基于豆瓣电影演员合作网络的社区发现

一、研究目的

本文通过利用豆瓣电影演员合作数据，从原始数据提取适合做社区发现的连通子网，作为本文主要研究的网络，先使用网络描述性统计指标，如从度分布、节点中心性等角度对本文研究的网络进行描述统计分析，再应用多种社区发现算法，如 fastgreedy、multilevel 等，通过模块度来评价社区发现算法对本文主要研究网络社区划分效果，从而检测每种算法的适用情况。

二、豆瓣电影演员合作数据的来源及说明

本文所使用的数据为在豆瓣爬取的 2015 年以来，评分在 7.5 分以上的电影的演员合作网络数据。原始数据为邻接矩阵，示例表 1 所示：

该邻接矩阵是对称矩阵，因此该网络为无向网络。矩阵对角线元素为演员在爬取的数据中的作品数，非对角线为演员之间的合作电影数，如（2,2）元素为 5，表明广末凉子在数据中的电影数为 5 部，（2,1）元素为 1，即她与本木雅弘共同参演了 1 部电影。元素为“.”则表示 0。

该网络共有 7025 个节点，14790 条边。即有 7025 个演员，14790 条合作关系，合作关系中包括一个演员与自己的合作关系（即电影作品数），和演员之间可能不止一次的合作关系。因此该网络图存在自环和多重边，为多重图。

表 1 邻接矩阵示例

	本木雅弘	广末凉子	山崎努	克里斯蒂安·贝尔	休·杰克曼
本木雅弘	1	1	1	.	.
广末凉子	1	5	1	.	.
山崎努	1	1	3	.	.
克里斯蒂安·贝尔	.	.	.	4	1
休·杰克曼	.	.	.	1	5

三、豆瓣电影演员合作数据数据预处理

（一）改多重图网络为简单图网络

考虑到社区发现算法多适用于简单图网络，因此需把原始数据形成的网络进行预处理，包括去除自环和把多重边转化成边的权重，得到包含 7025 个节点和 7765 条边的无向有权网络。

（二）提取合作演员数目大于 2 次的演员及其合作演员

原始数据中包含许多仅和 1 个演员有合作关系的演员，由此形成的网络中，节点分散，社区并不明显，且如果直接使用原始数据，应用社区发现算法时会因为数据量大使得程序运行速度非常慢，因此提取合作演员数目大于 2 次的演员及其合作演员。由此提取的子网络节点数为 919，边数为 1179，极大地缩小了数据量。

（三）取最大连通组件进行分析

图 1 为根据缩小后的数据绘制网络图，使用的布局为 kamada—kawai。每一个节点为一个演员，每一条边表明演员之间存在合作关系，不同的连通组件赋予不同的颜色。可以看到，中间的红色点构成最大连通组件，十分密集，通过肉眼观察大致能分为 3 大类，可能是合作关系十分紧密的演员；而周围散布一圈的多色点为规模很小的多个连通组件，这是爬取的数据中包含的少量的只在几个演员之间彼此合作的电影。

图 2 展示了图 1 中的细节，可以发现，中间一圈的红色点可粗略看成由 3 大群演员构成，其中两个规模较小些的群（Ⅰ区、Ⅱ区）均为日本演员，通过查阅资料可以发现，Ⅰ区均为日本电影演员，而Ⅱ区多为日本配音演员，另外一个规模较大的群（Ⅲ区）为欧美的演员，散布在周围一圈的极小群为多个国家的演员合作关系，如中国、韩国、印度等。

由于散布在周围一圈的极小群可能干扰社区发现的效果，因此决定以由红色点构成的最大连通组件为本文后续主要研究的网络。

四、最大连通组件的描述统计

最大连通组件包含 588 个节点和 1033 条边，以边的权重表示边的粗细，绘制网络图，从图 3 可以看到，权重较大的边集中在Ⅱ区，还有少量分布在Ⅲ区，

说明日本配音演员合作更为紧密。另外，Ⅰ区的节点最为密集，而Ⅲ区的节点最为稀疏但数量多，说明欧美演员圈子大，合作广泛。

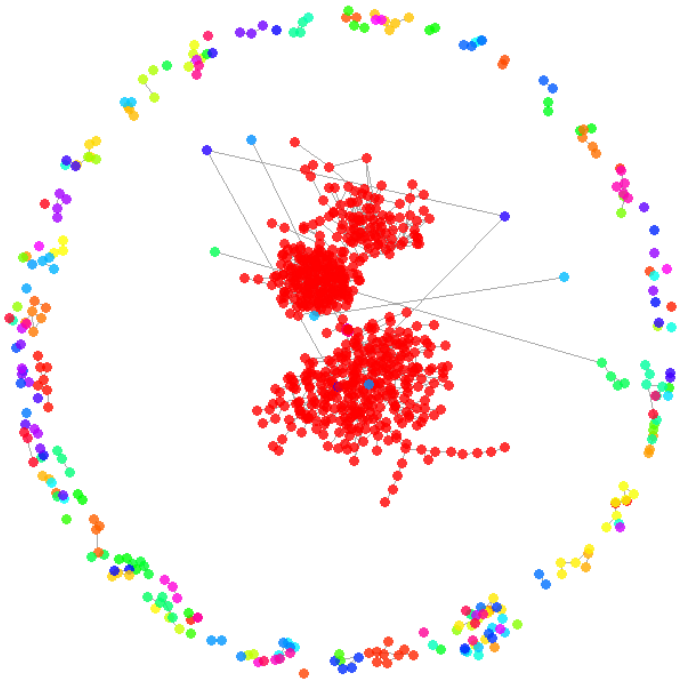

图 1 合作演员数在 2 人以上的演员合作网络图

（不同颜色为不同的连通组件）

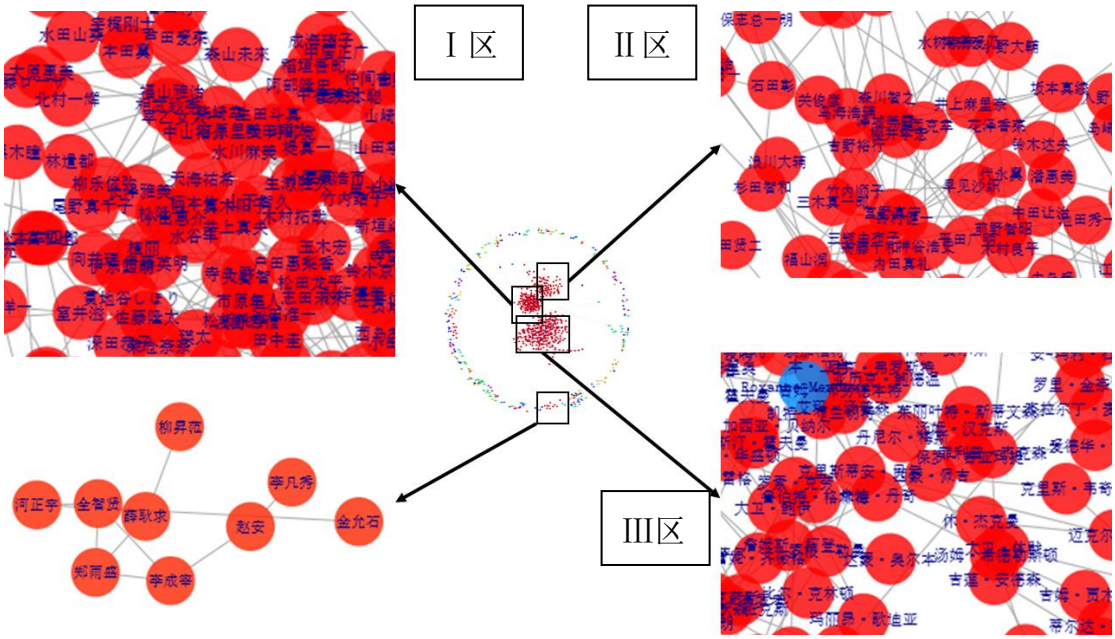

图 2 合作演员数在 2 人以上的演员合作网络图的细节展示

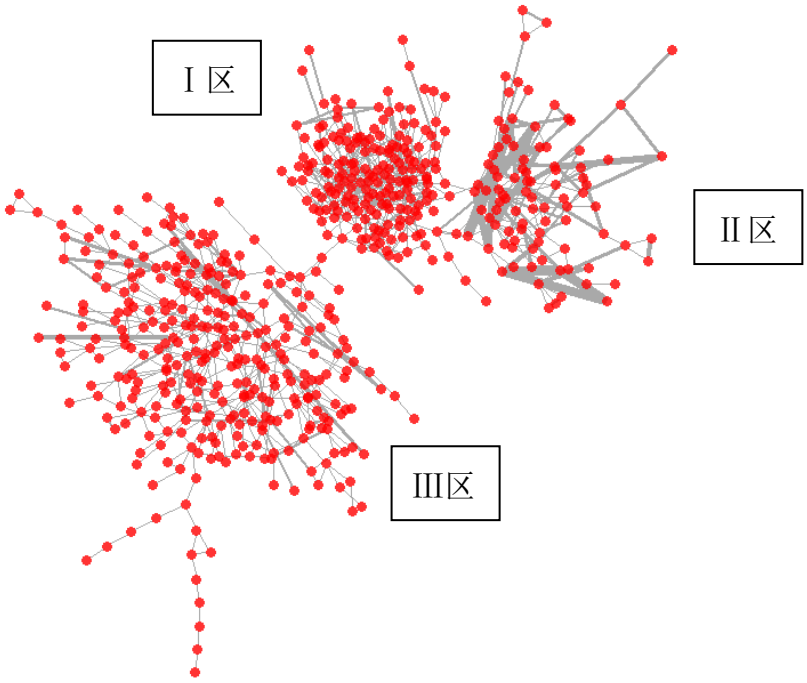

图 3 最大连通组件网络图

（一）度分布

对于含权网络，度的一个有用的推广就是节点的强度，即与某个节点相连的边的权重之和，强度分布有时也称为加权度分布。最大连通组件的强度分布如图所示。从图 4 可以看到，该网络大部分节点的度都很小，但也有一小部分节点具有很大的度，没有一个特征标度，强度分布大致服从幂律分布，非常符合现实网络的无标度特性。

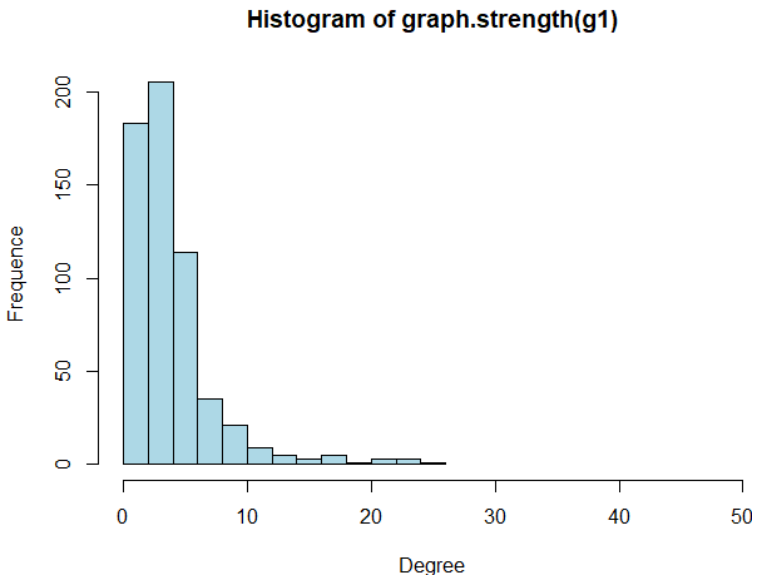

图 4 最大连通组件的度分布

（二）节点中心性

节点中心性定义了网络中一个结点的重要性。 节点中心性度量有：节点度中心性、接近中心性、介数中心性以及特征向量中心性。

其中，使用最为广泛的是节点度中心性，即与点相连的边的数量。

接近中心性的思想是，如果一个节点与许多其他节点都很”接近“，那么节点处于网络中心位置，定义为某节点到其他所有节点距离之和的倒数。

介数中心性指的是一个结点担任其它两个结点之间最短路的桥梁的次数。一个结点充当“中介”的次数越高，它的中介中心度就越大。

特征向量中心性认为，拥有很多的邻居的节点并不能确保这个节点就是重要的，拥有更多重要的邻居才能提供更有力的信息。

从表 2 可以看出，每种指标选出的中心性最大的前 10 名演员不尽相同。图 5 中，节点面积与节点中心性成正比，圆圈越大，说明节点越中心性越好。从图 5 可以看出，度中心性选出的节点中心性较大的节点多位于Ⅰ区，因为Ⅰ区在 3 个区中最为稠密，且表 2 显示度中心性选出的前 10 名全为日本电影演员；接近中心性选出的节点中心性较大的节点多位于Ⅰ区和与Ⅰ区连接的Ⅲ区的部分，因为只有在网络较为中心的位置的节点到其他点的总距离才会比较小；介数中心性选出的节点中心性较大的节点是连接Ⅰ区与Ⅲ区的节点，因为它们充当了连接两个区域的桥梁；特征向量中心性选出的节点中心性较大的节点在Ⅱ区，因为从图 3 可以看出，权较大的边多位于Ⅱ区，因此Ⅱ区有更多“重要”的节点，且表 2 显示特征向量中心性选出的前 10 名全为日本配音演员。

表 2 不同的节点中心性指标的选出的节点中心性最大的前 10 名演员

排名	度中心性	接近中心性	介数中心性	特征向量中心性
1	加濑亮	哈维尔·巴登	哈维尔·巴登	日笠阳子
2	户田惠梨香	乔什·布洛林	北野武	浅沼晋太郎
3	竹内结子	北野武	乔什·布洛林	佐藤聪美
4	二宫和也	二宫和也	二宫和也	坂本真绫
5	中村悠一	马特·达蒙	马特·达蒙	吉野裕行
6	堺雅人	丽贝卡·豪尔	大卫·田纳特	丰崎爱生
7	小栗旬	斯嘉丽·约翰逊	胜地凉	井口裕香
8	阿部宽	加濑亮	丽贝卡·豪尔	铃村健一
9	大卫·田纳特	汤米·李·琼斯	伍迪·哈里森	阪口大助
10	苍井优	西田敏行	丰崎爱生	中村悠一

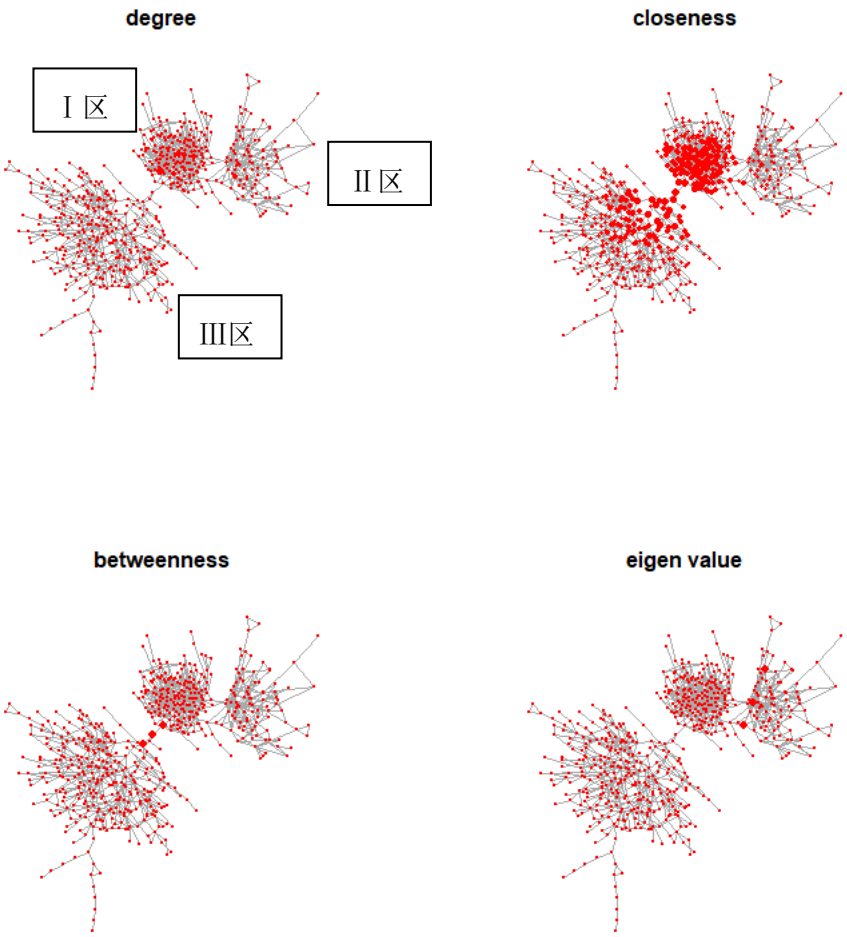

图 5 不同节点中心性指标的衡量的网络图

（节点大小为由指标数值大小决定）

（三）边介数

边介数定义为网络中所有最短路径中经过该边的路径的数目占最短路径总数的比例。图 6 中黑色的粗线为边介数最大的前 10 名的边。可以看到，黑色的边为连接Ⅰ、Ⅱ、Ⅲ区的桥梁，即黑色边连接的演员跟它周围两个区的演员均有合作。

表 3 边介数前 10 的边

排名	边名	
1	乔什·布洛林	--北野武
2	哈维尔·巴登	--乔什·布洛林
3	北野武	--二宫和也
4	马特·达蒙	--哈维尔·巴登
5	丽贝卡·豪尔	--大卫·田纳特
6	哈维尔·巴登	--丽贝卡·豪尔

7	丰崎爱生	--胜地凉
8	罗莎里奥·道森	--伍迪·哈里森
9	马特·达蒙	--罗莎里奥·道森
10	加濑亮	--北野武

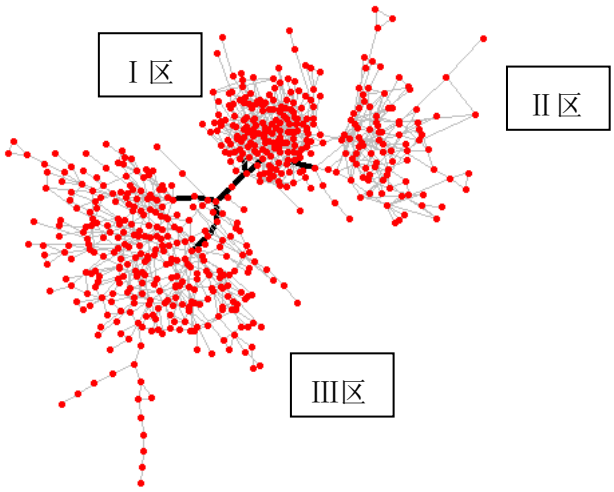

图 6 边介数大的边在网络中的位置

（黑色的粗线为边介数最大的前 10 名的边）

（四）凝聚性特征

团是一类完全子图，集合内所有节点都由边相互连接，因而是完全凝聚的节点子集。对网络进行所有尺寸的团的普查，可以快速了解图的结构。极大团是不被任何更大的团包含的一类团，在本文的最大连通组件的网络中，尺寸为 2 的极大团有 298 个，尺寸为三的极大团有 259 个，但尺寸为 4 的团仅有 2 个，分别是“真木阳子 福山雅治 堤真一 柴崎幸” 和“竹达彩奈 井口裕香 阿澄佳奈 仪武祐子”。说明这个网络符合现实网络的稀疏性。

另外，福山雅治、堤真一、柴崎幸一起合作了电影《嫌疑人 X 的献身 容疑者 X の献身》，真木阳子分别与另外 3 人合作过。竹达彩奈、井口裕香、阿澄佳奈、仪武祐子共同为动漫《玉响》配音过。

表 4 所有尺寸的极大团的普查结果

团的尺寸	2	3	4
包含节点数	298	259	2

网络的密度为实际出现的边和可能的边的频数之比，形容网络的结构复杂程度，越大说明网络越复杂，说明网络越能聚类块；聚类系数是对全局聚集性的度量，定义为连通三元组闭合形成三角形的相对频率，可以衡量网络中关联性如何，

值越大代表交互关系越大，说明网络越复杂，越能聚类。从表 5 可以看到，该网络密度仅为 0.006，平均度仅为 3.51，聚类系数为 0.20，因此该网络的总体聚集性并不高。

该网络的平均路径长度为 9.23，最长路径为 32，并不是很符合“小世界性”，网络聚集度与连通度不高。

表 5 密度、连通性指标

指标	密度	平均度	聚类系数	平均路径长度	直径
数值	0.0060	3.5136	0.1999	9.2386	32

五、社区发现算法效果评价

社区发现的算法比较多，大致可以分为两大类：拓扑分析和流分析。前者一般适用于无向无权网络，思路是社区内部的连边密度要高于社区间。后者一般适用于有向有权网络，思路是发现在网络的某种流动（物质、能量、信息）中形成的社区结构。这两种分析各有特点，具体应用取决于网络数据本身描述的对象和研究者想要获得的信息。8 种主流的社区发现算法及说明如表 6 所示，其中 Leading Eigenvector、Info map、Role-based community 三种方法不适用于本文研究的无向有权网络，因此只使用剩下的 5 种算法。

表 6 社区发现算法使用说明

	社区发现算法	优化目标	适用情况	局限
拓扑分析	Edge-Betweenness	最小化社区间连边的 betweenness	无向有权多分量	慢
	Leading Eigenvector	对拉普拉斯矩阵第二小特征根对应的特征向量聚类	无向无权多分量	
	Fast Greedy	使用社区合并算法来快速搜索最大 Q-modularity	无向有权多分量	不适用于小网络
	Multi Level	使用社区展开算法来快速搜索最大 Q-modularity	无向有权多分量	不适用于小网络
流分析	Walk Trap	最大化社区间的流距离	无向有权单分量	
	Label Propagation	每个节点取邻居中最流行的标签，迭代式收敛	无向有权单分量	不保证收敛，结果不稳定
	Info map	最小化随机流的编码长度	有向有权单分量	
	Role-based community	划分出在流中地位类似的节点	有向有权单分量	

从图 7 各种社区发现的算法效果图可以看出，每个算法划分的情况几乎完全

不同，且倾向于划分出非常多个小社区，如随机游走算法划分了 50 个社区，label.propagation 算法划分了 90 个社区，考虑到展示美观性，不在文中展示。但几乎都是把稀疏的III区划分为许多个小社区，划分的社区中几乎都是在三大区内划分，没有跨区。

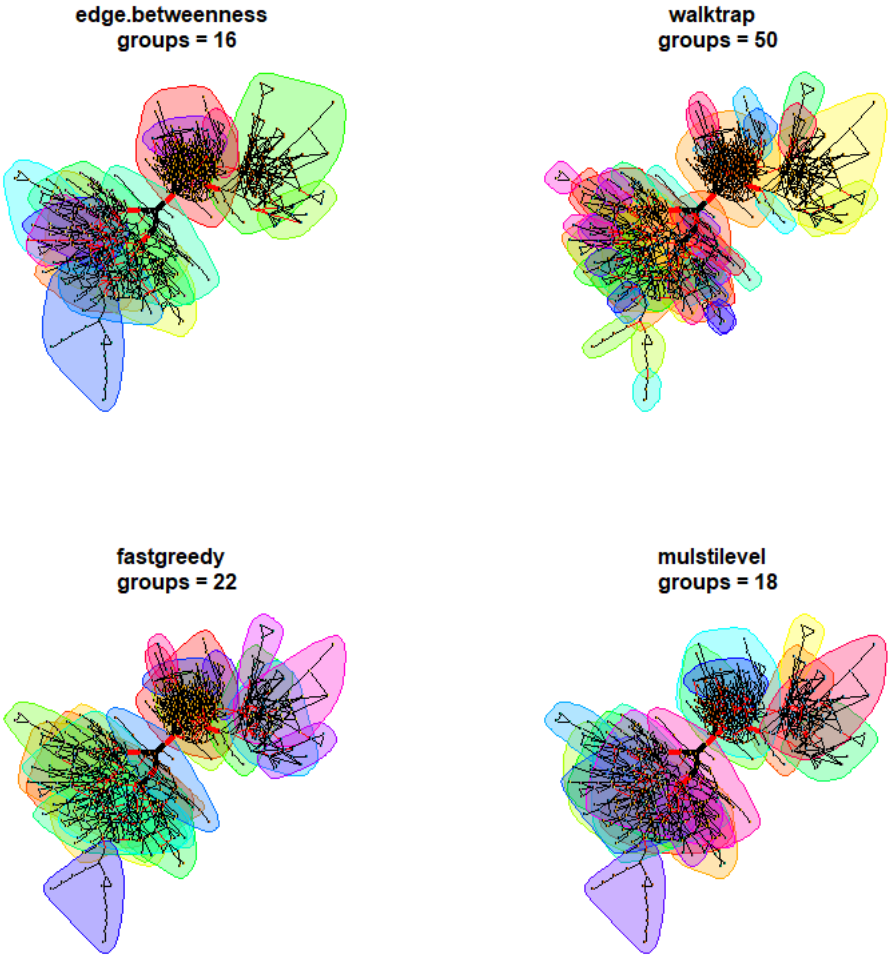

图 7 不同社区发现算法的社团效果图

（label.propagation 算法划分了 90 个社区，不予展示）

表 7 展示了不同社区发现算法的效果对比，以模块度来衡量社区划分好坏的指标。可以看出，模块度最高的是 multilevel 算法，它划分了 22 个社区，运行时间并不慢。模块度最低的是 label.propagation 算法，不到 0.7，它划分了 90 个社区，且由于算法的随机性，每次划分的情况并不相同。运行速度最快的是 fastgreedy，仅 0.01 秒就出结果，运行最慢的是 edge.betweenness，运行时间需要 3 秒。

表 7 中还显示，模块度低的算法为 walktrap 和 label.propagation，两者

均为流分析算法，说明拓扑分析算法更适用于本文研究的网络。

表 7 不同社区发现算法的效果对比

社区发现算法	发现社区数	模块度	运行时间/秒
edge.betweenness	16	0.7429	3.87
walktrap	50	0.7322	0.09
fastgreedy	22	0.7787	0.01
multilevel	18	0.7861	0.23
label.propagation	90	0.6942	0.32

参考文献

[1] E Kolaczyk G Csrdi. Statistical Analysis of Network Data with R[M]. Springer.2014

[2] 陈逸波.社会网络分析：探索人人网好友推荐系统.统计之都，https://cosx.org/2011/04/exploring-renren-social-network

[3] R 语言 | SNA-社会关系网络—igraph 包（社群划分、画图）（三），https://blog.csdn.net/sinat_26917383/article/details/51444536