社会网络分析期末报告

《红楼梦》社会网络分析

学　　院：信息与通信工程学院

班　　级：2018211120

姓　　名：吴限

学　　号：2018210120

一、开篇引入

《红楼梦》，《百年孤独》，《指环王》等等中外经典作品为读者构建了一个个人物关系复杂，故事情节跌宕起伏的家族背景。因为其人物与故事众多，读者热衷于深挖这些故事背后的人物关系。在信息化时代，我们如何利用计算机对这些故事进行另一个角度的分析？本文中，笔者将利用社会网络分析相关知识进行分析，以期得到红楼梦中最重要的人物，权势最大的人物，与红楼梦中的各个小集团。

二、相关工作

参考数据森麟《“水泊梁山”互联网有限公司一百单八将内部社交网络》[1]，《用 python 分析<三国演义中的社交网络>》[2]，《红楼梦主要人物词云》[3]等文章，确定了实验目标：

（1）人物社交网络

（2）人物地位、权势情况排名

（3）人物社区发现

三、方法

1. 问题定义

（1）人物社交网络

人物为 node 节点，人物之间关系使用边进行连接。构建全书人物之间的社交网络。

（2）人物地位、权势情况排名

根据全书人物的剧情，使用 pagerank 算法对人物地位、权势情况进行排名。

（3）人物社区发现

人以类聚，物以群分。红楼梦中贾、薛、王、史四大家族互为姻亲，彼此往来，形成了一个大“社区”。但是其中又存在不少小集团，有着故事分支。本项目中使用社区发现算法，进行社区发现。

2. 数据来源

曹雪芹《红楼梦》文本已超过版权保护年限，电子版通过在线阅读网站[4]爬虫下载。

第一回　甄士隐梦幻识通灵　贾雨村风尘怀闺秀

此开卷第一回也．作者自云：因曾历过一番梦幻之后，故将真事隐去，而借"通灵"之说，撰此《石头记》一书也．故曰"甄士隐"云云．但书中所记何事何人？自又云："今风尘碌碌，一事无成，忽念及当日所有之女子，一一细考较去，觉其行止见识

此回中凡用"梦"用"幻"等字，是提醒阅者眼目，亦是此书立意本旨．

列位看官：你道此书从何而来？说起根由虽近荒唐，细按则深有趣味．待在下将此来历注明，方使阅者了然不惑．

原来女娲氏炼石补天之时，于大荒山无稽崖练成高经十二丈，方经二十四丈顽石三万六千五百零一块．娲皇氏只用了三万六千五百块，只单单剩了一块未用，便弃在此山青埂峰下．谁知此石自经煅炼之后，灵性已通，因见众石俱得补天，独自己无材

一日，正当嗟悼之际，俄见一僧一道远远而来，生得骨骼不凡，丰神迥异，说说笑笑来至峰下，坐于石边高谈快论．先是说些云山雾海神仙玄幻之事，后便说到红尘中荣华富贵．此石听了，不觉打动凡心，也想要到人间去享一享这荣华富贵，但自恨

后来，又不知过了几世几劫，因有个空空道人访道求仙，忽从这大荒山无稽崖青埂峰下经过，忽见一大块石上字迹分明，编述历历．空空道人乃从头一看，原来就是无材补天，幻形入世，蒙茫茫大士，渺渺真人携入红尘，历尽离合悲欢炎凉世态的一

无材可去补苍天，枉入红尘若许年．

此系身前身后事，倩谁记去作奇传？诗后便是此石坠落之乡，投胎之处，亲自经历的一段陈迹故事．其中家庭闺阁琐事，以及闲情诗词倒还全备，或可适趣解闷，然朝代年纪，地舆邦国，却反失落无考．

空空道人遂向石头说道："石兄，你这一段故事，据你自己说有些趣味，故编写在此，意欲问世传奇．据我看来，第一件，无朝代年纪可考，第二件，并无大贤大忠理朝廷治风俗的善政，其中只不过几个异样女子，或情或痴，或小才微善，亦无班姑，

空空道人听如此说，思忖半晌，将《石头记》再检阅一遍，因见上面虽有些指奸责佞贬恶诛邪之语，亦非伤时骂世之旨，及至君仁臣良父慈子孝，凡伦常所关之处，皆是称功颂德，眷眷无穷，实非别书之可比．虽其中大旨谈情，亦不过实录其事，又

3. 算法

针对实验目标中确定的三个问题，笔者将进行分别分析。

（1）预处理

首先，针对文本内容存在言白混杂的中文文本情况，我们需要先进行分词。参考[5]利用 jieba 库进行中文分词，其中我下载了红楼梦的分词库--停用词.txt 导入 jieba 作为主词典。这里需要注意的是，半文半白的文章中，标点符号、虚词、数字等都将被认为是无用的词，与我们日常的停用词、文言文中的停用词一起进行移除。

下面对人物构建词典，参考[3]中提供的主要人物与其别称：

宁国府：*贾珍、尤氏（珍大奶奶）、尤二姐、尤三姐、贾惜春（四姑娘）、贾蓉、秦可卿（秦氏、蓉大奶奶）*

荣国府：*贾母（史太君、老祖宗、老太太）、鸳鸯、琥珀、傻大姐、贾赦、贾政、王夫人、金钏、贾宝玉（宝二爷）、袭人、晴雯、麝月、秋纹、碧痕、春燕*

贾琏、王熙凤（凤姐、琏二奶奶）、平儿、丰儿、林之孝、贾元春（娘娘）

贾迎春（二姑娘）、司棋、贾探春（三姑娘）、李纨（宫裁）、碧月、林黛玉（林姑娘、林妹妹）、雪雁

王家：*王子腾*

史家：*史湘云（史大姑娘）、翠缕*

薛家：*薛姨妈（姨太太）、薛宝钗（宝姐姐、宝丫头、宝姑娘）、薛蟠、夏金桂、宝蟾、香菱、薛蝌、邢岫烟、薛宝琴*

代码进行如下替换：

```
if w.word =='林黛玉' or w.word == '林妹妹' or w.word == '林姑娘':
    w.word = '黛玉'
if w.word =='贾宝玉' or w.word == '宝二爷':
    w.word = '宝玉'
if w.word =='薛宝钗' or w.word == '宝姐姐' or w.word == '宝丫头' or w.word == '宝姑娘':
    w.word = '宝钗'
if w.word =='王熙凤' or w.word =='凤辣子' or w.word == '琏二奶奶' or w.word == '凤姐':
    w.word = '凤姐'
if w.word =='珍大奶奶':
    w.word = '尤氏'
if w.word == '惜春' or w.word == '四姑娘':
    w.word = '贾惜春'
if w.word == '秦氏' or w.word == '蓉大奶奶':
    w.word = '秦可卿'
if w.word == '史太君' or w.word == '老祖宗' or w.word =='老太太':
    w.word = '贾母'
if w.word == '娘娘' or w.word == '贾妃':
    w.word = '贾云春'
if w.word == '二姑娘' or w.word == '迎春':
    w.word = '贾迎春'
if w.word == '三姑娘' or w.word == '探春':
    w.word = '贾探春'
if w.word == '宫裁' or w.word == '李宫裁':
    w.word = '李纨'
if w.word == '史大姑娘':
    w.word = '史湘云'
if w.word == '姨太太':
    w.word = '薛姨妈'
```

（2）人物社交网络

每个人物为一个 node 节点，为了分析人物之间的网络连接情况，需要对人物共同出现的段落篇章进行统计，每出现一次，两个节点间的关联权重将加一。利用 networkx 工具最终获得人物之间的社交网络图。

（3）人物地位、权势情况排名

利用排序算法 PageRank 进行统计最重要的人物。利用中心性来描述人物权势高下，通过中心度进行计算衡量中心性程度。

【1】PageRank 算法：

PageRank，简称 PR，是 Google 排名运算法则（排名公式）的一部分，是 Google 用于用来标识网页的等级/重要性的一种方法，是 Google 用来衡量一个网站的好坏的重要标准之一。PageRank 对每个链入(inbound)赋以不同的权值来计算页面的重要性。链接提供页面的越重要则此链接入越高。当前页的重要性，是由其它页面的重要性决定的。

PageRank 的原理

假设一个由 4 个页面组成的小团体：A，B，C 和 D。如果所有页面都链向 A，那么 A 的 PR（PageRank）值将是 B，C 及 D 的 Pagerank 总和。

$$PR(A) = PR(B) + PR(C) + PR(D)$$

继续假设 B 也有链接到 C，并且 D 也有链接到包括 A 的 3 个页面。一个页面不能投票 2 次。所以 B 给每个页面半票。以同样的逻辑，D 投出的票只有三分之一算到了 A 的 PageRank 上。

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$

换句话说，根据链出总数平分一个页面的 PR 值。

$$PR(A) = \frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)}$$

最后，所有这些被换算为一个百分比再乘上一个系数。由于“没有向外链接的页面”传递出去的 PageRank 会是 0，所以，Google 通过数学系统给了每个页面一个最小值：

$$PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \frac{PR(D)}{L(D)} + \cdots\cdots \right) d + \frac{1-d}{N}$$

最后通过对 PR 值的排序，我们就可以得到重要性排序。

【2】中心性(centrality)：

用于表达社交网络中一个点或者一个人在整个网络中所在中心的程度，这个程度用数字来表示就被称作为中心度，即通过知道一个节点的中心性来了解判断这个节点在这个网络中所占据的重要性的概念。

测定中心度方法的不同，可以分为度中心度（Degree centrality），接近中心度（或紧密中心度，Closeness centrality），中介中心度（或间距中心度，Betweenness centrality）等。

1. 度中心度，（也可以理解为"连接中心度"），顾名思义就是一个点与其他点直接连接的总和。

在一个社会网络中，谁遇其他人的连接最多，谁就具有最大的中心性。因此我们需要知道这个人和多少人有连接，度中心度对这一指标进行了度量。有时也用节点的大小（Size）来表达，一个节点的 size 越大，就说明其所占据的中心性越高。

在实际情况中，可能出现连接有方向的情况。这种情况下，就给连接加入了向量的概念，也就是说连接是有方向的。这种情况通过点入中心度（或入度，in-degree）和点出中心度（或出度，out-degree）进行衡量。

- ✓ 入度表现一个人的被关注程度。点入中心度高的人（B）是其他人都想与其形成关联的对象，也就是在这个网络中，B 被很多人认为很有必要与其取得关联，也就可以理解成 B 在这个网络中具有很高的声望（prestige），体现了一个人的吸引力。入度高的人有可能会引导这个网络圈交流的内容、视角、深度、广度等问题。
- ✓ 出度表现一个人关注他人的程度。点出中心度高的人（A）是在这个网络中，很努力并活跃地与他人取得关联的人，可以理解成 A 在这个网络中具有较强的交际性，体现了一个人的积极性。

2. 接近中心度，计算的是一个点到其他所有点的距离的总和，总和越小说明这个点到其他所有点的路径越短，也就说明这个点距离其他所有点越近。

接近中心度体现的是一个点（node）与其他点的近邻程度。Bavelas（1950）将接近中心性定义为距离的倒数：

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

一个具有高接近中心度的点，在空间上与中心位置上距离任何其他点都最近。

同样，在有方向的社交网络（directional social networking）中对接近中心度（Closeness centrality）的分析结果，会得出入接近中心度（In-closeness centrality）和出接近中心度

（Out-closeness centrality）

✓ 入接近中心度（In-closeness centrality）

入接近中心度是通过计算走向一个点的边来测量出其他点（nodes）到达这个点（node）的容易程度，一个点的入接近中心度越高，说明其他点到这个点越容易。

✓ 出接近中心度（Out-closeness centrality）

出接近中心度指的是一个点到达其他点的容易程度，通过一个点到其他点的最短距离的和的倒数，接近中心度越大，这个点到其他点越容易。

因此入接近中心度表达的是整合力（integration），出接近中心度表达的是辐射力（radiality）。

3. 中介中心度，计算经过一个点的最短路径的数量。经过一个点的最短路径的数量越多，就说明它的中介中心度越高，这个点处于其他点对相互之间的捷径上。

（4）人物社区发现

【1】社区

如果一张图是对一片区域的描述的话，将这张图划分为很多个子图。当子图之内满足关联性尽可能大，而子图之间关联性尽可能低时，这样的子图可以称之为一个社区。

【2】社区划分的标准

在社区发现算法中，几乎不可能先确定社区的数目，因此通过模块度（Modularity）用来衡量一个社区的划分是不是相对比较好的结果。一个相对好的结果在社区内部的节点相似度较高，而在社区外部节点的相似度较低。

模块度的物理含义是社区内节点的连边数与随机情况下的边数之差，取值范围是[-0.5,1)。可以简单理解为社区内部边的权重减掉所有与社区节点相连的边的权重和；对于无向图而言，即社区内部边的度数减去社区内节点的总度数。

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

$$\delta(u,v)=\begin{cases}1 & \text{when u == v}\\ 0 & \text{else}\end{cases}$$

其中，A_{ij} 为节点 i 和 j 之间边的权重，网络不是带权图时，所有边的权重可以看做是 1。$k_i=\sum_j A_{ij}$ 表示所有与节点 i 相连的边的权重之和（度数）；c_i 表示节点 i 所属的社区；

$m=\frac{1}{2}\sum_{ij}A_{ij}$ 表示所有边的权重之和（边的数目）。

模块度的公式可简化为：

$$Q=\sum_{c}[\frac{\sum in}{2m}-(\frac{\sum tot}{2m})^{2}]=\sum_{c}[e_{c}-a_{c}^{2}]$$

基于模块度的社区发现算法，都是以最大化模块度 Q 为目标。

【3】社区发现算法[6]

社区发现算法有很多，例如 LPA，HANP，SLPA 以及 Louvain，不同的算法划分社区的效果不尽相同。Louvain 算法是基于模块度的社区发现算法，该算法在效率和效果上都表现较好，并且能够发现层次性的社区结构，其优化目标是最大化整个社区网络的模块度，可以通过调用 community 包直接使用。

算法流程：

- 初始时将每个顶点当作一个社区，社区个数与顶点个数相同。
- 依次将每个顶点与之相邻顶点合并在一起，计算它们的模块度增益是否大于 0，如果大于 0，就将该结点放入该相邻结点所在社区。
- 迭代第二步，直至算法稳定，即所有顶点所属社区不再变化。
- 将各个社区所有节点压缩成为一个结点，社区内点的权重转化为新结点环的权重，社区间权重转化为新结点边的权重。
- 重复步骤 1-3，直至算法稳定。

四、实验与结果

- 预处理

预处理部分对人物出场次数也进行了统计。部分如图所示。

甄士隐 11
贾雨村 21
封氏 2
神瑛侍者 7
严老爷 1
霍启 2
冷子兴 9
林如海 11
黛玉 1376
贾代化 2
贾敷 1
贾敬 10
贾珍 284
贾蓉 176
贾政 433
贾珠 6
宝玉 3862
贾迎春 143
贾探春 432
贾惜春 151
贾敏 1
贾琏 679
贾母 2301
凤姐 1131
王夫人 1011
邢夫人 284
贾源 2
王嬷嬷 3
雪雁 105
尤氏 267
赖二 1
詹光 9
单聘仁 3
吴新登 13
戴良 2
钱华 1
紫鹃 406
贾代儒 4
李贵 29
贾瑞 48
贾蔷 17
贾菌 11
墨雨 7
胡氏 1
贾璜 3
金氏 5
冯紫英 58
贾敕 1
贾效 1
贾敦 1
贾琮 6
贾珩 2
贾琛 2
贾琼 4
贾菱 4
贾芸 140
贾芹 33
贾蓁 1
智善 2
胡老爷 2
金哥 2
李公子 1
云光 1
贾元春 3
夏守忠 4
赵嬷嬷 13
卜固修 1
山子野 2
赖大 30
林之孝 143
程日兴 9
妙玉 123
贾环 116
秋纹 57
史湘云 82
赵姨娘 132
周氏 1
卜世仁 8
倪二 30
绮霰 1
方椿 1
檀云 2
小红 18
马道婆 25
周姨娘 11
冯唐 1
蒋玉菡 29

图 0 预处理

- 人物社交网络图

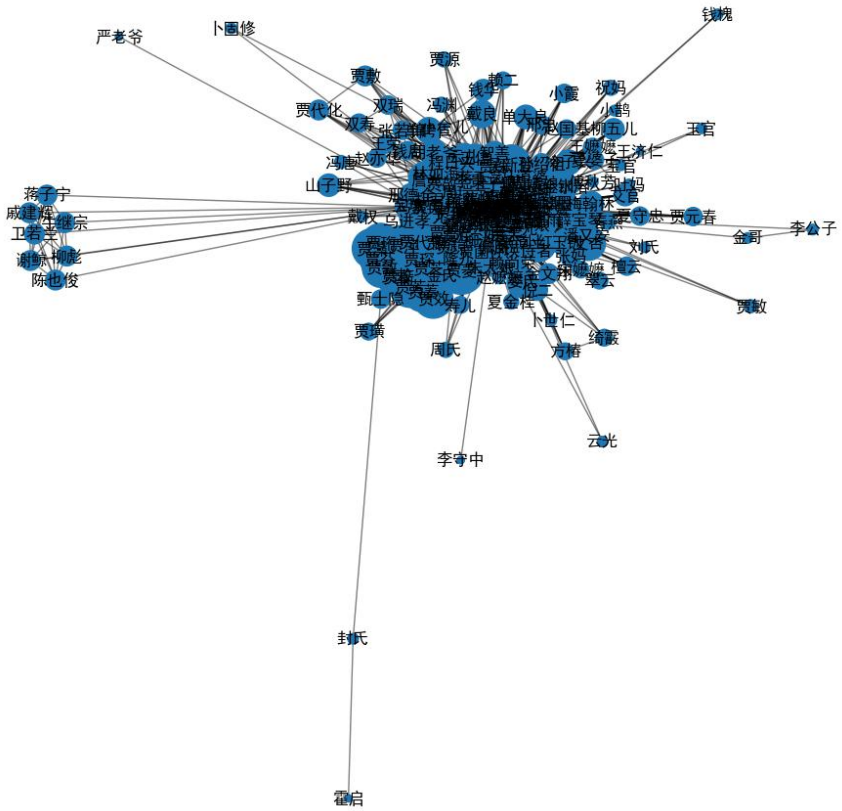


图 1 全书人物社交网络图

根据人物社会网络全图我们可以发现，核心人物之间关系紧密相连，而边缘人物几乎没有什么边相连，不具有很强的分析价值。

因此，我们重新选择连接数大于 30 的人物节点生成网络，进行研究。

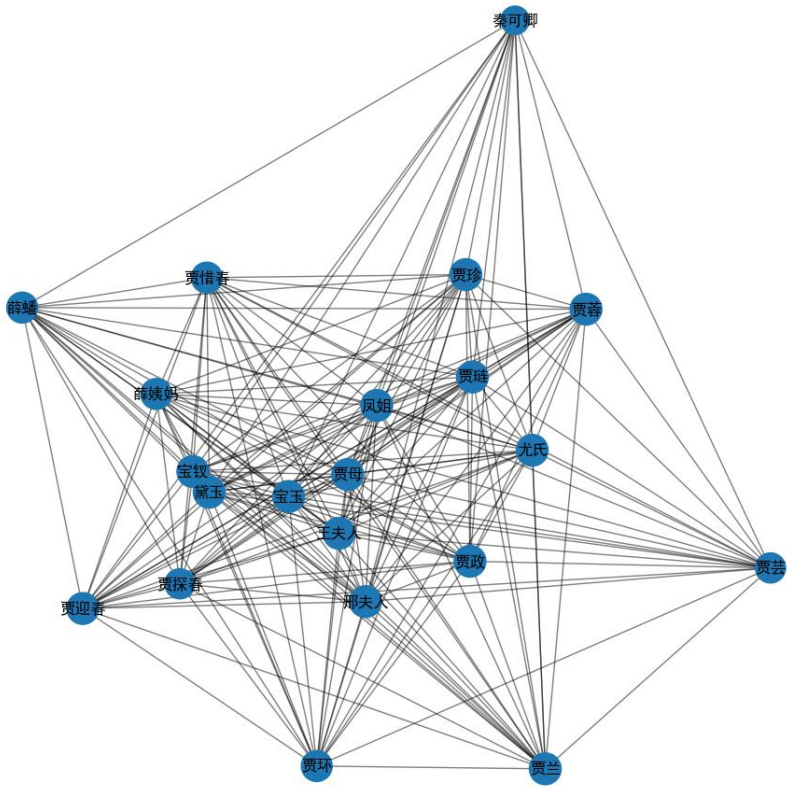


图 2 全书部分核心人物社交网络图

如图，我们可以发现故事的核心人物正是我们所耳熟能详的宝黛二人，宝钗，贾母，凤姐，三春等等。其余也都是四大家族的核心人物。

- 最重要的人物与最有权的人物

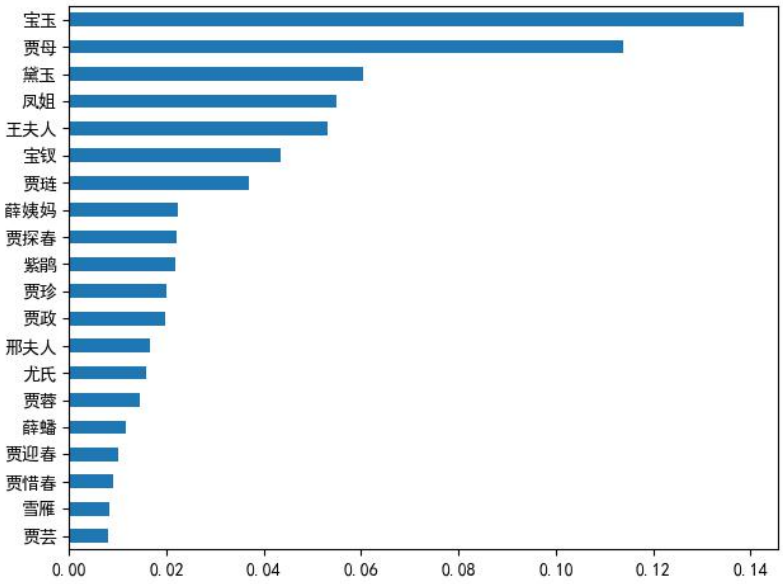


图 3 最重要的人物排名

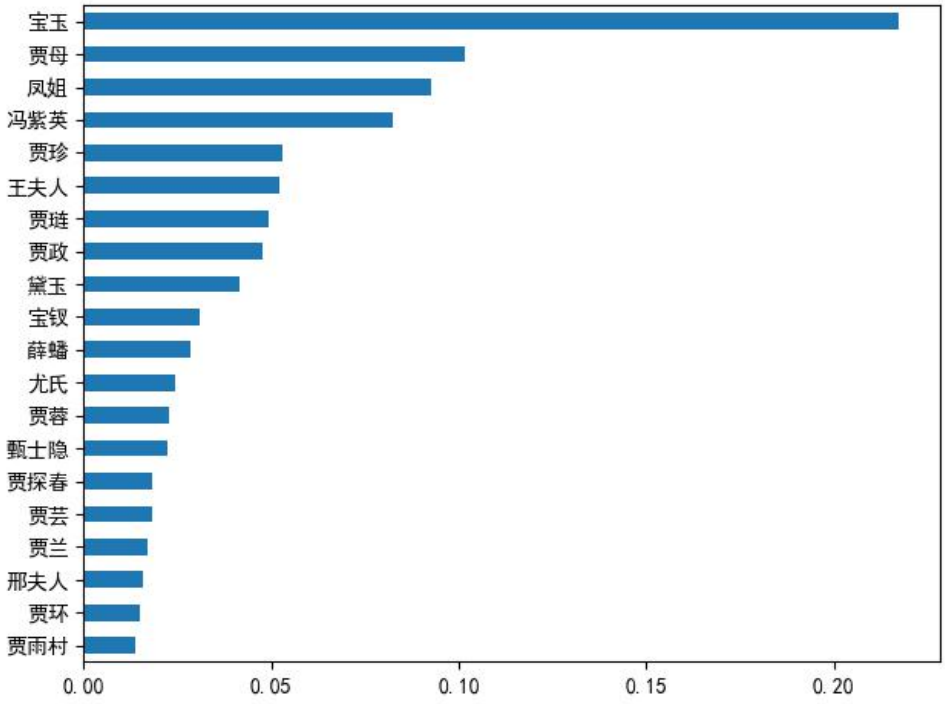


图 4 最有权势的人物排名

意外地发现，冯紫英、甄士隐、贾雨村等人并非核心人物，故事线也并不丰富，全文提到次数仅 10-30 次，与凤姐、黛玉、宝钗等人完全不在一个比较范围内，但是他们却在全书中更接近中心度。这暗示着他们对故事的发展有着极为重要的推动作用。

在红学领域，红学家们也都认为冯紫英为红楼梦中的一大神秘人物，提出了他实际上为年羹尧，隆科多等等猜测。贾雨村与甄士隐在全文头尾处多次出现，且他们每次的出现都标志着红楼故事四大家族的命运转折，在某种程度上确实体现了他们是全文核心人物。这些点也从侧面印证了我们这个项目中 pagerank 排序的有效性。

- 人物社区发现

社区发现算法每一次的结果都会有所出入。因此，我任意选取了其中两次的社区发现结果进行展示。可以发现，尽管有所出入，但是从文本角度而言，小集团的可解释性都很强。

如图 5 所示，community 0 所聚类的社区为各个家族核心掌权者与心腹；community 1 为玩乐多一些的纨绔子弟与艺人；community 2 为贾府晚辈子弟；community 3 为年轻一代，故事的集中发生人群，十二钗、宝玉、与他们交好的友人等等。

community 0: 贾母 凤姐 王夫人 贾政 贾珍 尤氏 贾琏 邢夫人 贾蓉 贾环 贾迎春 贾芸 秦可卿 林之孝 贾芹 赵姨娘 王子腾 丰儿 巧姐儿 赖大 尤二姐 司棋 小红 尤三姐 王仁 周瑞 李婶娘 佩凤 贾敬 周姨娘 秋桐 娄氏 张华 尤老娘 王钏 王信
community 1: 薛姨妈 薛蟠 冯紫英 詹光 文杏 柳湘莲 程日兴 蒋玉菡
community 2: 贾兰 秦钟 贾蔷 贾菌 贾蓉 贾琼 贾芝 贾萍 贾珩 贾芳 贾琛 贾荛 贾藻 贾效 贾芬 贾敦 贾藻 贾敕 贾瑞
community 3: 宝玉 黛玉 宝钗 贾探春 贾惜春 紫鹃 雪雁 史湘云 李纨 李纹 邢岫烟 吴新登 翠墨 金钏 李贵 薛宝琴 妙玉 贾雨村 秋纹

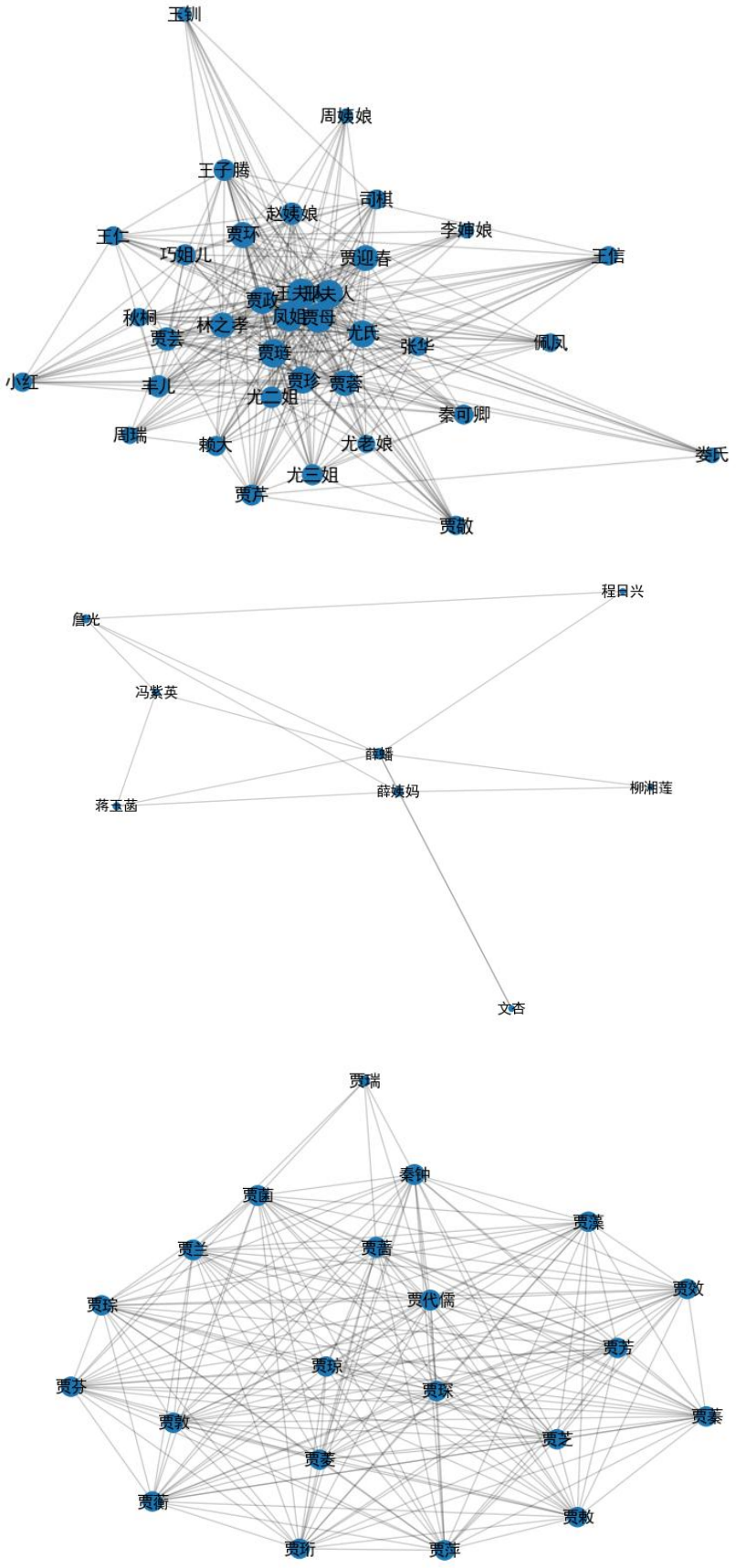
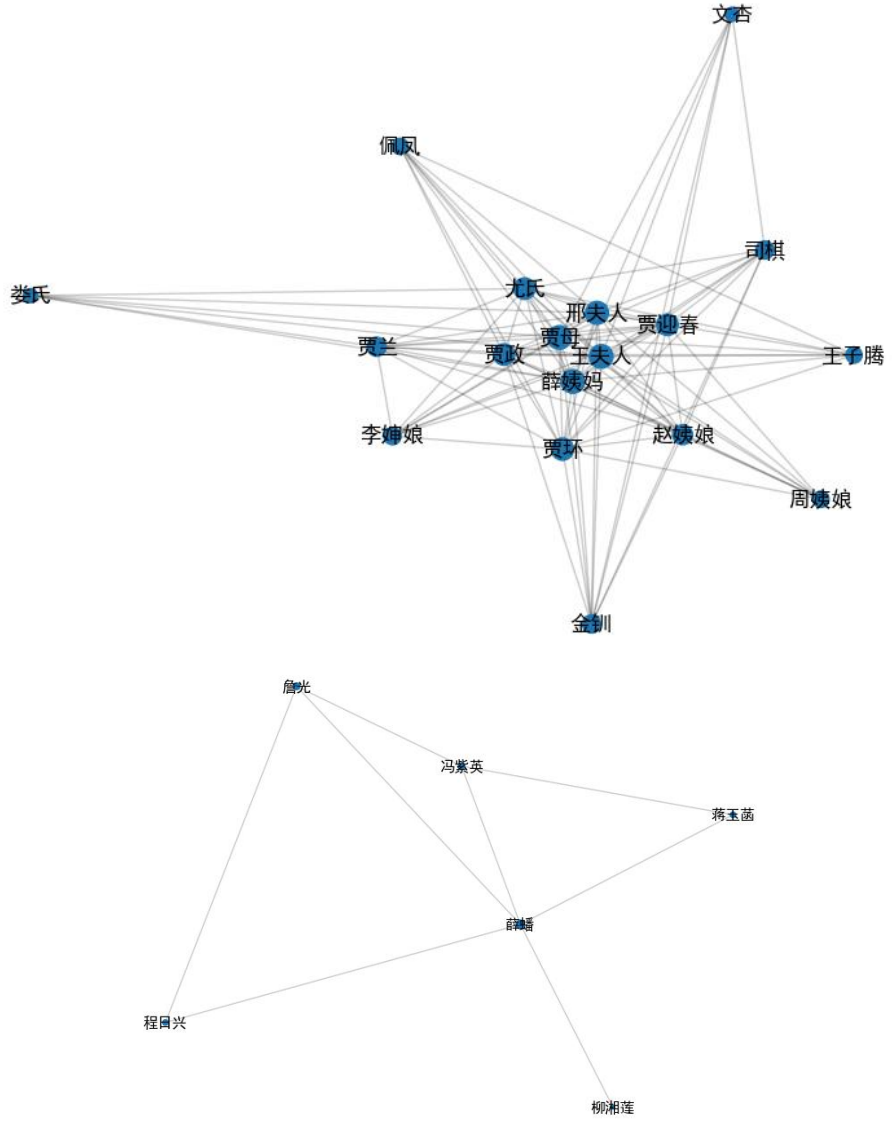


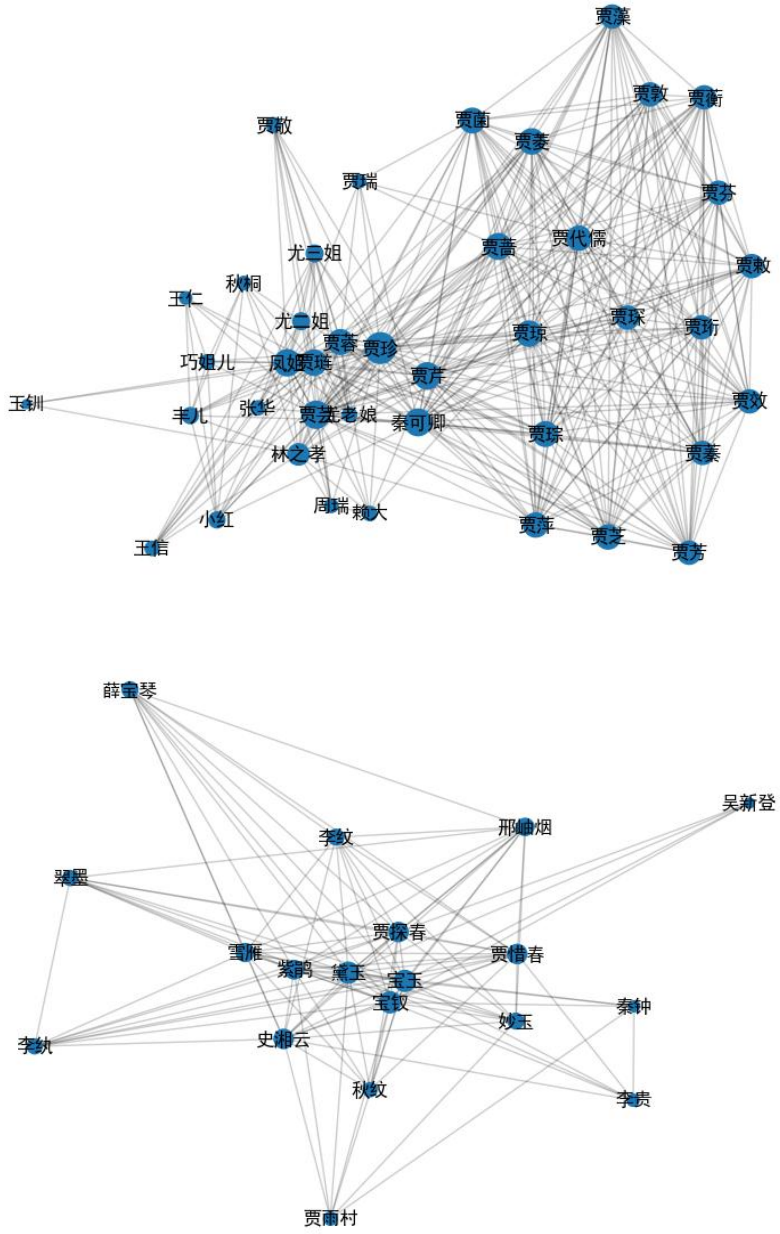
图 5 社区发现结果 1

而如图 6，我们可以发现第一次社区发现的结果有所改变，尤其是原来的 community 0

与 community 3。权势社区与贾氏子弟社区被打乱。同样，这一次的社区发现也具有很强的解释性。Community 0 为以贾母为首的小集团；community 1 位以凤姐为首的小集团；community 2 不变；community 3 为故事主体主角们，以宝玉黛玉宝钗为首的年轻人们。

其中，迎春从 3 社区变为了 0 社区，我们也可以从她的个性中推测一二，别名“二木头”，迎春身为庶出的小姐最终因父亲贾琏欠债被卖给“中山狼”抵债，新婚一年就虐待而亡。她在贾府中也颇为懦弱，并不像其他公子小姐一般风雅浪漫，而是更讨好贾母王夫人等人，为人胆小怯懦。






```
community 0: 贾母 王夫人 贾政 尤氏 邢夫人 薛姨妈 贾环 贾兰 贾迎春 赵姨娘 王子腾 司棋 金钏 李婶娘 文杏 佩凤 周姨娘 娄氏
community 1: 凤姐 贾珍 贾琏 贾蓉 贾芸 秦可卿 林之孝 贾蔷 贾芹 贾琮 贾蓼 贾琼 贾菌 丰儿 巧姐儿 赖大 贾代儒 贾芝 尤二姐 贾萍 贾蓁 贾藻 贾蘅 贾敦 贾敷 贾芬 贾珩 贾效 贾琛 贾芳 小红 尤三姐 王仁 周瑞 贾敏 秋桐 张华 王信 玉钏 尤老娘 贾瑞
community 2: 薛蟠 冯紫英 詹光 柳湘莲 程日兴 蒋玉菡
community 3: 宝玉 黛玉 宝钗 贾探春 贾惜春 秦钟 紫鹃 雪雁 史湘云 李纹 李纨 邢岫烟 吴新登 翠墨 李贵 妙玉 薛宝琴 贾雨村 秋纹
```

图 6 社区发现算法 2

其中的一次实验中，获得了这样的结果：

```
community 0: 贾惜春 妙玉
community 1: 凤姐 贾珍 尤氏 贾琏 贾蓉 贾兰 贾芸 秦可卿 秦钟 林之孝 贾蔷 贾芹 贾琮 贾蓼 贾琼 贾菌 王子腾 丰儿 巧姐儿 贾代儒 赖大 贾芝 尤二姐 贾周瑞 贾敏 佩凤 秋桐 娄氏 张华 王信 贾瑞 玉钏 尤老娘
community 2: 宝玉 贾母 王夫人 黛玉 贾政 宝钗 薛姨妈 贾探春 紫鹃 雪雁 史湘云 李纨 李纹 邢岫烟 翠墨 金钏 李贵 薛宝琴 李婶娘 贾雨村 秋纹 文杏
community 3: 薛蟠 冯紫英 詹光 柳湘莲 程日兴 蒋玉菡
```

图 7 社区发现算法 3

其中，惜春与妙玉成为了一个单独的社区。红楼爱好者对此一定不陌生，看起来觉得毫

无关联的两个人实际上在本书中被曹雪芹安排了一样的命运，冥冥之中都与佛法有缘。在许多红楼相关的文献中，红学家们都对二者进行了归类对比分析。而这是我们普通读者很难从自己的阅读中获得的信息。

多次社区发现能够帮助我们从更多角度去分析红楼故事。诸如此类，不同的社区发现，每一次的人物变化都可以引起读者的思考，并且从文中尝试找到合理的解释。而这正是社会网络分析的高效神奇之处。

五、总结

使用技术辅助中国传统艺术文学研究，将为研究打开新的思路，一方面提高了研究效率，一方面帮助我们发现许多未曾注意到的细节。作为一名红楼爱好者，在这次的实验中一边将红楼知识与社会网络分析技术相结合，一边将社会网络分析结果尝试用红楼知识进行解释。这种独特的体验让人记忆深刻，十分新鲜。如果时间允许，我将再拓展此项目，用于《百年孤独》《尤利西斯》等社会网络复杂的书籍中。

附录

代码见 github: https://github.com/XianWoo/SNA_Dream_of_the_Red_Chamber

参考文献

[1] 数据森麟《“水泊梁山”互联网有限公司一百单八将内部社交网络》https://mp.weixin.qq.com/s/OpR_FXt2pDdrj6U4JmIcDw

[2]《用 python 分析<三国演义中的社交网络>》https://blog.csdn.net/blmoistawinde/article/details/85344906

[3]《红楼梦主要人物词云》https://www.jianshu.com/p/1477bf14d2c9

[4] 曹雪芹《红楼梦》在线阅读 https://www.kepub.net/book/51001

[5]《用 python 绘制红楼梦词云图，竟然发现了这个》https://cloud.tencent.com/developer/article/1160886

[6]《Python 社区发现-louvain-networkx 和 community》https://blog.csdn.net/qq_34356768/article/details/104888579