互联网媒体数据的社交网络挖掘

杨昆霖 章晓慧 王勇程

摘 要 近年来，对社会网络的研究受到了广泛关注。现实中社会关系的演变、社会人物影响力的作用等社会科学领域的研究热点逐渐在数据科学领域得到了理论支持和理论验证。本文利用了新闻网站中跨度近十几年的新闻数据，以时事人物为节点构造网络图，利用了 PageRank 算法以及连通性、聚合系数、中心性等概念分析了网络图的静态特征并验证了社会网络的小世界理论。同时从信息效应与网络效应分析了网络的动力学特征，得到了关于网络特征与网络动力来源的相关结论。

关键词 社会网络图；小世界理论；网络动力学；信息效应；网络效应

Social Network Mining Based on Internet News

Kunlin Yang Xiaohui Zhang Yongcheng Wang

Abstract In recent years, research on social networks has received widespread attention. In reality, the research hotspots in social sciences, such as the evolution of social relations and the influence of social figures, have gradually gained theoretical support and theoretical verification in the field of data science. In this paper, we use news data of news websites spanning nearly ten years to construct a network graph with current characters as nodes, and use the PageRank algorithm and the concepts of connectivity, aggregation coefficient, centrality to analyze the static characteristics of network graphs and verify the social Small-world network theory. At the same time, the dynamic characteristics of the network are analyzed from the information effect and the network effect, and the relevant conclusions about the network characteristics and the source of the network power are obtained.

Key words Social network; six degrees of seperation; dynamics network; information effects; network effects

1. 引言

我们的数据来自于新闻网站，新闻跨度大约为从 2004 年到 2017 年十多年时间，其中提取出了新闻中出现的人名和机构名字。依据课堂上所学的社会网络动力学知识，我们希望借助该数据集分析时事人物网络图的特征以及信息效应与网络效应。

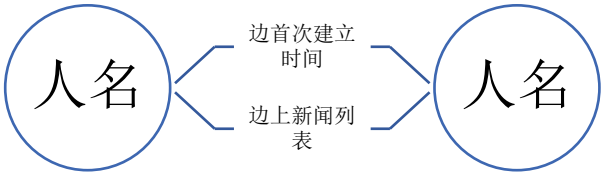
本文大致分为如下部分展开，首先是新闻报道到网络图的构建方法，其次是网络图的静态特征分析，再次是网络图的动态特征分析，最后做一总结。

2. 网络图构建

首先，先明确需求，我们网络图的构建目的是为了能够反应政府要员在网络图中节点重要程度的动态信息，其中的辅助因素是新闻中一起出现的组织名，动态因素的体现是新闻出现的时间，网络关系的构建依据是是否在一个网络中一起出现过，两人之间的关系不存在方向。

基于以上需求，我们需要的图设计大致如下：

- 节点为人名，无其他附加信息。
- 边来源与共现的新闻，只要两个人在一个新闻中出现过就代表这两个人存在一条边。边中需要包含边建立时间和边上的新闻列表，其中边的新闻列表数量作为边的权重。



网络图的结构如上所示。

同时，节点上还需要留出接口存放对应的PageRank、聚集系数、中心性等值，便于后续查询。

3. 静态网络分析

3.1 基于PageRank算法的用户影响力计算

随着互联网的发展，我们可以观察到许多网络大V在社交媒体上已经拥有巨大的影响力。这些大V的每一条微博都能引起巨量的粉丝转发、用户评论，每天微博热搜榜上起起伏伏的明星标签正是明星“影响力”的的具体体现。而节点的影响力也是网络图的分析要素之一。社会网络中蕴含着丰富的用户信息及用户间互相作用的链接关系信息。通常可以用图表示的多关系数据集来代表社会网络，图上的节点表示对象，边表示对象之间相互作用的关系。我们可以将社会结构用抽象的网络图表示出来，并结合数学方法具体、理性地计算和研究社会网络中特定成员的影响力。其中 PageRank 算法为我们提供了一种计算节点影响力的度量。

3.1.1 PageRank 算法

PageRank 是 Google 用来确定一个页面重要性的经典方法之一，其认为一个节点(网页)的重要性（PageRank 值）是由其他节点(网页)的重要性(PageRank 值)决定的。PageRank 值的定义为：假设网页 A 有网页$T_1, T_2, \ldots, T_n$中的链接指向它，则网页 A 的 PageRank 值的计算为:

$$PR(A) = \frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \cdots + \frac{PR(T_n)}{C(T_n)}$$

其中，$:PR(A)$ 即网页 A 的 PageRank 值，其大小表示该网页的重要性。$T_1, T_2, \ldots, T_n$为网页 A 的链入网页；$PR(T_i), (i = 1,2, \ldots, n)$ 是指网页 T_i 的 PageRank 值。$C(T_i), (i = 1,2, \ldots, n)$是指网页$T_i$的的链出网页的数量;$\frac{PR(T_n)}{C(T_n)}$为网页中有链接指向网页 A 的网页$T_n$投给网页 A 的网页级别值，也称为 mini—PageRank。

实际应用时，往往引入一个阻尼因子，以刻画“随机游走”(random walk) 行为。PageRank 认为，一个用户在通过点击链接来访问(浏览)下一个节点(网页)，这种行为终将停止，而以一个随机的概率跳转到其他任意一个节点。定义用户继续这种行为的概率为阻尼因子，记为d。而$(1-d)$则表示用户随机跳转到任意一个节点的概率。关于阻尼因子的取值，有许多相关的研究，一般取其值为0.85。于是，对于任何一个节点(网页) u，其 PageRank 值可计算如下: $PR(u) = (1-d) + d \cdot \sum_{v \in Bu} \frac{PR(v)}{C(v)}$

其中，Bu表示有链接指向节点u的所有的节点的集合，$PR(v)$表示节点v的 PageRank 值，$C(v)$表示节点v链出节点的数目。

由上述讨论可知，每个页面的 PR 值(PageRank 值)取决于指向它的页面的 PR 值。但是预先并不知晓其他指向它的页面的 PR 值，直到计算出指向它的所有页面的 PR 值，以此类推。总体上，网页间的链接是一个传递、循环的过程，在计算时需要用迭代算法。通常的做法是给各个网页赋一个初始 PR 值，这个初始 PR 值的大小无关紧要。在这个初始值的基础上进行迭代计算，记录下每次循环后算出的数值，随着循环次数的增加，这个数值越来越接近最终的估计。当每个页面的 PR 值计算完成时，所有页面的 PR 值平均值：“平均化概率分布”为 1。有些情况下最终得到的 PR 值平均值不等于 1，原因是有些网页没有回馈链接(反向链接)，降低了整体的 PR 值。

3.1.2 实验过程

我们首先用 python 工具包 networkx 构建了网络图，并调用了 networkx 中的pagerank(G,alpha)方法计算了图中各节点的 PageRank 值，其中取阻尼系数 alpha=0.85。

我们统计了 PageRank 值居于前 20 位的节点，分别是：

PR 值前 20 名节点及其 PR 值

节点名称	Pagerank 值
习近平	0.004400573849709015
胡锦涛	0.0025402544469366626

温家宝	0.0016185455366446828
李克强	0.0014675356517968134
徐建平	0.0010084754195465741
程永华	0.0009252695776524445
邱小琪	0.0008914963072860288
伍江	0.0008672370953726173
裴钢	0.0008496556980669535
杨洁篪	0.0008384861123903247
吴志强	0.0008353056156645532
罗林泉	0.0008171299876240572
杨厚兰	0.000812907427654458
杨贤金	0.0008066463700183865
孔子	0.0008032751141713536
邓小平	0.0007518406772033299
马锦明	0.0007414028333341172
诺贝尔	0.0007023361816094729
季志业	0.0006828418669212253
吴邦国	0.000660486953353377

我们又统计了 PageRank 值居于后 10 位的节点，分别是：

PR 值后 10 名节点及其 PR 值

节点名称	Pagerank 值
宋霖	1.9067724008916141e-06
凯瑟林·赫德尔斯顿	1.9067724008916141e-06
王书君	1.9076437046253516e-06
刘文方	1.9176752771008273e-06
金颖颖	1.918799393335631e-06
柏贵华	1.9229477961586306e-06
Chemins de Fer	1.9333164252127357e-06
黑国庆	1.9440733845425167e-06
谭力勤	1.946355807282094e-06
Ricardo Sealy	1.9470473633359597e-06

3.1.3 结果分析：

结合直观印象，这些 PageRank 值高的人物都有极高的政治地位和极大的政治影响力，因此他们的文章经常在政治新闻类的刊物中发表。这与我们的计算结果十分符合。

从网络结构图中，我们也可以发现他们与非常多的节点相连，并且处于图结构的中心位置。而 PageRank 值倒数的节点，它们只与少量的节点相连，并且处于图结构的边缘地带。

可以认为，PageRank 值从量化的角度具体表示了一个节点在网络结构中的影响力。

3.2. 小世界理论的验证

3.2.1. 小世界理论

在全球范围内，任意两个人相识的概率有多大？其中需要多少中间人才能将两人联系起来？1967 年，哈佛大学社会心理学家 Stanley Milgram 进行了一项连锁信件实验，提出了著名的“六度分离”假设，在学术上称为“小世界现象”。“小世界现象”目前还没有精确的定义，较为合理的解释为：网络中任意两点的平均距离 L 随网络大小（结点数 N）呈对数增长，即 L～lnN，也就是说网络中结点数量增加很快时，L 的变化相对较慢，这种现象称为“小世界现象”。在 Stanley Milgram 的实验中，信件传递的平均距离为 6，即通过 6 个人就可以将两个人联系起来。

六度分割（也叫“六度空间”）的概念由此而来。这个连锁实验，体现了一个似乎很普遍的客观规律：社会化的现代人类社会成员之间，都可能通过“六度空间”而联系起来，绝对没有联系的 A 与 B 是不存在的。这是一个更典型、深刻而且普遍的自然现象。

六度分隔的现象，并不是说任何人与人之间的联系都必须要通过六个层次才会产生联系，而是表达了这样一个重要的概念：任何两位素不相识的人之间，通过一定的联系方式，总能够产生必然联系或关系。显然，随着联系方式和联系能力的不同，实现个人期望的机遇将产生明显的区别。

3.2.2 实验过程

我们首先用 python 工具包 networkx 构建了网络图：两个人如果出现在同一篇新闻中，则假设这两个人有联系。两个人的联系强弱可以通过共同出现的文章的数目来表示。例如，假设 A 和 B 在 10 篇新闻中同时出现过，则 A-B 之间的边的权重为 10。

Networkx 的 shortest_path_length()方法为我们提供了求两节点间最短路径长度的方法。我们先将图划分成若干连通分量，并计算每个连通分量的平均最短路径长度，得到结果如下（选取最大的 10 个连通分量为例）：

连通分量平均最短路径长度

边数	节点数	平均最短路长
1729550	78005	3.578815
158	21	1.257142
45	10	1
36	9	1
36	9	1
28	8	1

22	8	1.214285
22	8	1.21428
21	7	1
21	7	1

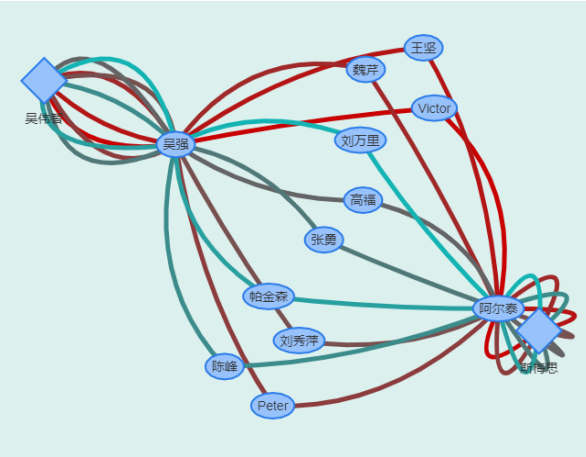
3.2.3 结果分析

计算结果表明，该网络图的各连通分量平均最短路径长度均小于 4，很好地验证了小世界现象。

同时，我们也自己利用 python 语言实现了方法 find_route(source,target)。在前端界面对输入节点名称 A 和 B，该方法可以找出 A 和 B 之间的前 10 条最优路径（路径越短越优，路径长度相同时，按照路径上权重总和由大到小排序）。

我们特地选取了 pagerank 值处于倒数且无直接联系的两个人物'吴伟智'和'斯梅思'，调用 find_route()方法求他们之间的前 10 条最短路径。最终得到结果，如图所示：

吴伟智-斯梅思前 10 条最短路径



表面上关系十分疏远且 pagerank 值均很小的两人，原来也可在 4 步内联系到彼此！

3.3聚合系数

3.3.1 聚合系数

设某个节点有 k 条边，则这 k 条边连接的节点之间最多可能存在的边的个数为$\frac{k(k-1)}{2}$。用实际存在的边数除以最多可能存在的边数得到的分数值，定义为这个节点的聚合系数。所有节点的聚合系数的均值定义为网络的聚合系数。聚合系数是网络的局部特征，反映了相邻两个人之间朋友圈子的重合度。我们可以在现实生活中找到例子：人们的朋友圈子在很大程度上重叠，也就是说你的两个朋友可能本来也相互是朋友，这个重叠性即是聚合系数。

3.3.2 实验过程

我们首先用 python 工具包 networkx 构建了网络图，并调用了 networkx 中的 clustering 方法计算了图中各节点的聚合系数值，并将结果按降序排列。

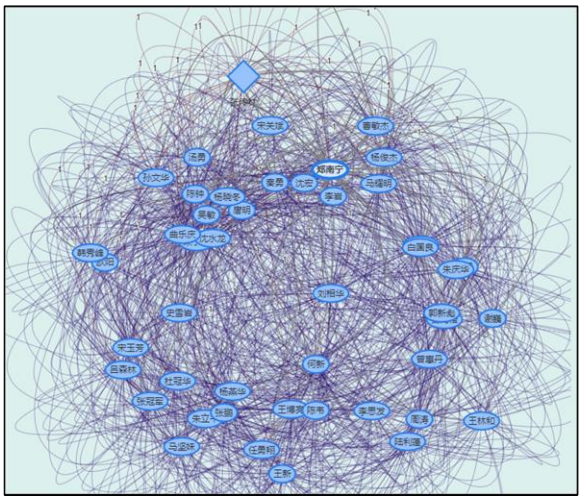
为验证聚合系数的实际意义，我们选取了聚合系数较大的节点“张传林”和聚合系数较小的点“马一里”，判断他们各自朋友之间的联系紧密程度。

例如，节点 A 有朋友$F_1, F_2, \ldots, F_k$，我们考察F_i与$F_j(i \neq j)$之间的联系。这个联系用F_i与F_j是否有联系以及最短路径长度来度量。同时，通过动态网络分析来研究他们朋友之间的社会网络联系演变。

3.3.3 结果分析

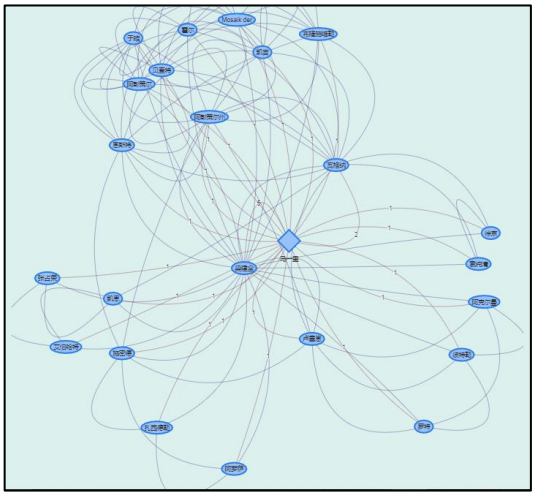
对于聚合系数较大的节点“张传林”，通过他的节点关系图，我们可以发现他的朋友之间联系较为紧密。并且通过动态网络分析，可以发现他原本没有联系的朋友也逐渐建立了联系。已存在的联系也逐渐加强。

“张传林”社会网络关系图



对于聚合系数较小的节点“马一里”，通过他的节点关系图，我们可以发现他的朋友之间联系较为松散，且朋友之间新建立的联系较少。

“马一里”社会网络关系图



以上实验结果印证了聚合系数的实际意义。

3.4 中介中心性

3.4.1 中介中心性

对于节点重要性的解释有很多种，不同的解释下判定中心性的度量指标也有所不同，但当前最主要的度量指标为点度中心性（Degree Centrality）、接近中心性（Closeness Centrality）、中介中心性 (Between Centrality) 、特征向量中心性（Eigenvector Centrality）四种。其中， 中介中心性的定义是在路径基础之上的。下式中，σ_{st}表示节点s到节点t的最短路径数量，$\sigma_{st}(v)$表示节点s到节点t的最短路径中经过节点v的数量。根据 Freeman(1977)的定义，节点v的中介中心性$C_B(v)$为：

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

3.4.2 实验过程

我们通过中介中心性的快速计算方法计算了原图中每个人物的中介中心性值，并将结果按降序排列。

3.4.3 结果分析

中介中心性前十位人物

人名	中介中心性
胡锦涛	143525
习近平	116060
宋克宁	49686
温家宝	20843
毛泽东	20420
奥巴马	20094
江泽民	14837
梁百忍	14588
王天明	14443

中介中心性衡量了一个人作为媒介者的能力，也就是占据在其他两人快捷方式上的重要位置的人，他拒绝媒介，这两人就无法沟通。占据这样的位置愈多，就愈代表这个人具有很高的中介性，愈多的人联络时就必须要透过他。

可以发现，中介中心性前十位的人物都是重要的政府官员或名人，属于社会政治网络的“枢纽”，肩负着完成两群隔离的人之间信息交互的责任，影响社会网络的信息传递。

4. 动态网络分析

4.1 动态网络定义

我们需要分析的目标是网络中的动态效应，先对我们的网络进行相关定义的阐述。

定义 1 个体行为

个体行为是指网络中的每一个时事人物，选择是否与另一位时事人物共同出现在一个新闻当中，也就是其是否会产生一条边。对于每一个个体，他们的选择空间都为该数据集中出现的所有时事人物，但我们只考虑个体选择之前未连接过边的行为。

定义 2 群体行为

群体行为指的是，个体是否有模仿他人的行为。也就是说，大众都倾向于选择建立关系的时事人物，个体会被大众行为所影响，从而选择与其建立关系。其中，动态的区间设置为一年，每一年的网络图中包含此前所有年份的关系，即网络图只会增加复杂性而不会减少。

定义 3 群体行为对象

群体行为的目标对象指的是网络图中，重要性最高的几位时事人物。只有个体对这些时事人物有建立关系的趋势并且此前没有建立过关系，才能算作是一个动态行为。

定义 4 信息效应

信息效应是动态行为产生的原因之一，也就是其他时事人物的行为传递了他们所知道的信息，个体观察别人的行为并模仿着去做，即便有时会违背自己的个人信息。反映在网络图上就是更多个体与重要时事人物连接，重要时事人物邻居的增加。

定义 5 网络效应

群体行为的产生原因之二是基于直接受益效应，也叫做网络效应，即对于某些决定，如果能

与他人的决策行为保持一致能够带来直接的利益，那么个体就会倾向于选择这个决策行为。

定义 6 个体利益

本网络图中所有个体追求的利益是使得自己在网络中的重要性增加。

4.2 动态网络行为衡量方法

在动态网络行为的定义之上，我们给出动态网络中的衡量指标，部分理论基于前文所述的静态网络分析。

定义 1 时事人物重要性

时事人物的重要性衡量基于网络静态分析中的 PageRank 值。PageRank 值以年为更新单位，每年 PageRank 值在前一百的，视为当年的重要时事人物，也就是动态行为对象。

定义 2 信息效应度量

由信息效应的定义可知，信息效应的对象是重要的时事人物，该时事人物重要程度位居前茅时，如果其边的增加速度大于往年的增加速度，就可以认为总体有与时事人物建立关系的趋势，那么就能推出个体有模仿总体的行为，从而体现出信息效应。

由于动态计算单位为年，增加速度即可定义为每年新增加的邻居数量$NewNeighbour_{year}$，如果当年$year_i$该人物成为重要时事人物，那么所要比较的就是下一年$year_{i+1}$与前一年$year_{i-1}$的新增加邻居数量，也就是下式：

$$\text{InfoCascade} = \begin{cases} True, & NewNeighgour_{year_{i+1}} > NewNeighgour_{year_{i-1}} \\ False, & NewNeighgour_{year_{i+1}} \le NewNeighgour_{year_{i-1}} \end{cases}$$

定义 3 网络效应度量

网络效应的对象是与重要时事人物新建立关系的个体，如果该个体因为与重要时事人物建立关系而自身的重要性得到增加，那么就认为该个体利益增加，否则利益减少。为了去除重要时事人物的影响，在计算中会先扣除个体与重要时事人物新建立的边。

也就是说，对于个体 *Person*，若其与若干重要人物 *Tops* 新建立关系，那么首先先计算其前一年的影响力$PageRank_{before}$，再将该 *Person* 今年所建立的所有新的边加上，并剔除与 *Tops* 建立的所有边以排除 *Tops* 的 PageRank 值所带来的影响，计算结果为$PageRank_{after}$，最终度量如下式：

$$\text{NetworkCascade} = \begin{cases} True, & PageRank_{after} > PageRank_{before} \\ False, & PageRank_{after} \le PageRank_{before} \end{cases}$$

4.3 动态网络行为实验结果

由于网络中年份新闻数量分布不同，早期新闻较少，从数据分布中我们决定将 2005 年以及之前的所有新闻构建得到的网络图作为动态网络行为分析的基础，在此基础之上，再分析网络动态行为，以年为单位进行考量。

4.3.1 信息效应度量结果

利用前文所述的衡量方法，我们计算 2005 年之后，每年重要的时事人物新增加的关系数量，得到结果如下表：

信息效应度量结果

是否存在信息效应	样本数量	样本百分比
存在	368	43.35%
不存在	481	56.65%
总计	849	100%

由此可见存在于不存在信息效应的时事人物数量较为接近，并且存在信息效应的任务比例比不存在的比例更小，说明该网络中信息效应不明显。

4.3.2 网络效应度量结果

与信息效应类似，我们在相同时间间隔上，计算了对应节点的网络效应，结果如下表：

网络效应度量结果

是否存在网络效应	样本数量	样本百分比
存在	11018	67.39%
不存在	5332	32.61%
总计	16350	100%

由上表可以看出，有网络效应的个体数量显著多于没网络效应的个体数量，同时，我们计算了网络效应的增加比例，也就是网络效应的增加量与原网络效应的比值，得到结果如下表：

网络效应增加比例

网络效应增加方向	平均变化比例
利益增加	68.81%
利益减少	12.71%

从上表可以看出，网络效应带来的利益增加的比例远远大于网络效应带来的利益减少的比例，并且利益减少的平均比例相对较低，可以视为正常波动。

为了进一步地验证我们的假设，我们取网络效应增加最多的前 20 名和减少最多的前 20 名做一比较：

网络效应改变前 20 名

正向改变前 20		负向改变前 20	
时事人物	改变比例	时事人物	改变比例
史明德	37.28	阿斯杜蒂	0.61
宋学锋	19.00	胡里奥	0.60
洪磊	18.29	袁方	0.58
李昕	18.20	谷岩昭	0.58
屈生武	14.17	徐会仲	0.58
许镜湖	14.08	帕萨特	0.58
李瑞宇	14.00	齐聚	0.57
杨贤金	13.88	江泽慧	0.56
达赖	12.81	段宁	0.56
李瑞佑	11.86	乐玉成	0.53
许仲林	11.75	徐建国	0.51
叶大波	11.45	刘少奇	0.50
刘海星	10.70	吴仪	0.50
张俊	10.57	爱迪生	0.50
丁伟	10.26	波罗申科	0.50
郑加麟	10.25	甘贾·普拉诺沃	0.50
张军	10.09	玛莎	0.50
杜青林	10.07	陈震	0.50
陈建福	9.75	盛华仁	0.49
方守恩	9.63	工勤	0.48

由上表可以看出，网络效应强的人，自身利益能得到高达 37 倍的增长，而网络效应最差的人，自身利益也仅仅只降低 0.61，更不及利益增长的平均水平。从而可以看出，网络效应是该网络中动力产生的主要原因，因为网络效应，大部分人的利益得到增加，且网络效应带来的利益下降并不一定是网络效应的反例，很有可能是自身产生的波动所造成的。

4.4 动态网络分析结论

经过上述实验分析，我们可以知道网络中信息效应不显著，但是网络效应显著。因此可以做出如下推断：

- 重要的时事人物并不会因为重要性排名前茅而获得级联效应，没有因此变得富者愈富。
- 个体与重要时事人物建立连接，就算去除重要时事人物对个体的影响，个体依然能获得不少利益。
- 综合信息效应与网络效应，可以得出网络效应是网络图动力的主要来源。
- 网络中的个体大多数是自驱动的，选择会给自己带来利益的人连接，而不会一味选择连接人数多的重要任务。

5. 总结

我们对新闻网络图进行了静态分析与动态分析，经过静态分析，利用了 PageRank 算法以及连通性、聚合系数、中心性等概念分析了网络图的静态特征并验证了社会网络的小世界理论。从动态分析中，我们发现网络图的动力来源是网络效应并且信息效应并不显著。

参 考 文 献

[1] Ulrik Brandes: A Faster Algorithm forBetweenness Centrality. Journal of Mathematical Sociology 25(2):163-177, 2001.

[2] TONG Ting-ting，SONG Yi: The small-world theory and its application in Internet

[3] ZHANGJing: A Review of the Research on the Small World Theory in China

[4] Ma Feng: Academic Impact Based on Page Rank Algorithm: an Empirical Study

组 员 分 工

名字	工作 1	工作 2	工作 3
杨昆霖	原始数据清洗与重构	静态图设计与基础任务实现	动态网络实验与结论分析
章晓慧	Web 服务器开发搭建	网页前端美化与处理逻辑	网络图 js 可视化实现
王勇程	用户影响力计算与小世界理论验证	节点中介中心性与聚集系数计算	图计算相关算法效率优化