Web信息处理与应用实验二程序说明——郭秋洋（PB15111650）

**社区发现算法实现与比较**

**【实验内容】**

实现 spectral clustering 等几个社区发现算法，并比较实验结果。

**【实验环境】**

编程语言：matlab

编程环境：matlab 2017b

运行环境：windows10 pro 2.20GHz jre1.8.0\_151

使用工具：matlab、Excel、gephi

**【实验步骤及方法】**

1. **五个社区发现算法的实现**

alinkjaccard 算法：

（1）调用 pdist 函数计算矩阵的 jaccard 相似度

（2）调用 linkage 函数使用 average 规则进行连接

（3）调用 cluster 函数将连接得到的矩阵聚类

girvannewman 算法：

（1）使用 betweenness\_centrality 计算每条边的介值中心性

（2）每次找到介值中心性最大的边，将其删去

（3）再调用 components 函数检查连通块的个数

（4）若连通块个数已达到要求的聚类个数，算法停止，否则算法继续迭代

rcut 算法：

（1）构造拉普拉斯矩阵 L = D – A，其中 A 为邻接矩阵，D 为度数矩阵

（2）调用函数 eigs 求拉普拉斯矩阵 L 最小的 k 个特征值所对应的特征向量

（3）利用 k 个特征向量构造 n×k 的矩阵

（4）将 n×k 的矩阵看成 n 个 k 维物体，调用 kmeans 进行聚类

ncut 算法：

（1）构造归一化的拉普拉斯矩阵 L = D^(-1/2) \* ( D – A ) \* D^(-1/2)

（2）调用函数 eigs 求拉普拉斯矩阵 L 最小的 k 个特征值所对应的特征向量

（3）利用 k 个特征向量构造 n×k 的矩阵

（4）将 n×k 的矩阵看成 n 个 k 维物体，调用 kmeans 进行聚类

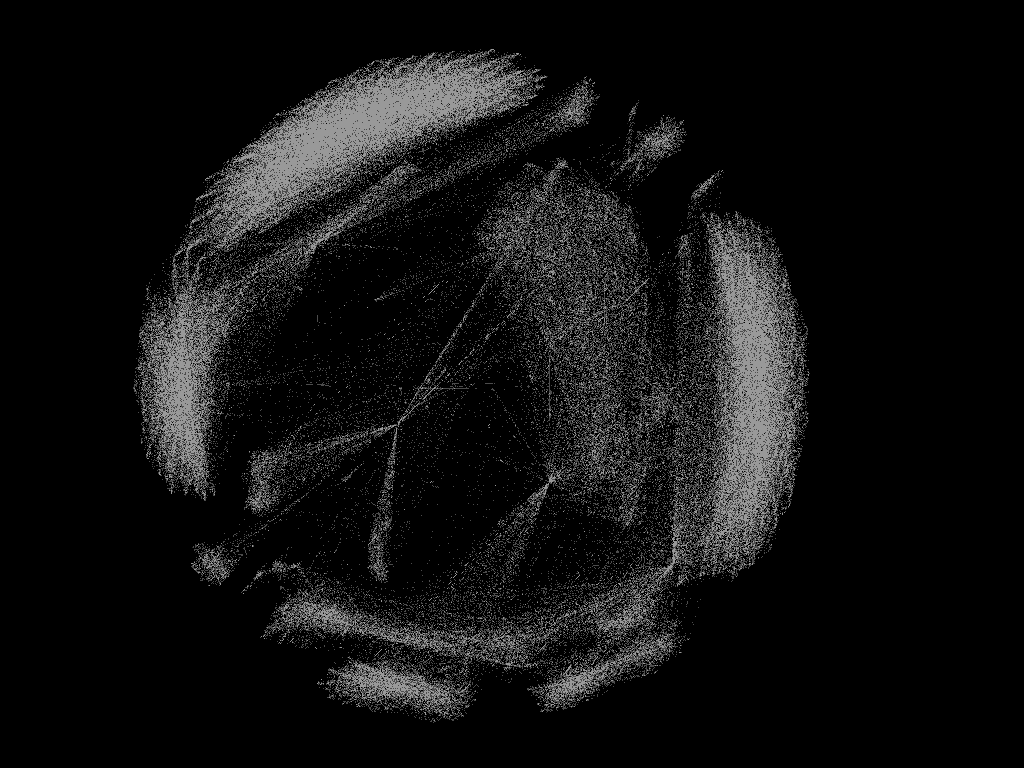
modularity 算法：

（1）构造矩阵 B = A - ddT/2m，其中 d 为度数列向量，m 为总边数

（2）调用函数 eigs 求矩阵 B 最大的 k 个特征值所对应的特征向量

（3）利用 k 个特征向量构造 n×k 的矩阵

1. 将 n×k 的矩阵看成 n 个 k 维物体，调用 kmeans 进行聚类
2. **Facebook-Egonet K的选择**

****将边数据导入 gephi 中运行布局算法，得到布局结果。粗略观察可得大约十五个社区，所以k选为15。

1. **可视化所需的csv文件的制作**

边数据： （Source，Target）

edge\_shift()：

（1）使用[ x, y ] = find( A )获得稀疏矩阵 A 非零的行列坐标

（2）此时 x 和 y 均为列向量，先对其转置变为行向量 x’和 y’

（3）拼接出坐标行向量 z = [ x’ ; y’ ]，再转置得到坐标列向量 z’

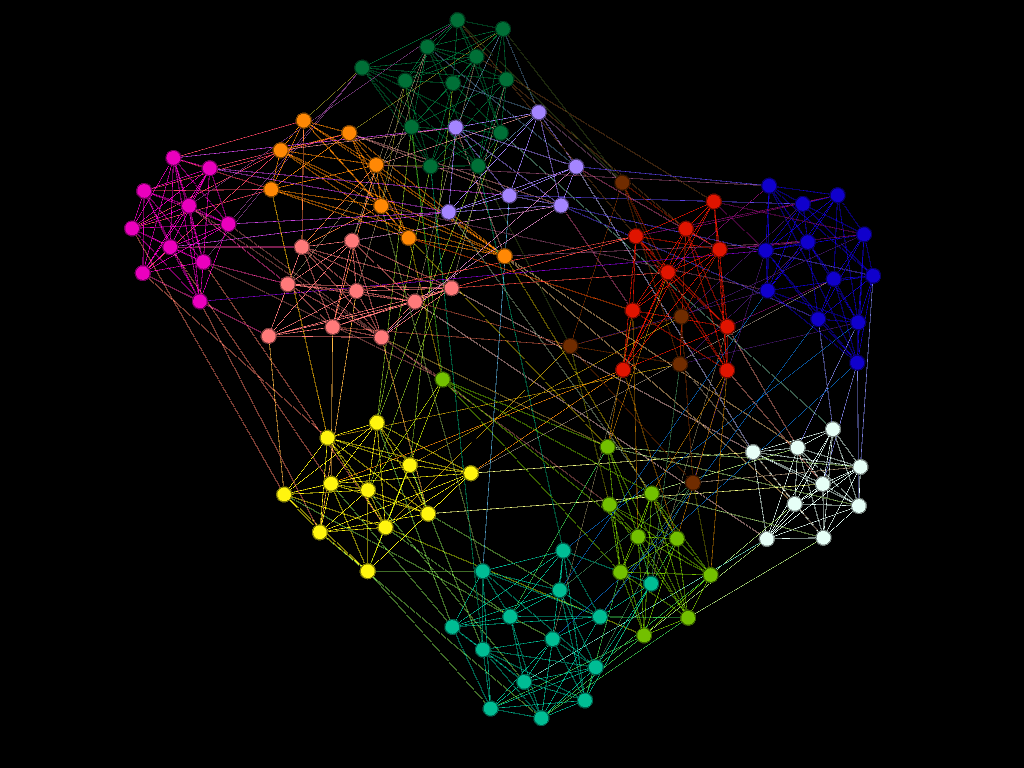
（4）使用 dlmwrite 函数将 z’写入csv格式文件

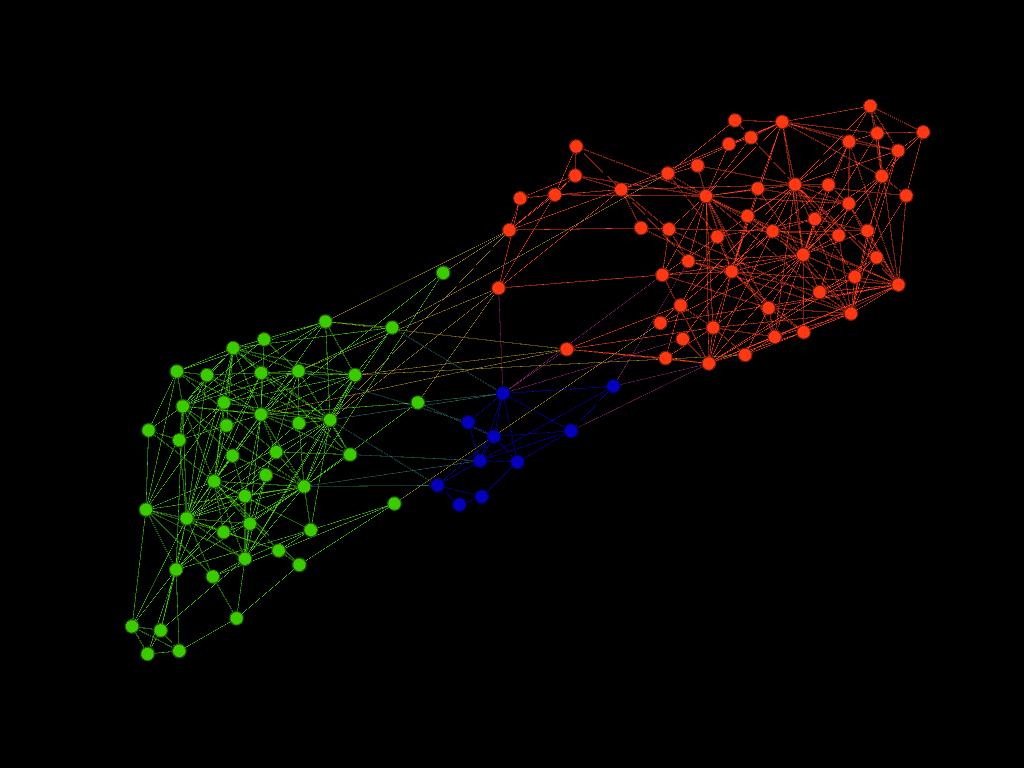
点数据： （Id，Modularity Class）

（1）直接将 matlab 运行后的输出文件拷贝一份，后缀名改为.csv

（2）将.csv 文件用 excel 打开，在其 Modularity Class 前添加顺序 Id 即可

**【实验结果说明及演示】**

1. **前两组数据与 ground truth 的比对结果**
2. **聚类的可视化**

football\_ncut的效果图

polbooks\_alinkjaccard的效果图

**3.实验分析**

1．五个算法在 football 数据集的 NMI 与 ACC 指标均很低，只有不到0.2，在polbooks数据集的NMI和ACC指标尚可，估计算法的准确度与数据集规模和社区的多少有一定关系

2. 大部分情况下 RatioCut 和 NormalCut 都能取得不错的结果，因为 Cut 的算法可以直接找出连接较薄弱的集合，分割出更加聚集的集合。

3．girvannewman 的复杂度最高，在Egonet数据集运行girvannewman算法所占的时间约为整个程序时间的99%，而 alinkjaccard 算法的时间复杂度最低。但对比两算法的NMI和ACC可知两算法在聚类效果没有明显差别，由此可见，alinkjaccard算法优于 girvannewman算法

4．观察NMI和ACC以及五个算法聚类效果图可知，modularity 算法比其他算法的聚类效果略差

5．本次实验所涉及的五个算法都需要人为确定 k 值，存在较大的局限性

**【实验总结】**

1．通过本次实验第一次接触 matlab 编程，掌握了 matlab 读写文件，矩阵运算，函数调用等的一些基本编程操作

2．实验要求的函数大多matlab或工具箱自带，直接调用即可，我深深感受到了matlab处理矩阵运算的方便性

3．通过对五大算法的直接实现和可视化聚类效果，加深了对社区发现的理解，认识到了各算法的优缺点

4．可视化聚类效果让我接触到gephi这样有趣而强大的软件

**影响力最大化算法**

**【实验内容】**

。。。。。

**【实验环境】**

编程语言、编程环境、运行环境、使用工具等等。

**【实验步骤及方法】**

写出实验的主要步骤及实现方法，给出关键部分的算法说明，参考实验要求中需要实现的几点内容。（可以写关键部分的伪代码，但是不允许粘贴源码）

**【实验总结】**

给出这次实验自我总结，可总结目前社区发现算法的不足，并提出创新性和改进。