关系。

红楼梦主要人物关系网络图（无权重）

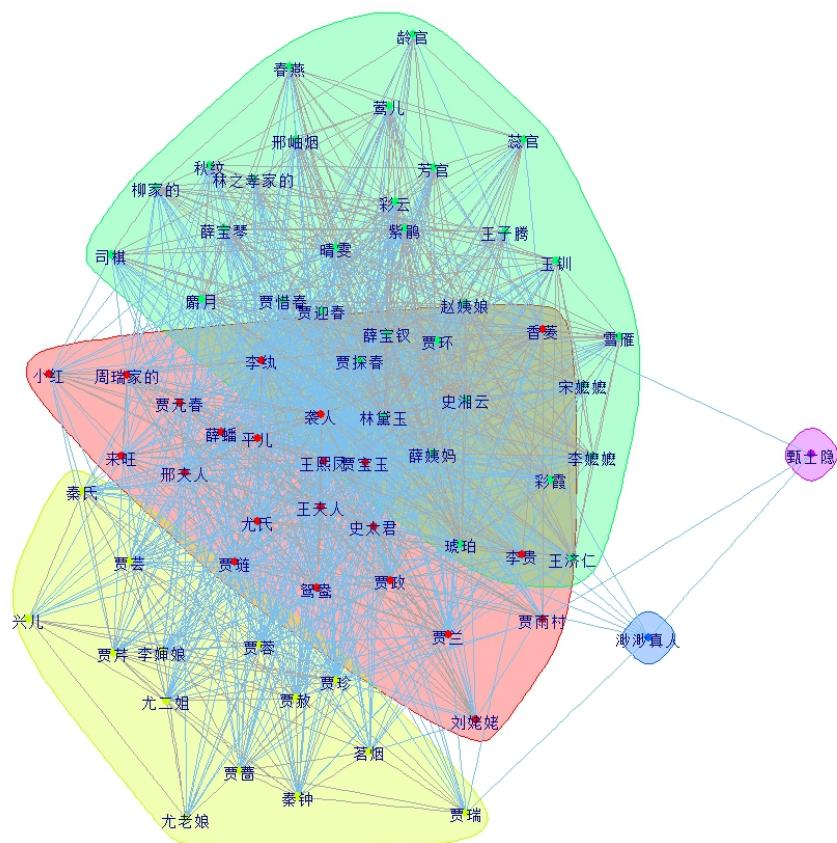


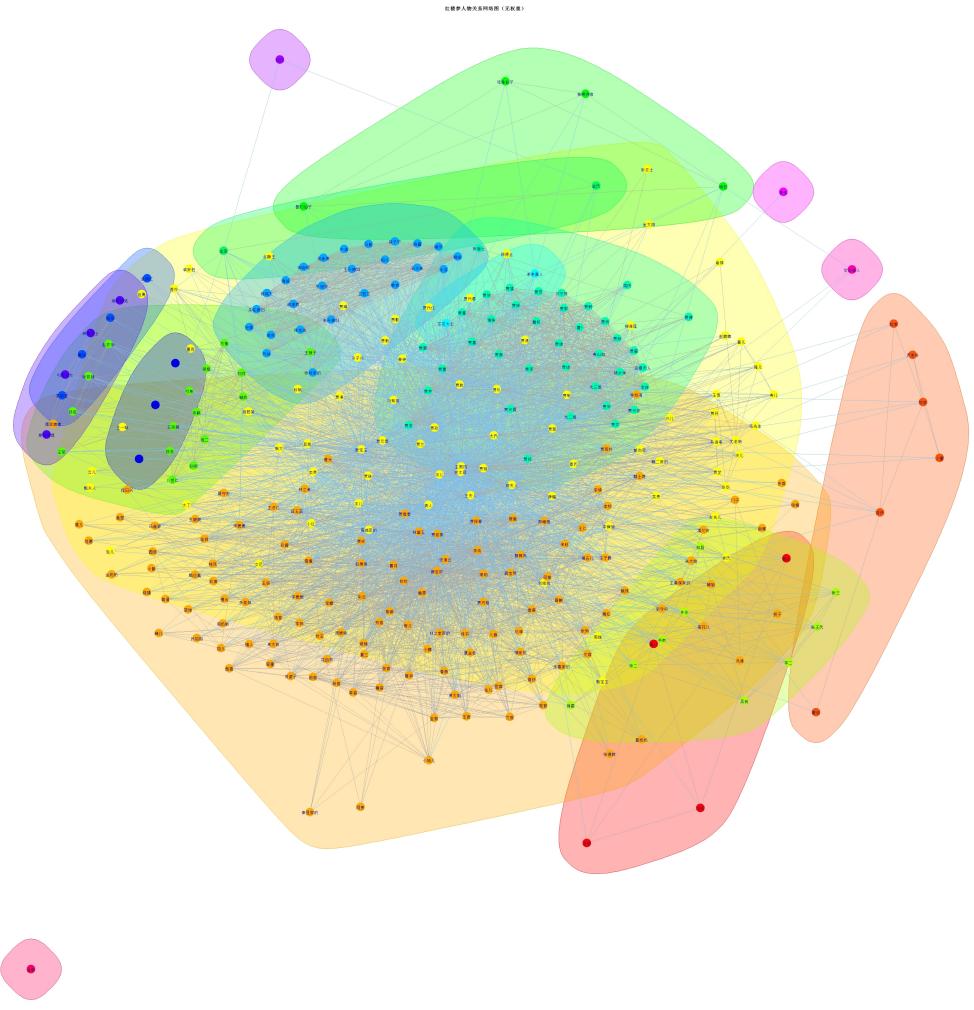
图 5: 《红楼梦》前80回主要人物关系无权重网络社区

图中节点的不同颜色代表所属社区，社区之间的边为浅蓝色，社区内的边为深灰色。

社区类别	规模	代表人物
1	22	贾宝玉、王熙凤、史太君、贾政、贾琏
2	13	贾蓉、贾蔷、秦氏、尤二姐
3	32	贾迎春、贾探春、史湘云、薛宝钗、林黛玉
4	1	渺渺真人
5	1	甄士隐

表 6: 《红楼梦》前80回主要人物无权重网络节点分类

如果对《红楼梦》前80回全部出场人物构建人物关系无权重网络 G_{all} , 同样对模型求解得到的人物社区如图6所示, 计算得 $Q = 0.3363465$. 对所有人物进行社区挖掘可以清晰的表现出一些边缘人物社区, 例如十二官 (芳官、藕官等)、神仙 (空空道人、神瑛侍者、绛珠仙子等)、香菱身世相关人物 (封肃、封氏、霍启等), 但是对于主要人物的社群关系有所掩盖。



图中节点的不同颜色代表所属社区，社区之间的边为浅蓝色，社区内的边为深灰色。

5.3 有权重网络上的社区挖掘

类似的，可以对《红楼梦》前80回文本的主要人物构建有权重人物关系网络 G^* 。这一网络不仅反映了人物之间的关系，还反映了人物之间的亲密程度，连接两个人物的边权重越大，两个人物之间的互动越多，两个人物之间的关系越亲密。使用类似的模型求解得到人物社区如图7所示，计算得 $Q = 0.1448009$ 。同社区的规模及代表人物如表7 所示。模型解得的分割方式可以将主要人物分为两个社区，一个是以王熙凤、史太君为代表的较为年长的贵族及家人社区，另一个是以贾宝玉、林黛玉为代表的较为年轻的贵族及家人社区。这两个社区的分割依赖于节点之间的亲密程度，即依赖于人物-事件之间的联系，而这一结果与上文主成分分析得到的结果相符合。

社区类别	规模	代表人物
1	34	王熙凤、史太君、贾政、贾琏、贾蓉、秦氏
2	34	贾宝玉、贾探春、史湘云、薛宝钗、林黛玉
3	1	渺渺真人
4	1	甄士隐

表 7: 《红楼梦》前80回主要人物有权重网络节点分类

如果对《红楼梦》前80回全部出场人物构建人物关系有权重网络 G_{all}^* ，同样对模型求解得到的人物社区如图8所示，计算得 $Q = 0.3055016$. 同无权重网络一样，对所有人物进行社区挖掘可以得到许多较为不活跃的社区，但是并没有掩盖住上文主要人物形成的两大社区，这为上文的结果提供了更高的可信度。

红楼梦主要人物关系网络图（有权重）

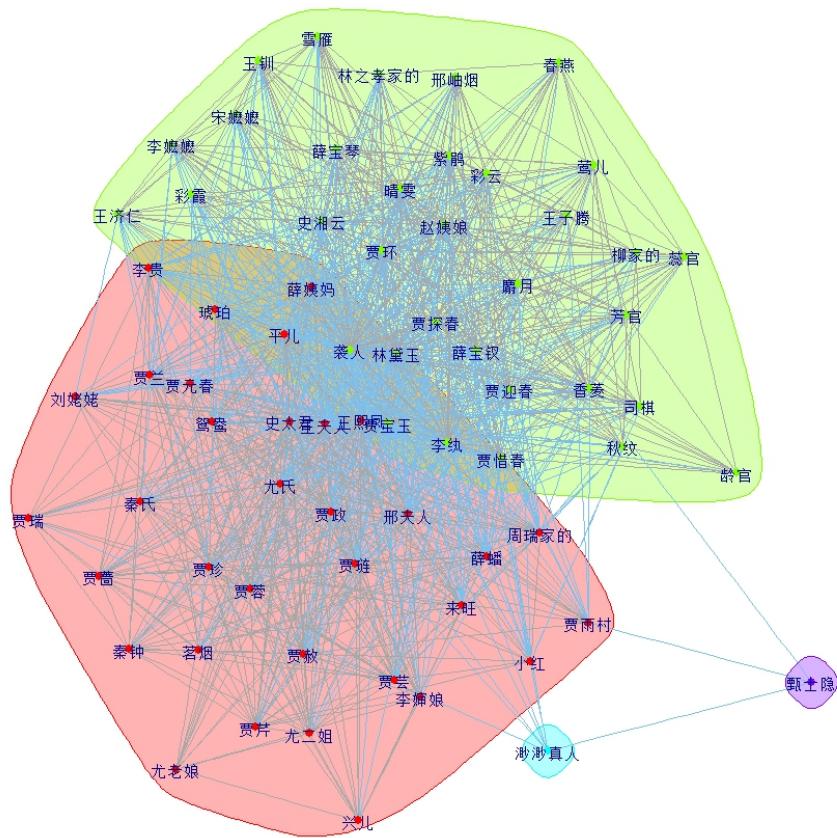


图 7: 《红楼梦》前80回主要人物关系有权重网络社区

图中节点的不同颜色代表所属社区，社区之间的边为浅蓝色，社区内的边为深灰色。

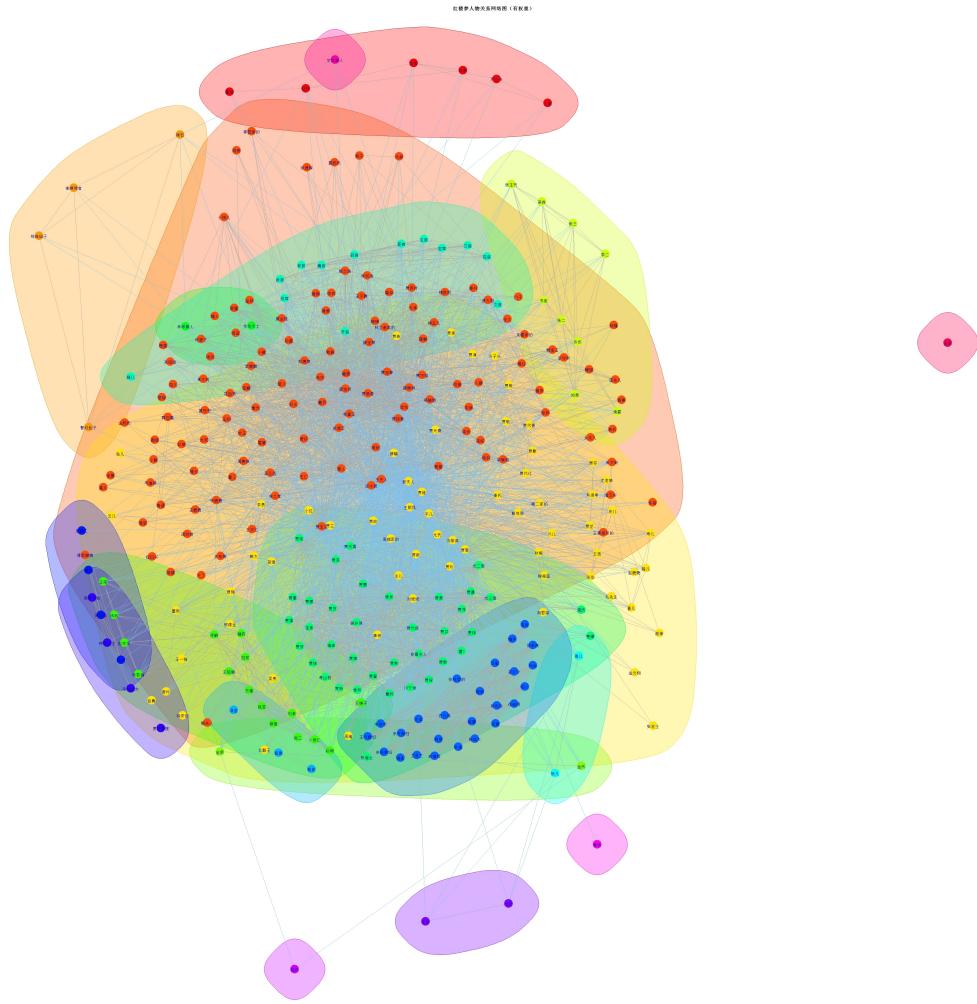


图 8: 《红楼梦》前80回全部人物关系有权重网络社区

图中节点的不同颜色代表所属社区，社区之间的边为浅蓝色，社区内的边为深灰色。

6 《红楼梦》中的人物关系发展

6.1 基本方法

在本节内容中，考虑将所有事件看做按照时间顺序顺次发生的序列，从而产生一个人物事件矩阵序列 $\{CharEve_i, i = 1, 2, \dots, n\}$ ，它包含了所有人物在第*i*+1个事件发生之前发生的所有事件的参与情况，即 $CharEve_i$ 是由 $CharEve$ 的前*i*列组成。由 $CharEve_i, i = 1, 2, \dots, n$ ，可以产生一系列的无权重人物关系网络 $G_i, i = 1, 2, \dots, n$ 。本节以此为依据，观察《红楼梦》的人物关系是如何伴随情节的发展而展开的。本节同时采用了同为四大名著之一的《西游记》的真实数据³，以及按照一定方式产生的随机数据作为对照组。

对于随机数据的产生遵循以下规则：

1. 估计出《红楼梦》参与每个情节的人数期望 $\lambda = \sum_{i \in Event} \frac{\sum_{j \in Ch} ce_{ij}}{|Event|}$ ，其中 $CharEve = \{ce_{ij}\}$ ；
2. 以 $Poisson(\lambda)$ 产生一个长为400的向量 $\vec{s} = \{s_1, s_2, \dots, s_{400}\}$ 作为每个情节的参与人数；
3. 从 $1, 2, \dots, 300$ 这一序列中抽取的 s_i 个数字，作为在第*i*个事件中出现的人物，并重复400 次，得到一个随机产生的 300×400 的人物事件矩阵；

6.2 人物数量伴随情节发展的变化

对于每一个人物事件矩阵 $CharEve_i, i = 1, 2, \dots, n$ ，可以计算其参与总人数，即非零横向量个数 s_i ，从而得到了一个伴随情节发展的参与人数序列 $\{s_i, i = 1, 2, \dots, n\}$ 。观察分别由《红楼梦》、《西游记》和随机数据产生的数组序列，其变化情况如图9所示。可以看出相比《西游记》，《红楼梦》与随机模拟数据的人数产生机制更为类似，其人数增长逐渐趋于平缓，说明在小说情节发展之初，人物分布还较为零散，而伴随着情节的发展，小说人物逐渐归于完整，在此将其称为多线索交叉发生机制。而《西游记》

³<http://www.math.pku.edu.cn/teachers/yaoy/Spring2013/>

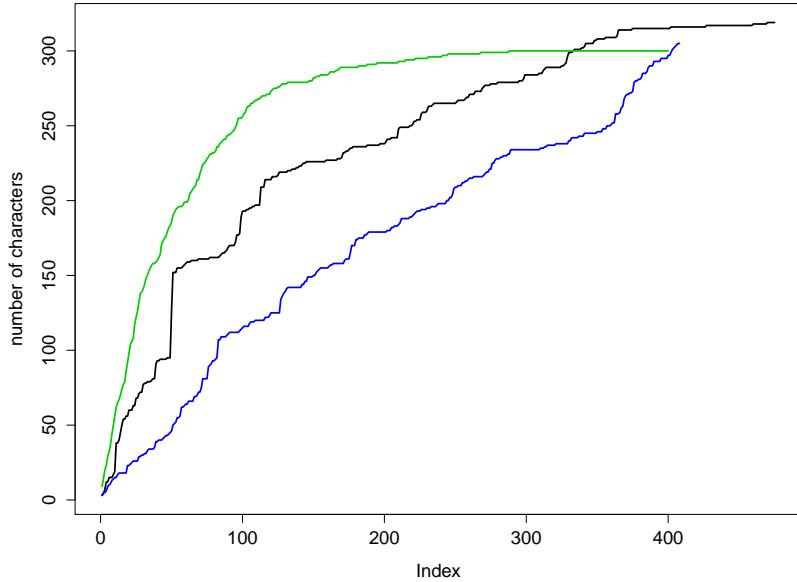


图 9: 《红楼梦》、《西游记》、随机数据随情节发展的参与事件人数趋势图

图中横轴表示事件编号，纵轴表示此前参与事件的总人数；图中黑色、蓝色和绿色线条表示《红楼梦》、《西游记》和随机数据的人数变化趋势。

在情节发展的过程中，极少数主要人物保持不变，同时不断地有新的人物产生，从而人数伴随情节发展逐渐增加，在此将其称为单线索递增发生机制。

6.3 人物关系网络伴随情节发展的变化

对于每个人物事件矩阵 $CharEve_i, i = 1, 2, \dots, n$ ，可以构造一个人物关系网络图 G_i ，并对其进行人物节点社区挖掘，得到相应的模块化指标，形成序列 $\{Q_i\}$ 。根据常识，在小说撰写过程中，一个基本的假设是：当情节发展到一定程度后，人物之间的关系趋于稳定，即在新事件中的人物互动对于关系网络的既有结构破坏不大。基于这样的假设，可以对《红楼梦》、《西游记》和随机模拟数据计算相应的序列 $\{Q_i\}$ 进行观察。得到结果如图10中的上方图片所示。由于在情节发展之初， Q_i 的波动较大，为了方

便观察后续情节 Q 值的趋势，可以选取 $i > 100$ 的部分如图10中的下方图片所示。

注意到随机产生的数据在 i 大于某个值后开始下降，其下降速度由快变慢，减慢的原因是此时的人物关系网络逐渐混合在一起，直到不可再分时， Q 值降为0。而对于真实的小说《红楼梦》、《西游记》的数据，并没有发现类似的情况，其 Q 值变化比较平缓，持续在一个大于0的水平，《西游记》的单线索递增结构甚至可以造成一段时间内 Q 值的增大。另一方面，对于图10中被描红的《红楼梦》后40回数据，发现了很小的阶梯型下降。为了进一步观察 Q 值的变化情况，我们计算了 $i > 100$ 时 Q 值的增量 $\Delta Q_i = Q_i - Q_{i-1}$ ，其结果如图11所示。考察 $\{i \in 100, 101, \dots, 475 \mid |\Delta Q_i| > 0.01\}$ ，找出编号为100的事件之后 Q 值变化比较大的那些事件为：祭饯花神（第二十七回， $\Delta Q = 0.01650038$ ）、藕香榭聚餐（第三十八回， $\Delta Q = -0.01075032$ ）、众伙计意欲出发（第五十二回， $\Delta Q = 0.01653596$ ）、矢孤介杜绝宁国府（第七十四回， $\Delta Q = -0.02017646$ ）、贾政升官（第八十五回， $\Delta Q = -0.01751388$ ）、贾家庆贺（第八十五回， $\Delta Q = 0.02193870$ ）、薛姨妈营救薛蟠（第八十六回， $\Delta Q = -0.01898191$ ）、宝玉静室诓功名-敛神定息止尘心（第一百一十八回， $\Delta Q = -0.01066699$ ）。可以发现， $|\Delta Q_i|$ 较大的那些事件大多发生在后四十回，尤其是前八十回与后四十回接壤的若干章回中，从而造成了 Q 值在末端出现了小幅度的阶梯型下降。一种可能的解释是后40回稿件是由另一位作者续写，对小说原有人物结构造成了破坏，从而出现了 Q 值轻微的断层。

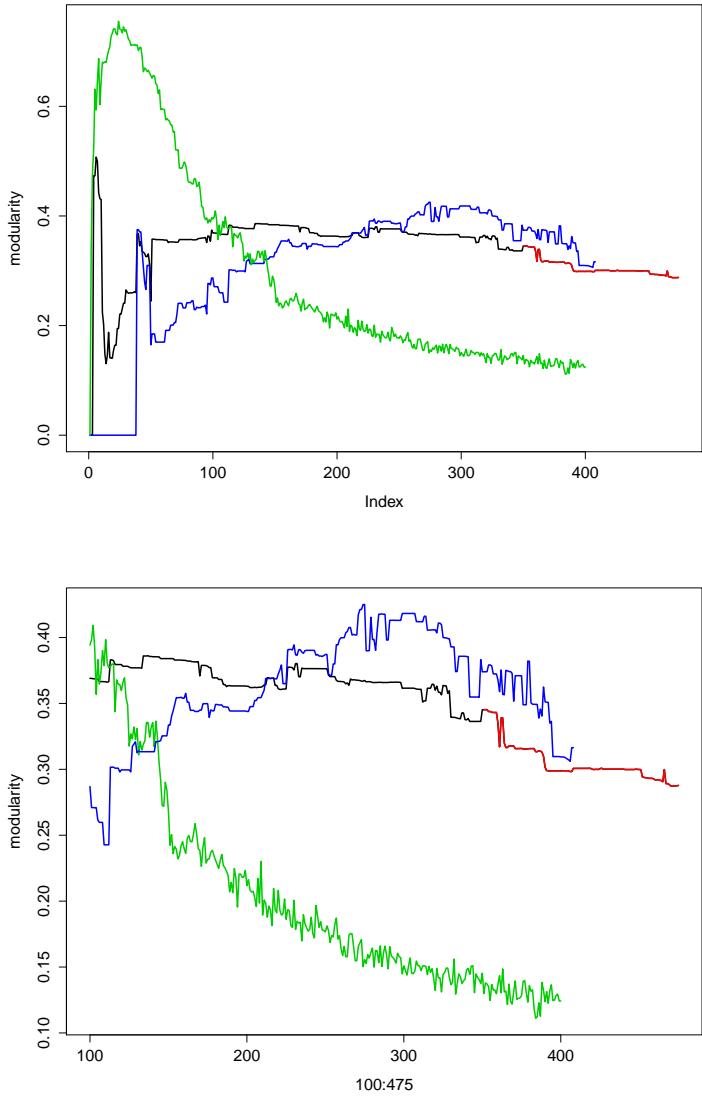


图 10: 《红楼梦》、《西游记》、随机数据随情节发展的 Q 值趋势图

图中横轴表示事件编号，纵轴表示相应的 Q_i ；上图为事件编号从1开始的 Q 值变化趋势，下图为事件编号从100开始的 Q 值变化趋势；图中黑色、蓝色和绿色线条表示《红楼梦》、《西游记》和随机数据的 Q_i 变化趋势；黑色被描红的部分为《红楼梦》后40回的 Q_i 。

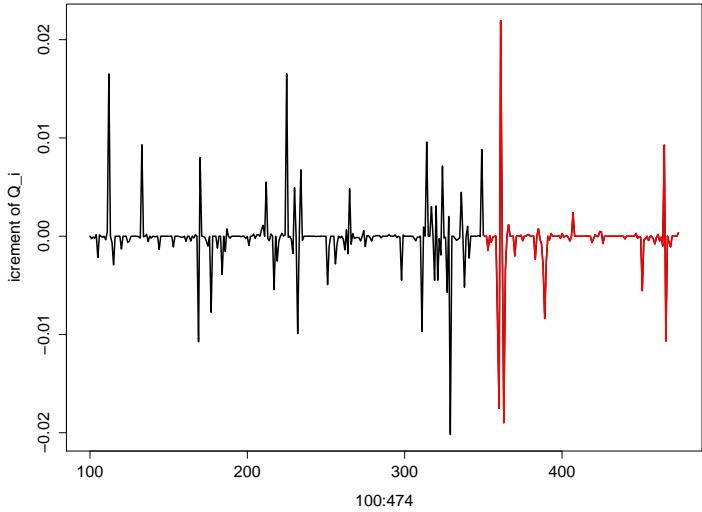


图 11: 《红楼梦》随情节发展的 Q 值增量趋势图

图中横轴表示事件编号，纵轴表示编号从100开始的事件相对应的 ΔQ_i ，被描红的部分为后40回事件对应的 ΔQ_i .

7 结论

本文运用主成分分析、模块化分类等方法讨论了《红楼梦》的人物情景关系及人物关系网络。通过主成分分析人物情景矩阵，本文构建起了《红楼梦》前八十回的基本人物情节结构，提出了以贾宝玉为中心的情感线索和以王熙凤为中心的事务线索并行发展的观点。而通过对于前八十回人物关系网络的模块化分类，也印证了上述观点。本文通过对比《红楼梦》和《西游记》的人物关系网络伴随情节发展的变化，指出《红楼梦》是多线索交叉结构而《西游记》则是单线索递增结构。这一结论与人们的阅读感受非常相符。另一方面，其 Q 值伴随情节发展在后40回发生阶梯型减少，也在一定程度上与后40回作者并非曹雪芹的研究结果相符。

参考文献

- [1] 李贤平. 《红楼梦》成书新说. 复旦学报(社会科学版), (05):3–16, 1987. 31-1142/C.
- [2] Ying Liu Hui Li. Language models and classification analysis for dream of the red chamber. In Proceedings of 2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems, page 6, epartment of Chinese Language and Culture,Tsinghua University, 2012.
- [3] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. Statistical Analysis and Data Mining, 4(5):512–546, 2011.
- [4] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. Physical review E, 70(6):066111, 2004.
- [5] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. Physical review E, 69(2):026113, 2004.
- [6] Donald Ervin Knuth, Donald Ervin Knuth, and Donald Ervin Knuth. The Stanford GraphBase: a platform for combinatorial computing. AcM Press New York, 1993.
- [7] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In Computer and Information Sciences-ISCIS 2005, pages 284–293. Springer, 2005.
- [8] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal, Complex Systems:1695, 2006.

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名： 日期： 年 月 日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校一年/两年/三年以后，在校园网上全文发布。

(保密论文在解密后遵守此规定)。

论文作者签名： 导师签名：

日期： 年 月 日