

基于复杂网络的大型互联网企业高管分析

郑喜亮, 苏湛, 艾均

(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘要: 采用复杂网络分析特定属性群体。以互联网企业高管简历作为原始数据, 以高管姓名和分词系统抽取实体关键词作为节点, 个人简历中是否包含关键词作为连接边的条件进行建模, 使用复杂网络理论对所建网络进行分析。实验结果表明, 部分关键词节点度值存在明显差异, 归一化后的特征向量明显大于介数。通过统计分析发现, 美国和北京相关背景很重要, 同时对比归一化后的介数和特征向量证明两点: 第一, 社会关系中个体涉及的实体对象比在社会关系网络中的位置更为重要; 第二, 跨行业跳槽人员的职业背景经历可能会给其在新企业中的个人发展带来不利影响。

关键词: 复杂网络; 分词系统; 中心性

中图分类号: TP 301 **文献标志码:** A

Analysis of Internet Enterprises' Top Executives Based on Complex Networks

ZHEN Xiliang, SU Zhan, AI Jun

(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Network theory can be used to analyze specific attribute groups. Using the executives' resumes of top Internet enterprises as data source, defining executives' name and entity keywords extracted by a word segmentation system as nodes and constructing links between the executive and keyword, when the keyword is in his resume, a network was built. Based on the network theory and the presented model, complex network characteristics were explored and analyzed. The experiment results show that the degree values of some keywords are significantly different, meanwhile normalized eigenvectors are observably larger than the normalized betweenness centrality. The statistical analysis prove that the United States and Beijing related backgrounds are extremely important for executives of top Internet enterprises, and the comparison between the normalized betweenness and eigenvector demonstrates that entities that individuals connect to in social experience are more important than the position of those individuals in social networks, the career experience of cross-industry job-hopping may not conducive to individuals' developments in a new enterprise.

Keywords: complex networks; segment system; centrality

人类对于网络的研究由来已久,18世纪伟大的数学家欧拉对于七桥问题(Konigsberg)的研究是目前公认的最早关于网络的研究。20世纪60年代由Erdos和Renyi两位匈牙利数学家建立的随机图理论(random graph theory)^[1]被公认是在数学上最早开创了复杂网络拓扑结构的系统性分析。在20世纪末,小世界网络模型^[2]和无标度网络模型^[3]的相继提出,使得复杂网络成为学术界热门的研究方向。而且小世界网络模型和无标度网络模型有着和现实世界网络^[4]相似的网络特性。例如,小世界网络的小世界性和高聚集度系数;无标度网络的小世界性和节点度分布服从幂律分布的特点。不过,这两种网络都存在一定的不足之处,如小世界网络的节点度值服从指数分布问题及无标度网络的聚集系数很小等问题,但实际上现实世界的网络也存在着不同程度的差异。针对以上网络,大量的学者进行了改进,如Newman和Watts对小世界网络模型作了进一步改进^[5],使得网络不存在孤立节点,也是目前最广为使用的小世界网络模型;Li等对无标度网络模型作了进一步改进^[6],由于网络中的节点并不能获取到网络的全局信息,从而提出了局域世界概念。

使用复杂网络来对现实世界中的一些问题进行研究,可以指导和预测现实世界中事物的发展趋势和走向,将现实世界中的有限资源合理利用,使得资源利用效益最大化。同时也可以利用复杂网络分析现实世界网络的脆弱边或是脆弱节点,从而有针对性地改善和防止问题的发生。肖盛等^[7]基于复杂网络理论对某地区真实电网进行脆弱性评估,结果表明,分析所得脆弱节点的分布情况与该地区电力网络脆弱节点的真实情况基本相符。谭阳红等^[8]对传统评估电网节点重要度的方法进行了改进,引入相对熵和灰色关联度,提高了评估结果的准确度。文献^[9]以武汉市路网作为原始数据,通过计算复杂网络的中心性,发现武汉市路网具有层次结构特征且结构构成符合“二八分率”的规律。文献^[10]分析了我国矿业并购演变过程,通过计算中心性指标,使用K-核分解算法发现矿业并购市场易受到外部环境影响等特点。这也就进一步证明了复杂网络不仅具有理论研究价值,同时也具备在现实中的应用价值。

现实世界中充斥着各种各样的信息,按照类型可分为文字、视频、图片及音频等,但最为常

见的还要属文本信息。而文本信息为了提高可读性和信息逻辑性,在表达上存在着冗余性,如文本信息中的介词、代词及标点符号等内容实际上一般并不包含有效的信息要素。为了提取出文本信息的有效信息要素,必须使用合理的分词方式来实现。分词是指将连续的字序列按照一定规范重新组合成词序列的过程。其中,分词的规范重组部分实际上是分词的核心部分,因为,无论中文还是英文都不能完全按照固定模式进行分词,文本信息中往往会含有大量专有名词、机构名、人名及地名等信息,如果按照完全固定模式分词,会使大量文本有效信息失真,所以,分词时必须进行文本语义分析,结合语义分析结果进行分词才能得到更为准确的分词结果,这样的结果对于研究也更具有现实意义。文献^[11]基于复杂网络理论对文本关键词进行提取,提出了综合考虑目标词汇以及相邻节点的贡献度来提取关键词的方法。文献^[12]通过词序统计组合来提取文本信息关键词,综合使用词序统计、词性标记、停用词过滤等方式提高关键词提取准确性。

在现代社会和企业发展中,人才是最宝贵的资源,很多企业和国家都在强调这一点,甚至出现企业为了得到一个关键人物或团队出手将整家公司收购的极端现象,由此看出现代社会对于有才能人士的重视。同时,人才也是不同国家、不同企业的必争之地。我国提出的“千人计划”、“万人计划”以及现在各个城市都在不断推出的引进人才计划和为留住人才而推出的各种优惠政策都从侧面证明了人才对于社会和企业发展的重要性。文献^[13]通过趋势外推法和灰色模型对湖北科技人才需求进行演化预测,文献^[14]分析了人才分析对战略人力资源管理的价值,总结了人才分析的主要用途并提出了实施人才分析需要注意的相关问题。本文基于复杂网络和NLPIR(natural language processing & information retrieval)分词系统以百度百科词条信息作为原始数据对几家互联网企业的高管进行研究分析。

1 高管复杂网络模型建立

本文使用的高管人员名单来自于以下互联网企业:百度、阿里巴巴、腾讯、滴滴、美团、京东、今日头条;每位高管人员的个人信息均来自于百度百科官方词条。

建模算法设计:

a. 由于使用爬虫技术获取到的文本信息不规范, 存在格式不统一、包含异常的符号等问题, 所以, 在获取到文本信息后, 根据预先设定的处理规则进行处理。例如, 删除多余的空格、换行符号及少部分的不规则网页标签等。经过预处理后原始数据成为姓名与个人百度百科词条一一对应的形式。例如, 通过预处理得到以李彦宏为文件名的文本文件。

b. NLPPIR(<https://github.com/NLPPIR-team/NLPPIR>) 是一套能够将给定文本进行分词并提取实体词(有意义的词汇, 如人名、地名、公司名称等)、关键词、敏感词的分词系统。本文使用该分词系统对处理好的文本信息进行实体抽取, 获得有意义的实体词汇作为关键词信息 k 。例如, 李彦宏对应百度、创始人、董事长、北京大学、华尔街等实体关键词。

c. 通过以上 a, b 这 2 个步骤, 获得了两类网络节点, 一类节点为个人姓名 v_p , v_p 代表本文网络所使用的高管人员姓名; 另一类节点为关键词 v_k , v_k 代表分词系统实体抽取得到的有意义的实体描述关键词。

最终得到节点集

$$V = \{V_p, V_k\} \tag{1}$$

经过简化处理后, 人物分析网络可以被抽象为一张具有 N 个节点和 M 条边的无向网络。设网络为 $G=(V, E)$ 。其中, G 表示有 N 个节点和 M 条边组成的网络, V 表示网络中的节点, 即 $|V|=N$; E 代表网络中的边, 即 $|E|=M$ 。 V 即是上文中通过数据处理获得的 V_p 和 V_k 的集合。而边则是 V_p 和 V_k 连接形成的, 两者建立边的原则: 若是 p_i 的个人信息描述中包含关键词 k_j , 则为 V_{p_i} 和 V_{k_j} 建立一条边; 否则两者不建立边, 即

$$(V_{p_i}, V_{k_j}) = \begin{cases} \text{有,} & p_i \text{ 个人描述中包含关键词 } k_j, \\ \text{无,} & p_i \text{ 个人描述中不包含关键词 } k_j, \end{cases}$$
$$i = 1, 2, \dots, p, j = 1, 2, \dots, k \tag{2}$$

边集合为

$$E_{ij} = (V_{p_i}, V_{k_j}), i = 1, 2, \dots, p, j = 1, 2, \dots, k \tag{3}$$

图 1 为本文建模算法设计流程图。按照上述规则建立复杂网络, 如图 2 所示。表 1 为网络的基本参数。图 3 为节点度分布图。其中, 横坐标为节点数, 纵坐标为节点的度。

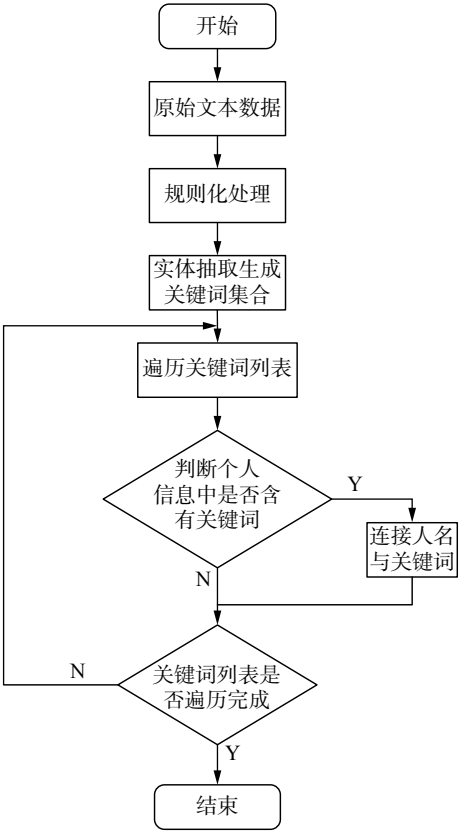


图 1 算法流程图
Fig. 1 Algorithm flow chart

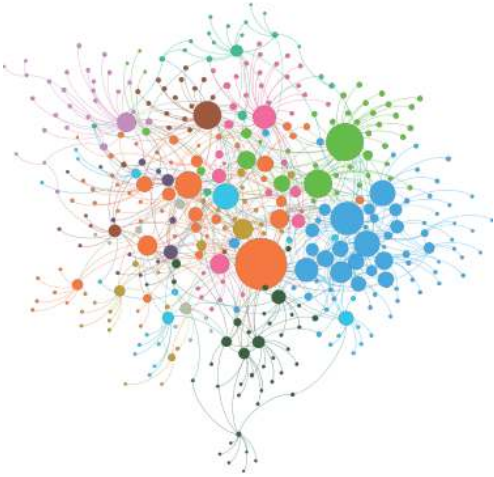


图 2 建模得到的复杂网络
Fig. 2 Complex network

通过图 3 和表 1 可以看出, 根据上述模型所建复杂网络有着与现实世界网络相似的网络参数^[15]。例如, 网络的平均路径长度为 3.641, 满足现实世界网络的小世界特性^[16], 而聚集系数为 0.092, 相对较大, 满足高聚集性, 图 3 为网络的度分布情况, 可以看出, 网络的度分布与现实世界网络度分布的无标度特点相对应。

表 1 网络各项参数指标
Tab.1 Network features

网络参数	节点数	边数	平均度	模块度	聚集系数	平均路径长度	网络密度	网络直径
值	382	671	3.513	0.553	0.092	3.641	0.009	6

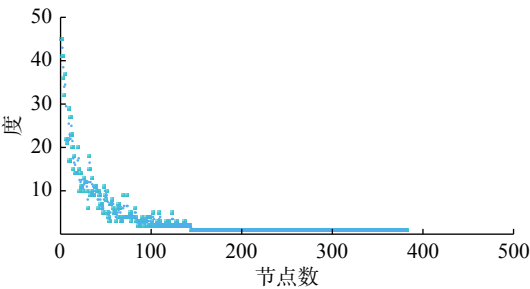


图 3 网络度值分布图
Fig.3 Distribution of degree

2 高管网络中心性

描述复杂网络的参数主要包含以下几种：网络密度、度、平均最短路径和网络的聚集系数。其中，网络密度主要刻画了网络连接的总体特征，网络的度主要描述个体节点的连接情况，平均最短路径能够体现网络信息传递效率，而网络的聚集系数则能够体现网络的局部聚集情况。

2.1 度

节点的度值为与该节点直接相连的节点数。可以看出，网络中一个节点的度在 $[0, N-1]$ 区间内， N 为网络 G 的节点数。节点度的大小反映了该节点在网络中的直接影响力^[17]。设 A 为网络 G 的邻接矩阵， a_{ij} 为 A 的第 i 行第 j 列元素，则

$$D_{x_i} = \sum_{j=1}^{N-1} a_{ij}$$

(4)

式中， D_{x_i} 为节点 x_i 的度值。

图 4 是经过处理后关键词度数排名前 25 的关键词度数分布情况。其中，横坐标为关键词节点，纵坐标为节点的度。由于网络使用的原始数据涉及的企业及人员均在中国，所以，度最大的节点为中国。同理，由于分析对象为高管人群，CEO 和总经理的高频出现也属于合理现象，本文对此不作进一步分析。但度数排名分别为第 4 和第 5 的北京和美国与同类关键词相比度数差别较大。故本文分别对北京和美国进行横向对比分析。

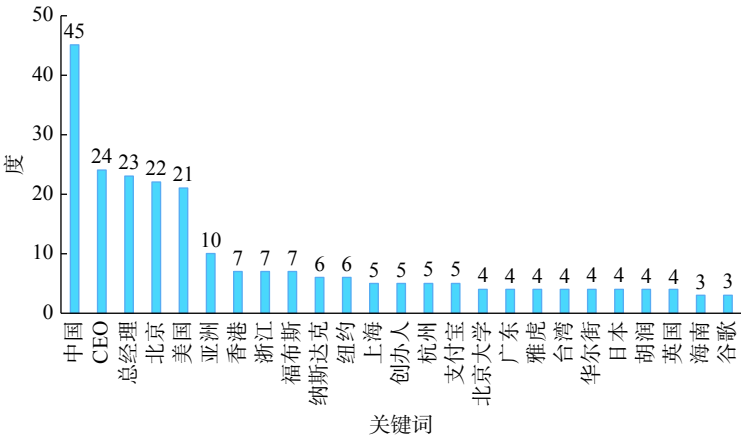


图 4 关键词度值
Fig.4 Degree of keyword

分别将度值排名在前 25 的关键词及关键词对应的度值进行横向对比分析，可以得到柱状图如图 5 所示。

从图 5 可以看出，横向对比下北京和美国这 2 个关键词的度值相比于同类关键词具有明显的优势，两者的度值都远大于同类关键词的

度值，其中，美国是同类关键词的 5~7 倍，北京是同类关键词的 3~7 倍。通过度值的横向对比可以看出，这些互联网企业的高管和北京、美国相关的背景信息显得尤为重要。这一点也与现实世界的基本情况相符。美国作为自二战以来的世界强国，对于计算机、互联网这类高

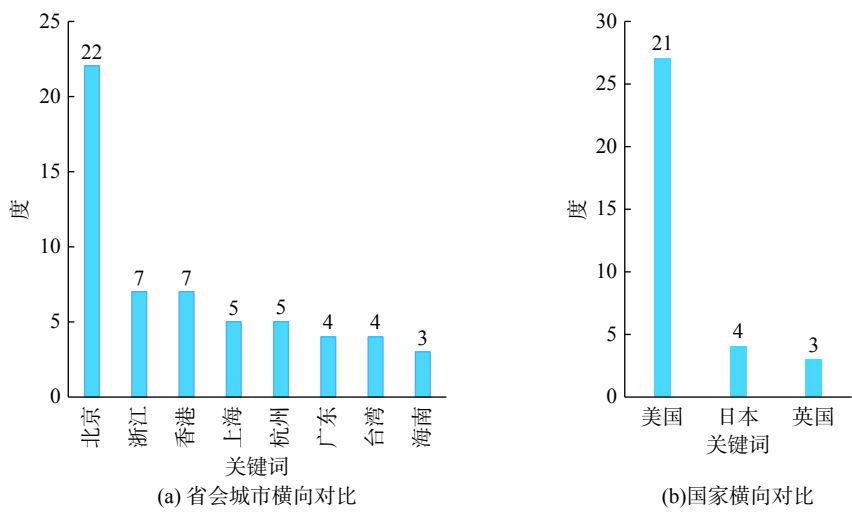


图 5 度值横向对比
Fig.5 Degree comparison

科技行业在全球范围内都称得上绝对领先，所以，我国大量互联网人才在选择留学、交流时，美国往往是他们的首选，文献[18]中表明，美国仍然是多数人留学的首选目的地。另一方面，虽然我国大型互联网公司在国内发展情况很好，但往往也会在美国硅谷等地设立研发中心等，网络中涉及的 7 家(百度、阿里、腾讯、滴滴、美团、京东、今日头条)公司中有 71.43% 的公司在美国有分公司、办事处或研发中心。文献[19]指出，设立海外研发中心具有从东道国获取新信息、新技术服务于母公司的技术开发战略的作用，而这时的东道国主要有两类：一是拥有“技术高地”的技术发达国家，二是拥有“人才富地”的人力资源优质国家，而恰好美国在这两方面都符合。

北京作为我们国家的首都及政治中心，对于任何一个企业而言，北京相对于我国的其他城市有着完全不同的意义，企业管理人员背景信息和北京有着较大的关系也属于合理。另一方面，北京作为一线城市，有着众多优秀的高校和人才，企业和一线城市有着较大的亲密度，这对于属于前沿行业的互联网公司也是必不可少的，网络中涉及的 7 家企业 100% 均在北京有分公司、研发中心等。其中，5 家公司总部就在北京。文献[20]针对我国互联企业分布进行分析。图 6 为互联网企业数量前五的省市的统计分析图。可以看出，北京互联网企业数量有明显优势，这也进一步证明了本文所得分析结果的合理性。

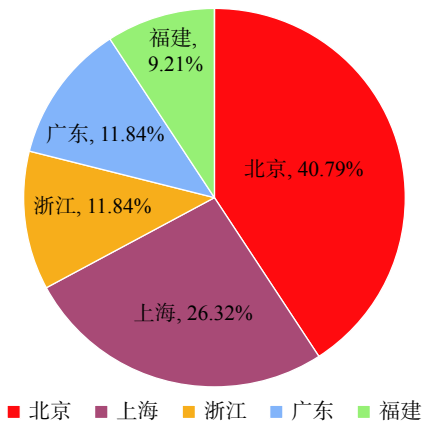


图 6 互联网企业省市分布占比图
Fig. 6 Internet enterprise ratio of provinces

2.2 特征向量和介数

2.2.1 特征向量

一个节点的度指标仅仅描述了该节点对于其他节点的直接影响力，若一个节点与另一个度值很高的节点之间存在连接^[17]，则该节点的影响力也会受到影响而增强，这与现实世界中的很多情况是一致的，如果把每个人看作是一个节点的话，一个节点的信息占有率在很大程度上取决于它与什么样的节点有连接。但是，这种类型的特征无法用度值进行衡量，故本文使用特征向量这一网络指标来衡量节点这种特征的显著与否。设网络 G 的邻接矩阵为 A ， a_{ij} 为矩阵第 i 行第 j 列的元素， a_{ij} 的取值集合为 $\{0, 1\}$ ， λ 为 A 的主特征值，是一个常量， $e = (e_1, e_2, \dots, e_N)$ 为矩阵 A 对应 λ 的特征向量，即

$$\lambda e_i = \sum_{j=1}^N a_{ij} e_j, \quad i = 1, 2, \dots, N \tag{5}$$

即

$$e_i = \lambda^{-1} \sum_{j=1}^N a_{ij} e_j, \quad i = 1, 2, \dots, N \tag{6}$$

式(6)即为特征向量的计算公式。

2.2.2 介数

介数指标主要体现了网络中的节点对于信息流动的影响力。设网络 G 中含有 N 个节点，则节点 x 的介数指标 B_x 定义为

$$B_x = \frac{\sum_{j < k} g_{jk}(x)}{\sum_{j < k} g_{jk}} \tag{7}$$

式中： g_{jk} 表示网络中节点 j 和节点 k 之间的所有最短路径数； $g_{jk}(x)$ 为经过节点 x 的最短路径的数量。

图 7 通过可视化的图形来对比同一节点在图 7(a)和 7(b)中的相对大小。按照节点的度大于等于 4 对网络节点进行过滤，得到 85 个节点(过滤前为全部节点，未进行过和类型相关的区分)。以节点的介数和特征向量相对大小决定图中的节点半径作为自变量，大小作为应变量作图；由于介数和特征向量的计算公式不同，为保证公平性，以归一化后的特征值作为决定节点半径大小的唯一自变量，节点的半径与对应特征值归一化后的值为线性关系。按照以上规则得到图 7。图 7(a)为按照节点的介数作图，图 7(b)为按照节点的特征向量作图。

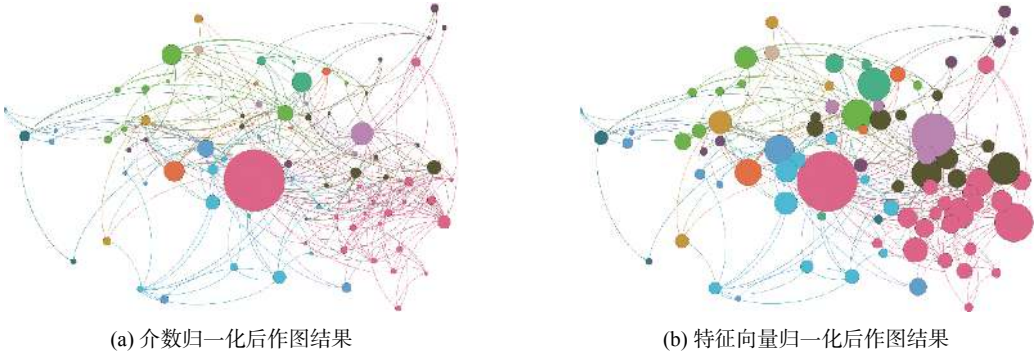


图 7 介数与特征向量归一化可视化对比

Fig.7 Visual comparison between the normalized eigenvector and betweenness

图 8 为曲线图，横坐标为网络 G 中的所有的高管节点，纵坐标为网络指标归一化后的常数。从图 8 中可以看出，绝大多数高管节点的特征向量较介数有明显的优势，部分节点的特征向量

值远大于介数值。图 8 中为部分典型节点加入了数据标签，可以看出，李彦宏、马云、张亚勤、马化腾、戴珊等人的特征向量明显比对应的介数高。

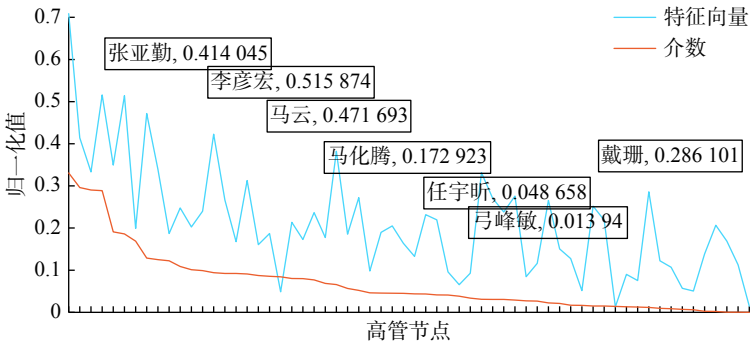


图 8 特征向量与介数归一化对比

Fig.8 Comparison between the normalized eigenvector and betweenness

在不考虑极少的特殊节点后，通过对比图 7 和图 8 发现，高管节点的特征向量明显大于他们的介

数，即特征向量这项指标更为显著。考虑到介数主要用来评价节点对网络中信息传播的影响力^[21]，

特征向量用来评价节点的相邻节点的度值大小, 而节点的度值又能判断出该节点对于网络的直接影响力, 从而可以看出, 相比于个体传递信息的能力而言, 认识什么样的人对于个人的成功与否显得更重要, 这与现实世界中的社交关系有着明显的一致性。如果一个人在社会交际中认识的人都是比较有影响力的, 那么, 往往这个人也比较有影响力, 而且比起那些不具备这一特点的人而言也更容易成功。

2.3 异常节点分析

从图 8 中发现, 极少部分的异常节点的归一化介数反而大于归一化后的特征向量的值。通过对原始网络进行分析, 发现任宇昕、弓峰敏相对于其他节点而言, 关键词信息和个人经历与这些互联网公司中的其他人员确实有较大的差别。

2.3.1 模块度分析

图 9 中, 在模块化后, 任宇昕独立划分为一块。造成这一结果的原因是本文研究对象限制在互联网公司高管人员范围之内, 而企业高管会有不同类型企业之间跳槽情况。例如, 弓峰敏曾就职于网络安全公司, 而图 9 中任宇昕曾就职于华为公司, 华为公司是一家做电信服务和手机业务的企业, 不属于互联网类公司。这就造成了这种类型的人的个人信息与其他人员存在较大的差异。



图 9 异常节点局部网络
Fig.9 Local network of abnormal nodes

2.3.2 接近中心性分析

接近中心性是通过计算节点与网络中其他所有节点的距离的平均值来消除特殊值的干扰。一个节点与网络中其他节点的平均距离越小, 该节点的接近中心性就越大^[22]。通过计算网络节点的接近中心性, 分别得到任宇昕和弓峰敏的接近中心性为 0.544 643 和 0.458 647, 在高管节点中分别位于倒数第 3 位和最后 1 位。接近中心性倒数第 2 位的节点为姚星, 通过查看原始数据发现, 造成该节点接近中心性、特征向量和介数值较低的原

因是从百度百科抓取到的词条内容非常少, 仅有一句话。不同于姚星, 任宇昕和弓峰敏的词条内容非常丰富, 而词条内容量类似的曾鸣、董本洪等节点的接近中心性非常大, 曾鸣和董本洪接近中心性分别排在第 11 和第 20 位, 即造成任宇昕和弓峰敏的接近中心性较低的原因可能是跨行业跳槽造成的。

综上所述可以发现, 跨行业跳槽的人员与一直从事于该行业的人员在多项中心性值的比较中存在较为明显的差异, 这与其长期的职业背景经历有较大的关系, 这种差异可能为未来的职业发展带来不利影响。

3 结 论

以复杂网络为理论基础, 对几家大型互联网企业高管人员进行建模分析, 通过复杂网络的理论分析研究对象是否具有共性、具有哪些共性以及该类人员具有哪些显著特点。以百度百科官方词条作为原始数据, 通过分词系统抽取实体关键词, 获得有信息要素的实体关键词; 以高管姓名和关键词作为网络节点, 高管个人信息中是否包含关键词作为连接边的条件构建复杂网络; 最后通过复杂网络的理论分析得到一些与现实世界具有一致性的网络特征, 同时这些一致性也证明模型的合理性。

分析网络后, 得到以下主要结论: 首先, 将度值较高的美国和北京分别整理统计对比后发现, 美国和北京相对于其他同类别的关键词确实有着明显的度值优势, 证明对于大型互联网企业高管而言美国和北京相关的背景信息确实很重要; 其次, 将特征向量指标值和介数指标值归一化后作图对比, 发现一个与现实世界具有一致性的特点, 一个人社交关系中的社交对象相比于他在社交关系中起到的信息传递的作用更为重要, 也让他更容易获得成功; 最后, 对图 9 中的异常节点进行分析, 发现跨行业跳槽时过去不同的职业经历可能会在未来的个人发展中产生不好的影响。

参考文献:

[1] BOLLOBÁS B. The evolution of random graphs[J]. Transactions of the American Mathematical Society, 1984,

- 286(1): 257–274.
- [2] WATTS D J, STROGATZ S H. Collective dynamics of ‘small-world’ networks[J]. *Nature*, 1998, 393(6684): 440–442.
- [3] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509–512.
- [4] ALBERT R, BARABÁSI A L. Statistical mechanics of complex networks[J]. *Reviews of Modern Physics*, 2002, 74(1): 47–97.
- [5] NEWMAN M E J, WATTS D J. Renormalization group analysis of the small-world network model[J]. *Physics Letters A*, 1999, 263(4/6): 341–346.
- [6] LI X, CHEN G R. A local-world evolving network model[J]. *Physica A: Statistical Mechanics and its Applications*, 2003, 328(1/2): 274–286.
- [7] 肖盛, 张建华, 肖河. 基于复杂网络理论与风险理论的地区电网脆弱性评估[J]. *电网与清洁能源*, 2013, 29(6): 21–28.
- [8] 谭阳红, 张婧, 李肖. 基于复杂网络理论的电网节点重要度评估[J]. *计算机工程*, doi: [10.19678/j.issn.1000-3428.0053015](https://doi.org/10.19678/j.issn.1000-3428.0053015).
- [9] 王庆国, 张昆仑. 复杂网络理论的武汉市路网结构特征[J]. *测绘科学*, 2019, 44(4): 66–71.
- [10] 武羨慧, 郭琛. 基于复杂网络的我国矿业并购演变过程分析[J]. *中国矿业*, 2018, 27(12): 58–64.
- [11] 刘通. 基于复杂网络的文本关键词提取算法研究[J]. *计算机应用研究*, 2016, 33(2): 365–369.
- [12] 苏祥坤, 吾守尔·斯拉木, 买买提依明·哈斯木. 基于词序统计组合的中文文本关键词提取技术[J]. *计算机工程与设计*, 2015, 36(6): 1647–1651.
- [13] 毛磊, 李成标. 湖北省科技人才分析与预测[J]. *科技信息*, 2007(23): 1–5.
- [14] 朱伟峰. 人才分析: 一种日趋流行的战略人力资源管理技术[J]. *中国商贸*, 2010(25): 61–62.
- [15] NEWMAN M E J. The structure and function of complex networks[J]. *SIAM Review*, 2003, 45(2): 167–256.
- [16] 刘涛, 陈忠, 陈晓荣. 复杂网络理论及其应用研究概述[J]. *系统工程*, 2005, 23(6): 1–7.
- [17] 王林, 张婧婧. 复杂网络的中心化[J]. *复杂系统与复杂性科学*, 2006, 3(1): 13–20.
- [18] 刘萍, 袁昕. 中学生出国留学的意向调查与思考[J]. *科教导刊*, 2018(13): 182–184.
- [19] 柯银斌, 康荣平, 沈泱. 中国企业海外研发的功能与定位[J]. *经济界*, 2011(6): 22–31.
- [20] 王亮. 我国互联网企业的区域分布情况分析——基于不同省市分布的差异研究[J]. *科技·经济·市场*, 2018(3): 194–196.
- [21] RAJASINGH I, RAJAN B, ISIDO D F. Betweenness-centrality of grid networks[C]//Proceedings of 2009 International Conference on Computer Technology and Development. Kota Kinabalu, Malaysia: IEEE, 2009: 407–410.
- [22] 任晓龙, 吕琳媛. 网络重要节点排序方法综述[J]. *科学通报*, 2014, 59(13): 1175–1197.

(编辑: 石 瑛)