基于新闻共现的社交网络分析与挖掘

付廷琛[1)]

[1)](中国人民大学信息学院, 北京市 中国人民大学 100872)

摘 要 近年来，有关社交网络的分析的挖掘受到了广泛关注，将图论知识和大数据方法应用于社会学研究之中的研究方法和模式也受到推崇。本文以中国政府门户网站中采集到的 2016-2020 年间的约三万条新闻作为数据集，以两个人物在同一新闻共现即认定两者有朋友关系为方法，构建出社交网络。本文所做的研究工作主要有：通过设计算法找到任意两个人之间的若干条最短路径，通过计算平均路径长度验证小世界现象；计算所有结点的聚集系数并就其中的个例进行分析；计算所有结点的中介中心性并排名得到中心性最高的若干个节点。特别地，本文使用 Jaccard 距离来度量社交网络上任意相邻两个人之间的距离，并证明了 Jaccard 距离在此应用场景之下是良定义的。

关键词 社交网络；小世界现象；节点中心性；聚集系数；Jaccard 距离

Social Network Analysis and Mining Based on News Co-occurrence

Tingchen Fu

[1)](School of information, Renmin University of China, Beijing City 100872, China)

Abstract In recent years, the mining of social network analysis has received widespread attention, and the research methods and models that apply graph theory knowledge and big data methods to sociological research have also been praised. This article uses about 30,000 pieces of news collected from the Chinese government portal website from 2016 to 2020 as a data set, and uses the method that two characters co-occur in the same news to determine that they have a friend relationship to construct a social network. The research work done in this article mainly includes: finding several shortest paths between any two people through design algorithms, verifying the small world phenomenon by calculating the average path length; calculating the aggregation coefficients of all nodes and analyzing individual cases; Calculate the betweenness centrality of all nodes and rank the nodes with the highest centrality. In particular, this article uses Jaccard distance to measure the distance between any two adjacent people on social networks, and proves that Jaccard distance is well-defined under this application scenario.

Key words social network; small world; node centrality; clustering coefficient; Jaccard distance

1. 构造社交网络

我们研究的数据来源于中央人民政府门户网站(www.gov.cn)中 2016-2020 年之中的 29699 条新闻数据，从中提取出每一条新闻内容之中出现的所有人名、地名、机构名等实体，并合理假定出现在同一篇新闻中的实体之间具有社交网络上的联系，以此作为研究前提来构建社交网络。

1.1 数据预处理与实体抽取

由于分词工具 jieba 内含的 posseg 在实体识别方面的表现欠佳，更确切地说对于人名的识别准确率偏低，会将“黑名单”，“小微”，“司”等明显不属于人名的名词或者单字错误地识别为人名。这样的结果显然无法满足我们的研究要求，会极大地降低我们研究成果的可信度。因此，我转而使用 thulac 作为预处理阶段进行分词和实体识别的工具。为了验证数据的正确性以及 thulac 进行分词和实体识别的准确率，统计分词结果之中所包含的热门人物和机构。为了避免在某一篇特定的新闻稿之中某一个人物或者地点大量重复多次，影响整体词频排序的结果，所以在统计的时候只统计每一个专有名词在

不同的新闻内容之中出现的次数。最终得到的人名的出现频率前十名如下表所示：

人物	出现频率
习近平	5777
李克强	4352
王毅	1645
何立峰	706
杨洁篪	622
丁薛祥	481
韩正	469
小微	439
汪洋	419
胡春华	396

从表中可以看到，出现次数最多的人名为“习近平”，出现在 5777 篇新闻内容之中；其次为“李克强”，出现在 4352 篇新闻内容之中，除了“小微”，等极少数的人名识别错误之外，绝大多数的人名实体识别都是准确的，并且词频排序结果符合预期。

地点	出现频率
中国	13952
北京	10001
上海	2496
美国	1554
河北	1283
浙江	1213
河北省	1211
山东	1094
中华人民共和国	1089
江苏	1070

机构	出现频率
新华社	17581
国务院	8511
党中央	2940
中共中央	2587
财政部	1465
联合国	1298
中央军委	770
公安部	747
教育部	612
外交部	465

对于地名和机构名的词频统计排序结果同样符合预期和常识，但是 thulac 无法识别同一个实体的不同名字，如“中国”和“中华人民共和国”，“北京”和“北京市”等。不过我后续的研究并不会进一步地依赖和使用有关地名的信息以构建网络，所以 thulac 同一实体的别名识别问题并不会对我的实验和研究造成太大的影响。

1.2 社交距离重定义

为了构造社交网络，将所有出现过的人名视为节点，所有在同一篇文章之中共现的人名对应的节点之间有无向边相连，并定义边的“关联度”为该边所关联的节点对应的人名在不同的文章中共现的次数。

但是在后续的研究之中我们发现，仅仅使用“关联度”这一属性并不能很好地描述社交网络之中一条边的特征。“关联度”的计算是没有“归一化”步骤的，在计算两个人的关联度时，只是单纯考虑了两个人在新闻中共现的次数，而没有考虑两个人在新闻中出现的总次数。这就会导致，两个人的关联度非常高，可能仅仅是因为两个人都是热门人物，本身各自在新闻之中出现的次数都特别多，但是两个人之间的关联程度并不高，对研究结果造成偏差。其次，“关联度”这种指标并不能很好地对应到图中边的“距离”。在后续的工作中，计算两个人的最短路时需要使用到每一条边上的距离，但是“关联度”和“距离”之间，缺乏简洁又明确的转

化手段。

基于上述论断，我在构造社交网络的过程中，除了计算每一条边的关联度之外，定义并计算了每一条边的“距离”属性。借鉴“邻里重叠度”和两个集合之间的 Jaccard 相似度的概念，我们定义，在新闻社交网络挖掘之中，两个人的 Jacard 相似度为两个人在同一篇新闻中共现的次数与两个人出现过的新闻中的总数的比值。设两个新闻社交网络中的人物是 p1 和 p2，f 为一个一元函数，从人名到集合的映射，能够将一个人物映射到出现该人物的所有新闻的集合，sim(p1,p2)为两个人 p1 和 p2 之间的 Jaccard 相似度，dis(p1,p2)为两个人 p1 和 p2 之间的 Jaccard 距离。那么有

$$sim(p1,p2)=\frac{card\big(f(p1)\cap f(p2)\big)}{card\big(f(p1)\cup f(p2)\big)}$$

$$dis(p1,p2)=1-sim(p1,p2)$$

一般而言，一个距离函数 d(x,y)应当满足非负性，同时公理，对称性，三角不等式，这几个要求的形式化表达如下：

$$d(x,y)\geq 0$$
$$d(x,x)=0$$
$$d(x,y)=d(y,x)$$
$$d(x,y)+d(y,z)\geq d(x,z)$$

下面，我们需要证明 Jaccard 距离符合距离函数的一般要求。

非负性。由于两个集合的交集所包含的元素数量一定小于等于两个集合的并集所包含的元素数量，所以两个人物之间的 Jaccard 相似度一定是介于 0 和 1 之间，因此两个人之间的 Jaccard 距离一定是非负的，介于 0 和 1 之间。

同时公理。由于 sim(x,x)=1，因此有 dis(x,x)=0.

对称性。距离公式之中的集合交运算和并运算都具有对称性，因而 Jaccard 距离本身也具有对称性。

三角不等式。对于三角不等式的证明， Sven Kosub 等人从集合函数的角度给出了一个严谨的数学证明。而我在本文之中将给出另外一种基于哈希的证明。

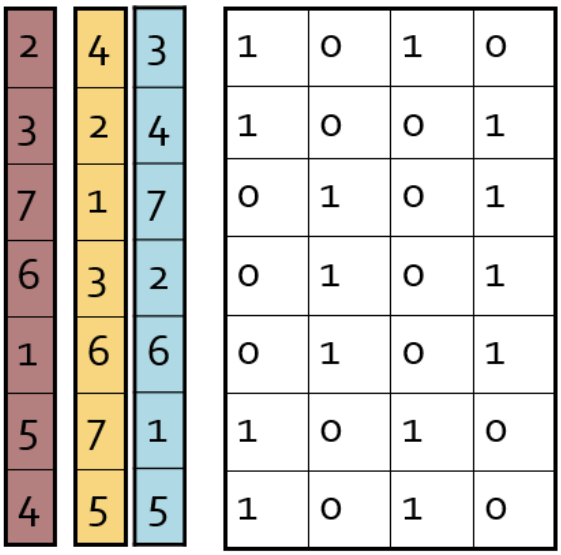
首先，将所有人名与所有新闻稿的共现关系使用矩阵 **A** 来呈现，该矩阵共有 N 行 M 列，其中 N 是新闻总数，M 是所有的人名总数，A_{ij}只会取两个值 0 或者 1。$A_{ij}=1$当且仅当第 j 个人名出现在第 i 篇新闻内容之中。下面我们定义一个特殊的，基于随机重排的“哈希”函数。假设π是一个对矩阵 A 的各行之间的随机重排

$$h_\pi(C)=min_\pi\pi(C)$$

其中 C 是 A 的其中一个列，或者说是一个人名。这个哈希函数简单地理解，就是在经过行随机重排π之后，从上往下数遇到的第一个“1”。下面我们需要证明对于两个不同的人名，或者说两个不同的列，有：

$$Pr[h_\pi(C_1)=h_\pi(C_2)]=sim(C_1,C_2)$$

这里借用 *Data Mining* 之中的一个简单的例子说明。如下图所示，



右侧矩阵是人名-新闻共现矩阵，共有 7 个新闻文档，4 个人名，如果某一个人名在某一篇文档之中出现，那么相应的位置为 1，否则为 0。

左侧三个不同的颜色是三个不同的行随机重排。对于每一个重排序列$a_1,a_2,a_3,a_4\ldots a_7$代表的是现在重排之后的第 i 行就是原来的第$a_i$行。比如说，红色序列<2，3，7，6，1，5，4>代表的是重排后的第一行是原来的第二行，重排后的第二行是原来的第三行，重排后的第三行是原来的七行，其余以此类推。经由哈希函数处理之后，对于第一列，原来的第二行就出现了 1，所以$h_{\pi_1}(C_1)=1$，对于第二列，原来的第三行才出现 1，所以$h_{\pi_1}(C_2)=2$。对于第三列来说，原来的第二行、第三行都没有出现 1，指导第七行才有 1 出现，所以有$h_{\pi_1}(C_3)=3$.对于第四列，同理，在第二行就有 1 出现，所以有$h_{\pi_1}(C_4)=1$。所以，在第一个随机重排后，这四列的哈希值分别是 1，2，3，1。同理，在黄色标注的行随机重排后，这四列的哈希值应当是 2，1，3，1.

在蓝色标注的随机重排之后，这四列的哈希值为 3，1，3，1.

接下来证明方法比较巧妙。C_1和C_2在随机重排π下的哈希值相同，意味着从第一行开始往下遍历的时候，C_1和C_2两个列在同一行中遇到了第一个数字“1”.而我们考虑C_1和C_2两个列从第一行开始往下遍历时遇到的所有可能的情况，总共有以下几种情况：

	C_1	C_2
Case1	0	0
Case2	1	0
Case3	0	1
Case4	1	1

如果在第 i 行遇到情况是 Case1，那么会继续向下遍历第 i+1 行；如果遇到的是 Case2，Case3 或者 Case4，就至少有一列的遍历会停下。而这三种情况下，只有遇到 Case4 时，才会有$h_\pi(C_1) = h_\pi(C_2)$。由于π是一个对矩阵 A 的所有行的一个随机重排，所以我们只需要关注 Case4 在 Case2、Case3 和 Case4 之中发生的概率即可。仔细观察发现，Case4 代表C_1和C_2两个名字同时出现在同一篇新闻之中的情况，Case2 和 Case3 涵盖了只有其中一个人名出现的新闻的情况。因此有：

$$Pr[h_\pi(C_1) = h_\pi(C_2)] = \frac{|Case4|}{|Case2| + |Case3| + |Case4|} = sim(C_1, C_2)$$

还是用刚才的例子来说明。在这个只有三个随机重排的小样例之中，第二列和第四列在黄色和蓝色的随机重排下哈希值相同，哈希值相同的概率是$\frac{2}{3}$。而这两列实际上的 Jaccard 相似度应当是$\frac{3}{4}$.随着随机重排的数量不断增加，两者应当逐渐趋于相同。

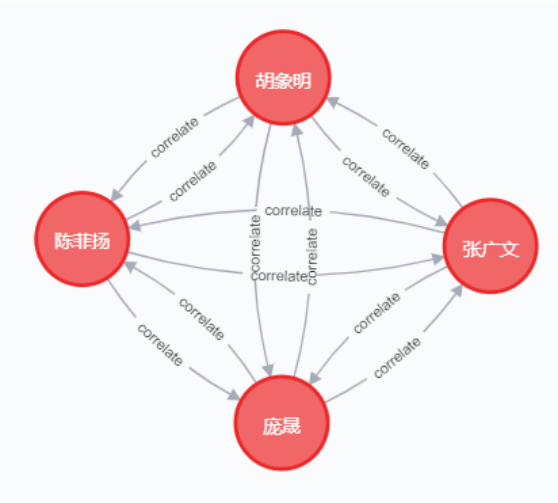
2. 图的基本信息统计分析

跟据上述方法构建生成社交网络，并确定每一条边的“关联度”和“距离”两个属性之后，我们首先对图的基本信息进行了统计和分析。

2.1 图中点数、边数、度数与联通分量

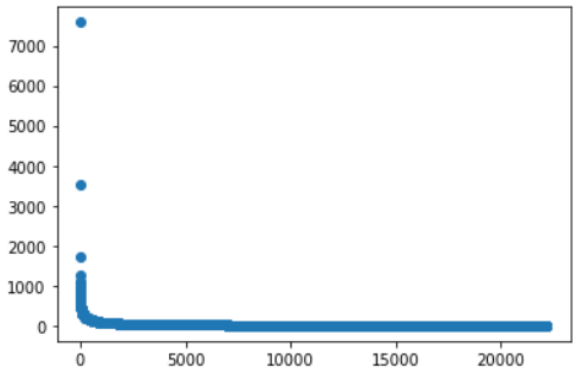
在建立的社交网络图之中，总共有 22187 个节点，255060 条边，由此可以计算稀疏因子保留到小数点后四位的结果是 0.0005，总体来说是一个稀疏图。经过计算，总共具有 184 个连通分量，因此该社交网络不是连通图。最大的连通分量具有 21648 个节点，最小的连通分量仅具有两个节点。通过对各个连通分量的对比观察发现，整个图的大致结构大致可以描述为，中心有一个巨大的连通分量，涵盖了图中绝大部分的顶点和边，其余没有被连通到这个连通分量的节点三三两两之间组成一些特别小的簇，图中没有度数为 0 的孤立点。这个结果佐证了“超大连通分量”的概念假设。“超大连通分量”是一个非形式化的对于包含其中大部分节点的连通分量的称谓。一般来说，一个社交网络之中只会包含有一个“超大连通分量”。因为，假如有两个超大连通分量，很难想象这两个大团体之中的任意两个人都不认识。而一旦有两个来自不同团体的人成为朋友，两个“超大连通分量”就会融合成为一个，这是非常容易做到的。

我随机选取了一个除了超大连通分量之外的一个连通分量，其局部网络结构如图所示：

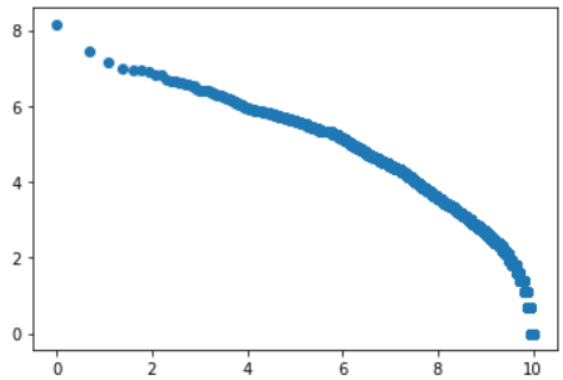


可以看到，这个连通分量中只有四个节点并且彼此之间有联系，而所有节点都与外界其他节点没有任何联系，形成了一个相对封闭的小圈子。

为了详细观察图中的边的稠密程度与边的分布均衡程度，统计每一个节点的度数，依据度数做出所有节点的度数散点图，如下所示：



其中横轴代表的是节以度数作为排序标准，节点的序数，纵坐标代表节点的实际度数。为了让图像更加利于观察，对横纵坐标都分别进行取对数操作，得到处理之后的新图像如下所示：



通过这张图可以大致推测，节点的序数值的对数值同节点的度数的对数值呈现线性相关关系，也就是：

$$log(rank(i)) \propto log(degree(i))$$

或者等价地有:

$$rank(i) \propto degree(i)^k$$

其中 k 为比例系数且有$k < 0$

2.2 节点影响力度量

每个人在社交网络中的地位和影响力并不相同，为了描述和度量这种差别，我们利用图中的无向边来描述节点之间的联系和约束。具体而言，我们使用 PageRank 算法来计算各个节点在社交网络中的影响力大小。PageRank 是 Google 在为他们的搜索引擎返回的网页搜索结果排序时所使用到的网页排序算法，由 Google 的创始人之一 Larry Page 提出并命名。PageRank 最初是一种链接分析算法，为相互之间通过超链接指向的网页分配一个权重，用以衡量这个网页在网页集中的相对重要程度。

PageRank 使用随即冲浪者模型，得到的权重可以解释为一个概率分布，用于表示打开一个网页并随机点击网页中的任意一个链接的人经过若干次点击之后到达任意页面的可能性。我将 PageRank 的思想理念迁移到新闻社交网络之中，为图中的每一个人计算其相对影响力的大小，其基本思想是，一个有影响力的人，会尽可能地同更多的人建立社交联系；一个有影响力的人，其邻居也应当包含有许多有影响力的人；通过类似于投票和选举的方式让社交网络上的所有人找出其认为最具有影响力的人。

PageRank 初始化每一个节点的分数都是相同的，并且所有人的分数之和为 1，也就是说

$$PR(p_i) = \frac{1}{|V(G)|}$$

之后在每一轮的迭代过程中，每一个节点都将自己现有的权重平均分配给所有的邻居，也就是他的每一个邻居都将收到自己权重的一个等分。每一个节点下一次迭代开始时的权重不仅来源于在上一轮的迭代过程中收到的来自邻居的权重，还有一小部分来源于重新分配以避免某些节点过度吸收全局所有节点的权重。

PageRank 可以使用公式表达如下，假设某一轮迭代的结果使用矩阵的形式表达为 R，

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

那么下一轮迭代之中有：

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

其中 N 代表图中的节点总数，d 代表 dangling factor，$l(p_i,p_j)$表示节点p_i分配给节点p_j的权重占p_i自身的比重。如果$(p_i,p_j) \notin E$,那么$l(p_i,p_j) = 0$，否则

$$l(p_i,p_j) = \frac{1}{deg(p_i)}$$

我们利用 PageRank 算法得到的分数为所有人进行排名，排名得到结果如下表所示。

人物	rank score
习近平	0.0136
李克强	0.0063
王毅	0.0028
小微	0.0019
何立峰	0.0016
杨洁篪	0.0014
里巴巴	0.0014
惠民生	0.0013
丁薛祥	0.0012
汪洋	0.0012
韩正	0.0011
苏宁	0.0010
王沪宁	0.0010
刘鹤	0.0010
王晔	0.0010
徐昱	0.0010
孙春兰	0.0009
李涛	0.0009
肖捷	0.0009
赵文君	0.0009

3. 社交网络的进一步挖掘分析

3.1 小世界现象

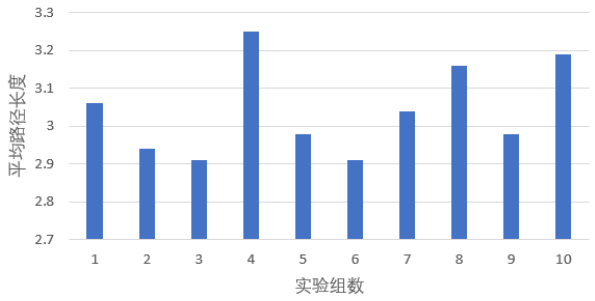
小世界现象，或者说六度分割理论，指的是现实中的社交网络存在大量和丰富的短路径。通俗地说，在社交网络中的每一个个体，除了同自身周围的同学、同事和邻里之间存在较为同质化的联系，并构成了社交网络图上大量的“三角形”之外，还存在着数量可观的远程线，也就是通过偶然的原因产生的联系，可能成为图中的捷径或者在图上搜索路径时发挥着捷径的作用。

为了验证这一现象，首先要实现的给定社交网络中两个不同的人，找出这两个人之间的前 k 条最短路径。这一问题可以被归结为 KSP 问题。虽然现有的 Dijkstra，Bellman Ford，Jonson 算法可以解决图中的单源最短路径问题或所有节点对之间的最短路径，但是这些算法的输出结果都只是最短的一条路径，而无法找到次短路径，第 k 短的路径等，有些情况下不能满足应用场景的特定需求。KSP 问题是对最短路径问题的推广。一个正确的 KSP 算法的返回结果是一个路径组成的长度为 k 的路径序列，序列中的每一个元素都是一条从源点到目标结点的路径，并且按照路径的长度进行排序。

目前比较成熟的 KSP 算法主要有以下几种：标号算法，删除路径算法，偏离路径算法，智能改进算法。而我的工作基于 Yen 等人在 1971 年提出的用于解决 KSP 问题的 Yen's algorithm，实现该算法并将其应用于新闻社交网络之中。总体而言，Yen‘s algorithm 属于偏离路径算法的范畴。现将我在社交网络上寻找两个人之间最短路径的实现思想阐述如下：

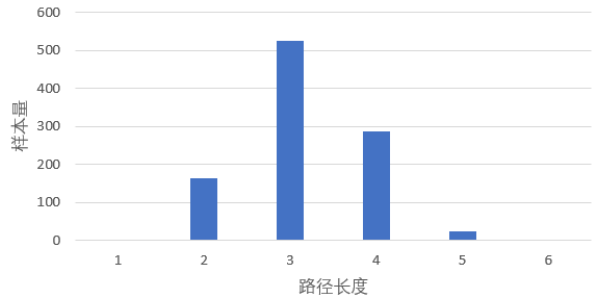
输入的起点 s 和终点 t，以及所需要寻找的路径的数量，采用 Dijkstra 算法求出从起点到终点的最短路径，将其记为p_k(k=1)。接下来，我们需要利用p_k来求得p_{k+1}。设置一个候选集合 S，初始为空，用来存放所有可能成为p_{k+1}的路径。将p_k上除了目标终点 t 之外的所有节点分别看作“偏离点”，记作$v_1, v_2, v_3, \dots, v_x$,对于每一个偏离点$v_i$，使用 Dijkstra 算法计算从$v_i$到终点 t 的最短路径，记作$p(v_i, t)$，将这条路径与$p_k$之中从起点 s 到偏离点$v_i$的路径$p(s, v_i)$进行连接，得到一条从 s 到 t 的完整路径。将这条路径加入到候选集合 S 之中。当遍历完$v_1, v_2, v_3, \dots, v_x$所有的偏离点之后，从候选集合 S 之中挑选中权重最小的路径，作为p_{k+1}，并将这条路径从候选集合 S 之中删除。重复上述流程，从p_{k+1}中得到p_2, p_3, p_4 ...,直至找到我们所需要的总共 k 条路径，程序结束并将结果输出。

要想在社交网络之中精确地检验小世界现象和六度分隔理论，最严谨的做法应当是枚举图中的每一对节点对(u,v),计算这两个节点之间的最短路径的长度，最后统计所有的最短路径的长度的平均值是否为 6.但是，估计这种方法的时间复杂度就可以发现，计算任意两个节点之间的最短路径需要的时间复杂度至少是$O(|V|log|E|)$级别的，而枚举图中的所有节点对需要的时间复杂度为$|V|^2$,因此这种验证算法总的时间复杂度将是$O(|V|^3 log|E|)$,效率非常低下，不具有可扩展性。与这个验证为了验证小世界现象，我们首先采用随机抽样的方法，每一次从社交网络所有的节点之中任意抽取 200 个节点组成 100 个节点对，使用 Dijkstra 算法计算这 100 个节点对之间的最短路径的长度，重复多次实验，记录每一次实验得到的所有路径长度和平均路径长度。几次验证的结果如下所示

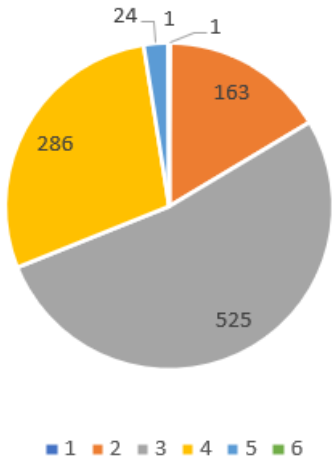


可以看到，10 次实验的平均路径长度大约都在 3 左右，很少会超过 3.2，更不用说到达六度分隔理论所提出的任意两个人在社交网络上的平均距离是 6.

接下来，我们把 1000 次随机抽样得到的点对之间的路径长度进行统计，统计得到的频度图如下所示



可以看到，绝大多数的路径长度都是 3，长度为 2 的路径和长度为 4 的路径占比稍微少一些。而长度为 1，5，6 的路径则非常少，没有长度超过 6 的路径。



饼状图也反映了类似的信息，长度为 3 的路径占比超过了所有采样路径的一半。

在新闻社交网络上进行实验得到的结果同小世界理论和六度分隔理论有一些出入。我分析认为，之所以认为出现这种情况，主要有两方面的原因。

首先是数据来源的问题。几乎所有出现在新闻报道中的名字都是热门的公众人物和政府高级官员，这意味着他们所形成的网络本身就是非常稠密的。我通过随机抽样选取到的两个人物很有可能都是国家部委的高级官员，或归属于同一行政机构，在这种情况下两个人的路径长度显然会偏低，不能同真实社交网络上的任意两个人相比。

其次是采样的问题。采样的过程中会遇到两个人属于不同的连通分量。这个时候两个人的距离实际上是正无穷，为了统计的方便，采样时遇到处于不同连通分量的节点时，直接废除此次采样，不计入采样次数并进行重新采样。

3.2 聚集系数

聚集系数，顾名思义是指图形之中所有的节点趋向于聚合在一起的程度的度量，反映了一个图中的连通程度和局部边的稠密程度。通俗地说，一个节点的聚集系数表现了这个节点的所有邻居之间，彼此也是邻居的情况所占的比例。用公式可以表示如下：

$$C_i = \frac{2\left|\{(v_j, v_k), (v_j, v_k) \in E, v_j, v_k \in N_G(i)\}\right|}{\left|N_G(i)(N_G(i)-1)\right|}$$

其中，C_i代表的是第 i 个节点的聚集系数，$N_G(i)$代表的是第 i 个节点的邻域，E 代表的是图中的所有边的集合。

之所以在公式之中会出现 2 的系数，是因为我们的社交网络图是一个无向图，无向图中不区分两个节点之间的边的方向。这一公式还可以使用更加简洁的形式表达如下：

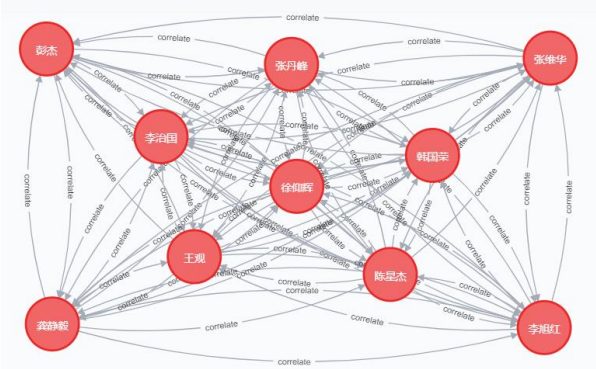
$$C_i = \frac{\lambda_G(v_i)}{\tau_G(v_i)}$$

其中$\lambda_G(v_i)$代表的是节点v_i所处的“三角形”的数量，也就是具有三个顶点和三条边，并且其中一个节点是v_i的 G 的子图的数量。而$\tau_G(v_i)$则是具有三个顶点和两条边，并且节点v_i和另外两个顶点都相邻的 G 的子图的数量。

我借助 networkx 之中相关函数计算每一个节点的聚集系数，得到的结果如下图所示：

人名	聚集系数
徐仰辉	1.0
龚静毅	1.0
陈星杰	1.0
向拜登	1.0
纳塔莉	1.0
布莱思	1.0
熊奕	1.0
唐治宇	1.0
吴德顺	1.0
谢龙星	1.0

聚集系数的统计结果显示，处于前几名的人物的聚集系数全部都是 1.0，推测可能是因为这几个人恰好组成一个小的连通分量或者是一个团。以徐仰辉为例，对其所处的局部网络结构进行绘图发现：



徐仰辉刚好处于以张维华，张丹峰，彭杰，龚静毅，李治国，陈星杰，李旭红，王观，韩国荣的 10 人小团体之中，团体之中的任意两个人都有联系，所以可以恰好达到聚集系数为 1.

3.3 中介中心性

度量一个节点在网络图中的中心性有很多方法，不同的方法侧重于使用不同的指标。常见的方法如度数中心性（Degree Centrality，DC），计算节点的邻居数目与图中所有节点总数的比值。

$$DC(v_i) = \frac{|N_G(v_i)|}{|V(G)|}$$

紧密中心性（Closeness Centrality，CC），计算的是该节点到所有节点的距离的倒数。

$$CC(v_i) = \frac{1}{\sum d(v_i, v_j)}$$

中介中心性（Betweenness Centrality，BC）,计算的是这个节点处于多少对节点的最短路径上。

$$CB(v_i) = \sum_{v_j, v_k \in V(G)} \frac{\sigma(v_j, v_k | v_i)}{\sigma(v_j, v_k)}$$

我计算新闻社交网络时使用的中心性指标是中介中心性，并对图中所有节点的中介中心性按照从高到低进行排序，排序的结果如下所示：

人名	中心性
习近平	0.4261
李克强	0.1129
小微	0.0270
王毅	0.0213
里巴巴	0.0188
惠民生	0.0143
苏宁	0.0107
赵文君	0.0088
李鑫	0.0083
齐中熙	0.0079

4. 结论

本文通过挖掘中国政府门户网站 gov.new.cn 中约 30000 条新闻中隐含的人际关系，构造社交网络，并针对这个社交网络进行一系列分析和实验。首先对图中的节点数、边数、度数、联通分量等基本信息量进行统计，以及计算每一个节点的 pagerank 影响力大小。接下来，我们继续深入分析和挖掘社交网络中的深层次信息。我设计算法找到了任意两个人之间的最短的十条路径以及任意两个人之间的平均路径长度，以此来验证小世界现象。为了度量每个节点的中心性和局部拓扑信息，我还计算了每一个节点的聚集系数和中介中心性等信息。

参 考 文 献

[1] jieba, https://github.com/fxsjy/jieba

[2] thulac, http://thulac.thunlp.org/

[3] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry(1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[4] betweenness centrality https://en.wikipedia.org/wiki/Betweenness_centrality

[5] Kosub S. A note on the triangle inequality for the Jaccard distance[J]. Pattern Recognition Letters, 2019, 120: 36-38.

[6] Buhler J. Efficient large-scale sequence comparison by locality-sensitive hashing[J]. Bioinformatics, 2001, 17(5): 419-428.

[7] Hand D J, Adams N M. Data Mining[J]. Wiley StatsRef: Statistics Reference Online, 2014: 1-7.

[8] Yen J Y. Finding the k shortest loopless paths in a network[J]. management Science, 1971, 17(11): 712-716.

[9] Brander A W, Sinclair M C. A comparative study of k-shortest path algorithms[M]//Performance Engineering of Computer and Telecommunications Systems. Springer, London, 1996: 370-379.