**聚类技术——复杂网络社团检测**

学号：18030100101 姓名：张帅豪

1. 实验背景

复杂网络是描述复杂系统的有力工具，它不仅是一种数据的表现形式，同样是也一种科学研究手段。钱学森对于复杂网络给出了一种严格的定义：具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络成为复杂网络。

复杂网络社团结构定义为内紧外松的拓扑结构，即一组节点的集合，集合内的节点交互紧密，与外界节点交互松散。复杂网络社团结构检测广泛的应用于信息推荐系统、致癌基因识别、数据挖掘等领域。近年来，社区检测得到了快速的发展，这主要是由于Newman提出了模块度(modularity)的概念，从而使得网络社区划分优劣可以有一个明确的评价指标来衡量。模块度越大，对应的社区划分越合理。

社团检测就是在复杂网络上做聚类，聚类出来的就是社团。

1. 实验内容

某跆拳道俱乐部数据由34个节点组成，由于管理上的分歧，俱乐部要分解成两个社团。

**该实验的任务即：要求我们在给定的复杂网络上检测出两个社团。**

1. 分析与设计

**实验思路分析如下：**

1. 聚类算法通常可以描述为用**相似度**来衡量两个数据的远近，**搜索**可能的划分方案，使得**目标函数**达到极值。目标函数通常与相似度关系密切，例如目标函数是同类中数据相似度的平均值。

2. 类似的，对于社团检测（复杂网络上做聚类），我们有三个关键问题：

·希望得到什么样的社团？

·如何衡量数据的相似度？

·如何搜索得到最优解？

下面我们围绕解决这三个问题进行本实验算法的说明：

**问题一：**在本实验中，由于复杂网络的数据结构特点，我考虑从社团结构而不是两点之间的距离去定义社团。我希望检测到“内部链接密集，外部链接稀疏”的两个社团。

**问题二：**明确了希望得到什么样的社团，下面解决如何**衡量数据的相似度**以及**目标函数**的构造。

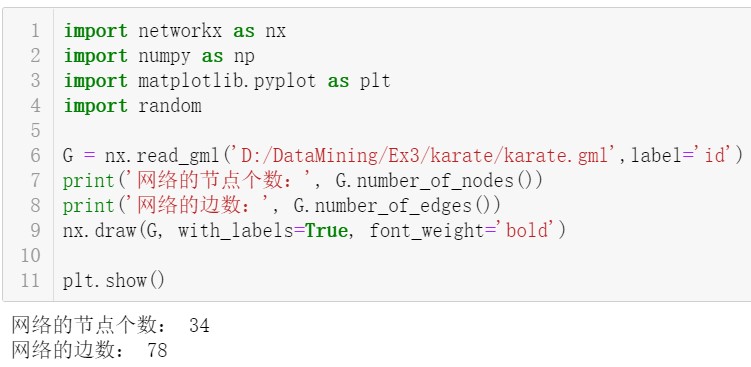
1. 给定节点i, 其邻居节点定义为与该节点相链接的所有节点组成的集合，给定一对节点（i,j）,其**相似度**定义为这个两个节点的公共邻居节点个数与邻居节点的并的个数的比值， 即： ，其中表示集合中元素的个数。该相似性度量很好的体现了我们所定义社团“类内密集，类间稀疏”的特性，值越大，表明两个节点之间的联系越紧密。
2. **目标函数构造。**定义平均相似度函数，其值的大小表示了一个社团内部的紧凑程度即社团密度。

**问题三：**采用贪心算法（爬山算法）搜索社团。

随机选择一个未聚类的节点作为当前社团C，提取出社团C所有未聚类的邻居节点。选择使得**社团密度降低最小**的那个节点v添加到社团C中，更新当前社团为。持续该过程直到当前社团的密度**小于某个阈值**。该搜索算法得到的是局部最优解而并非全局最优解。

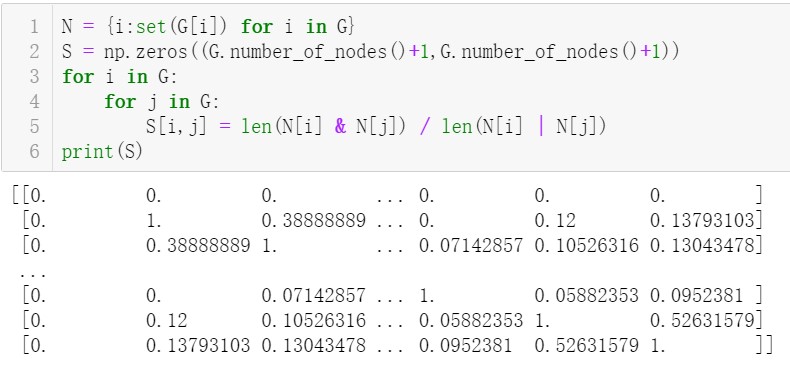
1. 实验详细

**1. 导入复杂网络数据集**



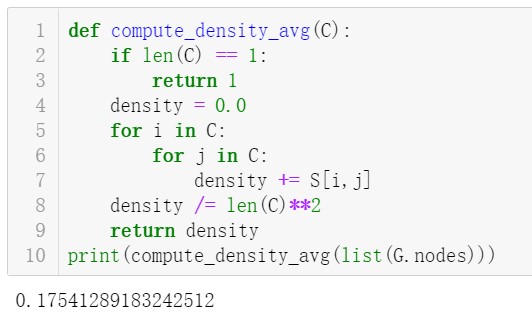
Networkx是一个用Python语言开发的图论与复杂网络建模工具，内置了常用的图与复杂网络分析算法，可以方便的进行复杂网络数据分析、仿真建模等工作。调用nx.read\_gml()函数读取实验给出的.gml数据，构建复杂网络G。

**2. 计算相似度矩阵S**

****

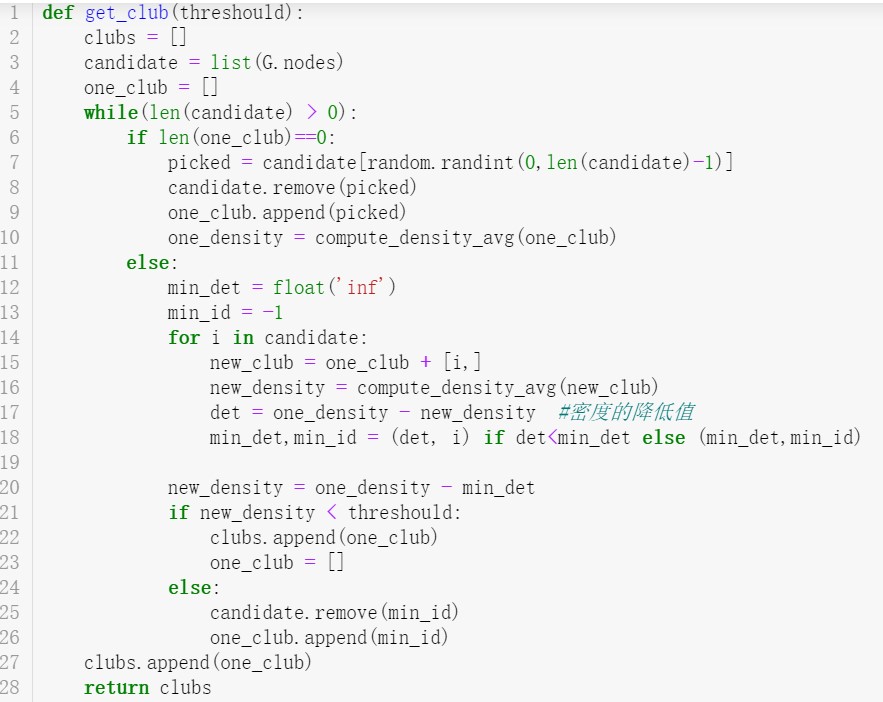
N为一个字典，“节点id—邻接点id集合”为其键与值对应关系，所以N[i]即返回的节点i的邻接点的集合，可以进行集合运算实现公式从而得到两节点的相似度。最终得到相似度矩阵S。

**3. 定义目标函数**

****

实现目标函数，调用该函数返回该社团内部的紧凑程度即社团密度。

**4. 贪心算法搜索社团**

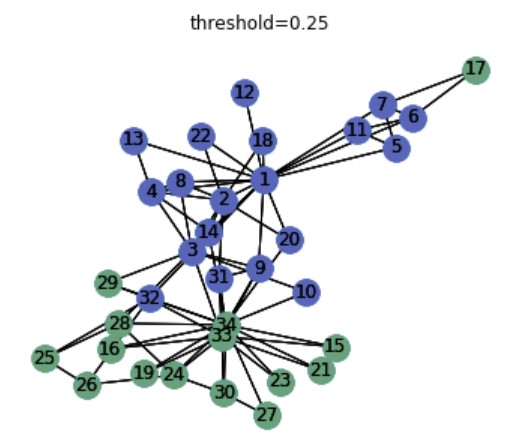


传入阈值threshould，贪心地给社团中加入使得社团密度降低最少的点，当该社团的社团密度小于阈值时，该社团搜索成功，然后继续搜索其他未加入任何社团的节点，直到所有节点加入社团。函数最终返回一个列表，列表中每个元素是保存了一个社团内节点id的列表，元素个数即等于所分得的社团个数。

**5. 可视化社团检测结果**

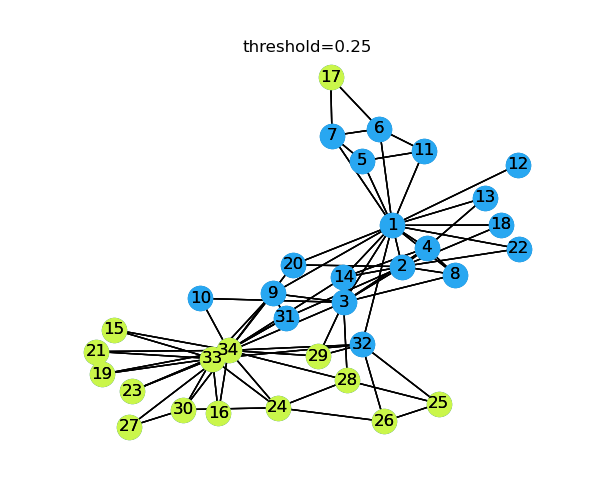


实验要求将网络分为两个社团，经过阈值threshould的调整得到当阈值=0.25附近时，复杂网络能较好地划分为两个社团。最后用networkx画图相关函数绘制出划分好的社团。其中一个结果如下图：



1. 实验结果

实验代码和注释最先在Jupyter上完成，可以查看每一步的结果。最终的.py代码在PyCharm编辑器中完成，最终完成了实验任务，能较好地将实验给出的跆拳道俱乐部数据划分为两个“内部链接密集，外部链接稀疏”的社团，一个结果如下图所示。



1. 心得与反思
2. 完成实验之前，我学习了networkx库的基本操作。认识了networkx在图论和复杂网络数据分析领域的强大功能，学会了创建图、访问网络节点、访问网络的边、读取写入文件以及用networkx提供的画图相关函数绘制出网络图等操作。
3. 深刻理解了复杂网络的概念、复杂网络上进行社团检测的算法思想与聚类技术的联系与区别。社团检测过程中最关键的两个问题是：

·希望检测到什么样的社团？

·如何衡量数据间的相似度？

1. 动手实现了一个简单的社团检测问题，解决了社团检测算法中的两个关键问题，得到了较好地检测结果。
2. **问题反思：**该实验中搜索社团的算法是满足局部最优解的贪心算法，该算法不能得到全局最优解。