西南交通大学
Southwest Jiaotong University

基于集成学习的社区发现算法研究

信息学院与技术学院

指导老师：陈红梅

组员：黎家昊 史晨阳
黄晋涛 邱凯

目 录

CONTENTS

1、研究背景及目的

- 研究背景
- 研究目的

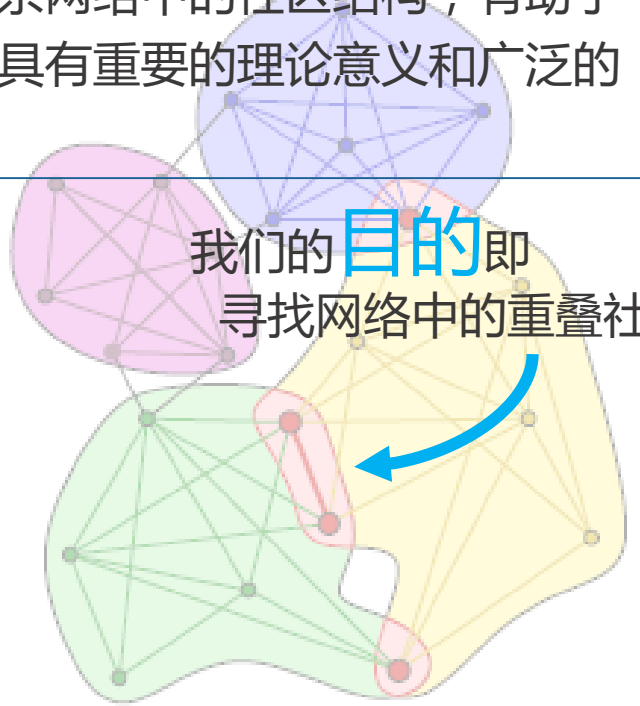
研究背景及目的

研究背景

复杂网络是复杂系统的抽象，几乎所有的复杂系统都可以用网络建模进行研究，如万维网、社交网络、交通运输网等。研究表明，复杂网络具有小世界现象等特性，其中网络簇也被称为社区，是复杂网络的最重要性质之一。所谓“社区”是指网络中那些联系比较紧密的节点组成的点集。社区发现能帮助挖掘复杂网络中的社区结构，有助于更深入地理解复杂网络地特性和功能，具有重要的理论意义和广泛的应用前景。

什么是重叠社区？

重叠社区发现允许一个节点同属于多个社区。重叠社区具有更现实的意义，首先重叠节点是必然是社区中重要的节点， 其次重叠社区反映了更加真实的社会网络结构。

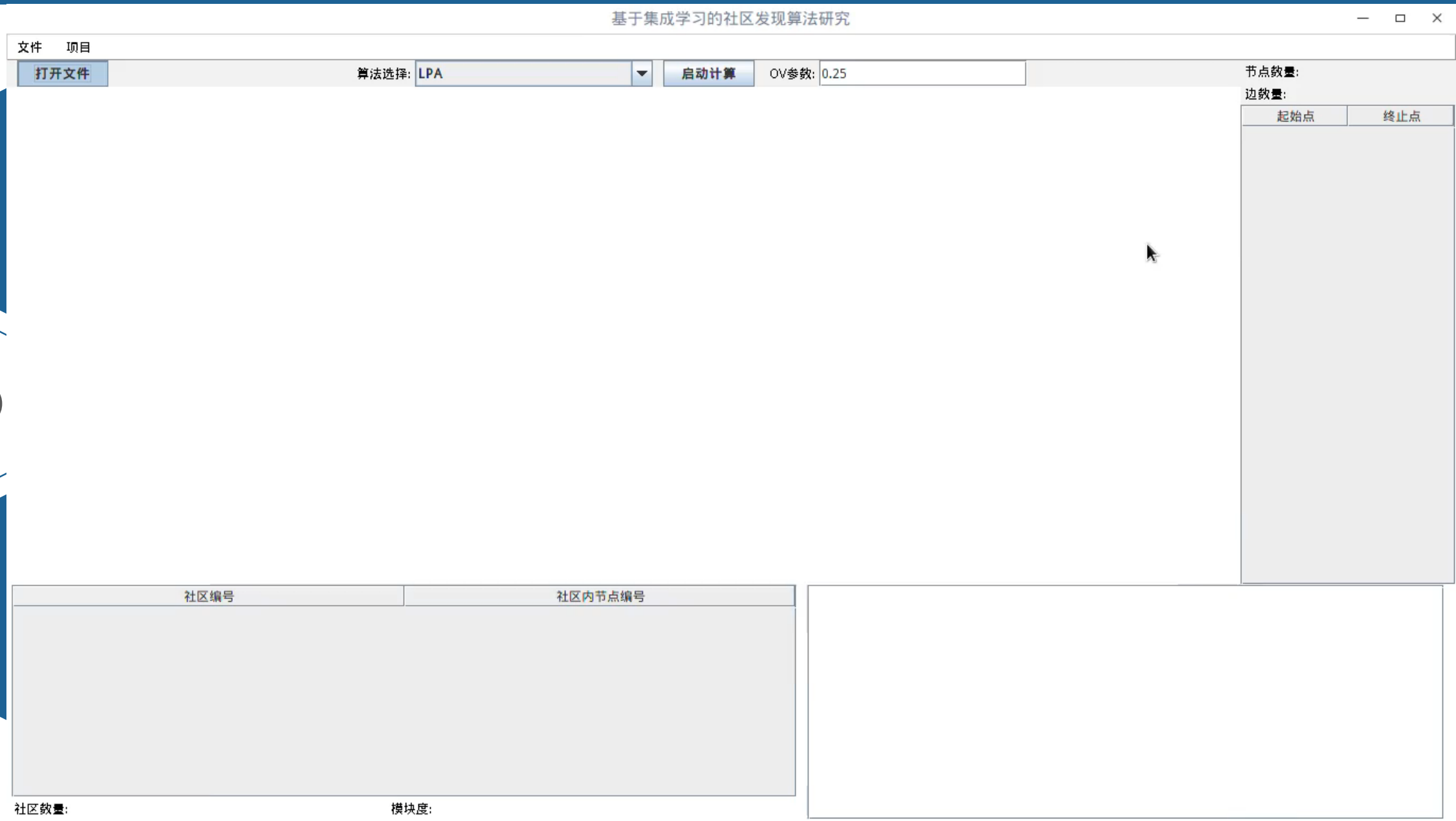


2、成果演示

- Demo

成果演示

DEMO


基于集成学习的社区发现算法研究
文件
项目
打开文件
算法选择:
LPA
启动计算
OV参数:
0.25
节点数量:
边数量:
起始点
终止点
社区编号
社区内节点编号
社区数量:
模块度:

3、关键技术及难点

- 改进LPA
- 改进CPM
- 创新点——集成学习

关键技术及难点

改进的标签传播算法

01

改进思路之第一步

节点间标签的传播不应该是对等的传播关系，高影响力节点的标签应该更容易地传递给低影响力的标签，反之，低影响力节点的标签的标签更加难以传播。

02

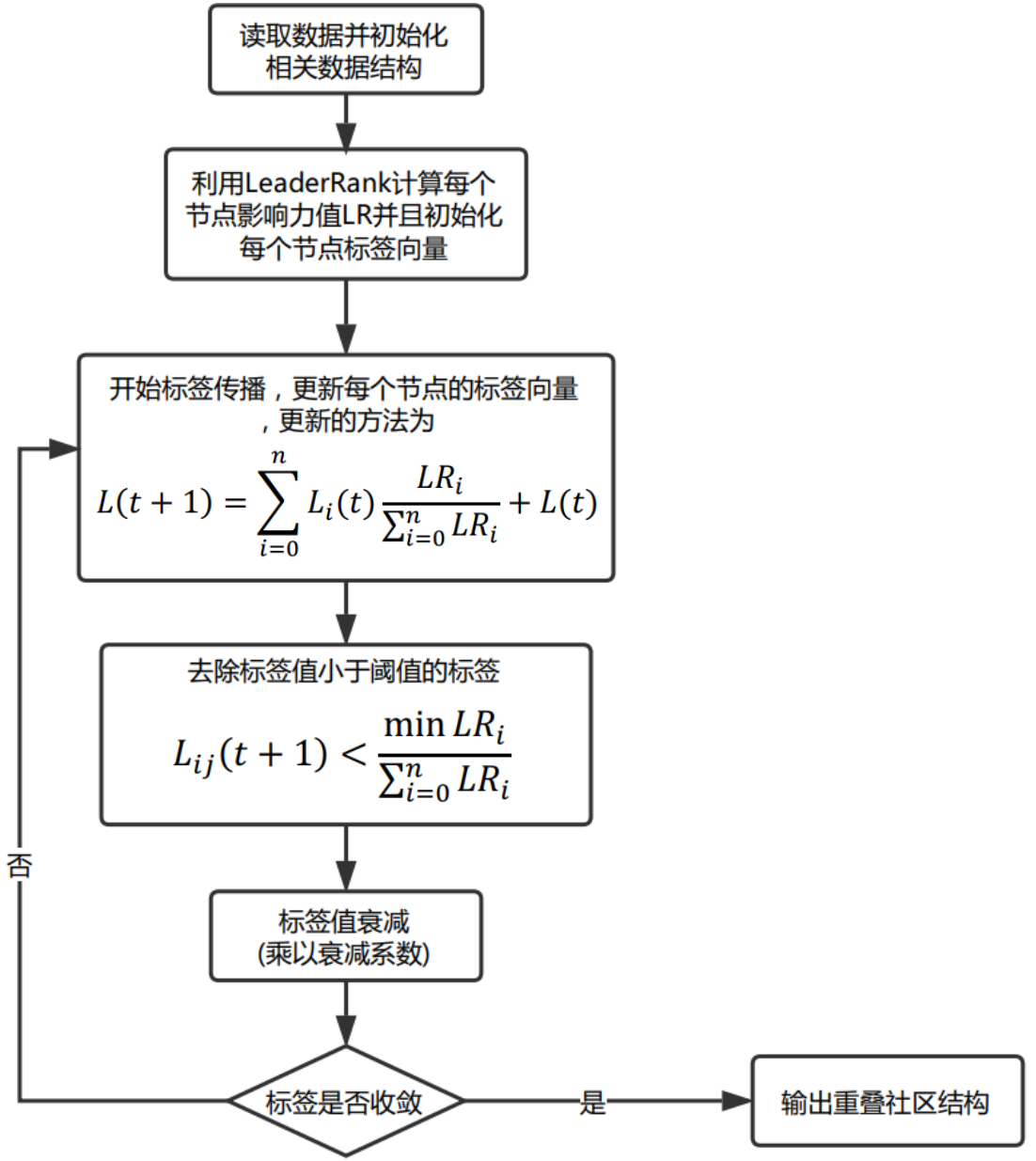
改进思路之第三步

为了发现重叠社区，每个节点不应该只保留一个标签，应该使用一个数组来保存接受到的所有标签信息，这样处理也可以保留标签传播过程中的历史信息。

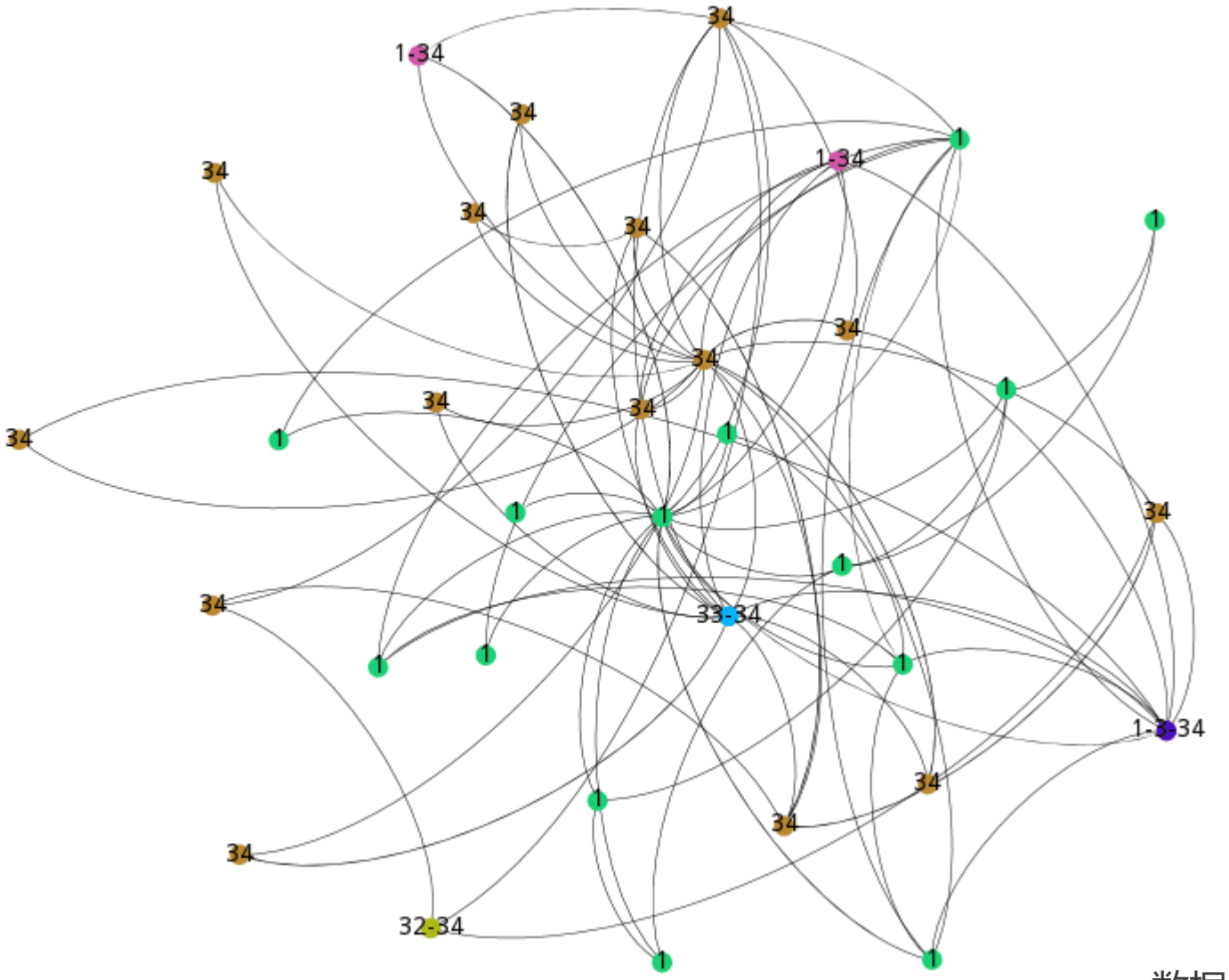
03

改进思路之第五步

在迭代终止时，高影响力节点的标签会以高标签值存在于它周围的所有节点中，这样的话，这些持有高影响力节点标签的节点也就自然成为了一个社区。

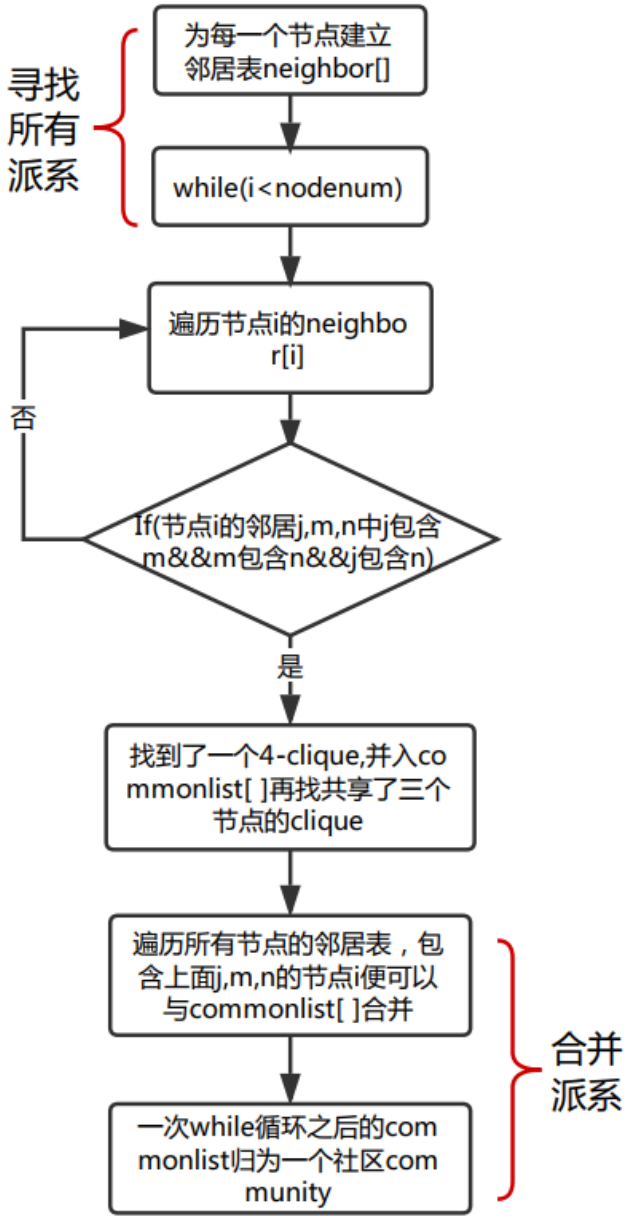


改进的标签传播算法结果可视化



数据集：karate.gml

关键技术及难点



派系传播算法

研究思路之第一步

01

FindAllClique是一个NP问题，非常消耗内存、时间。因此为了降低计算成本，应该首要解决如何快速找到所有clique的问题。

研究思路之第二步

02

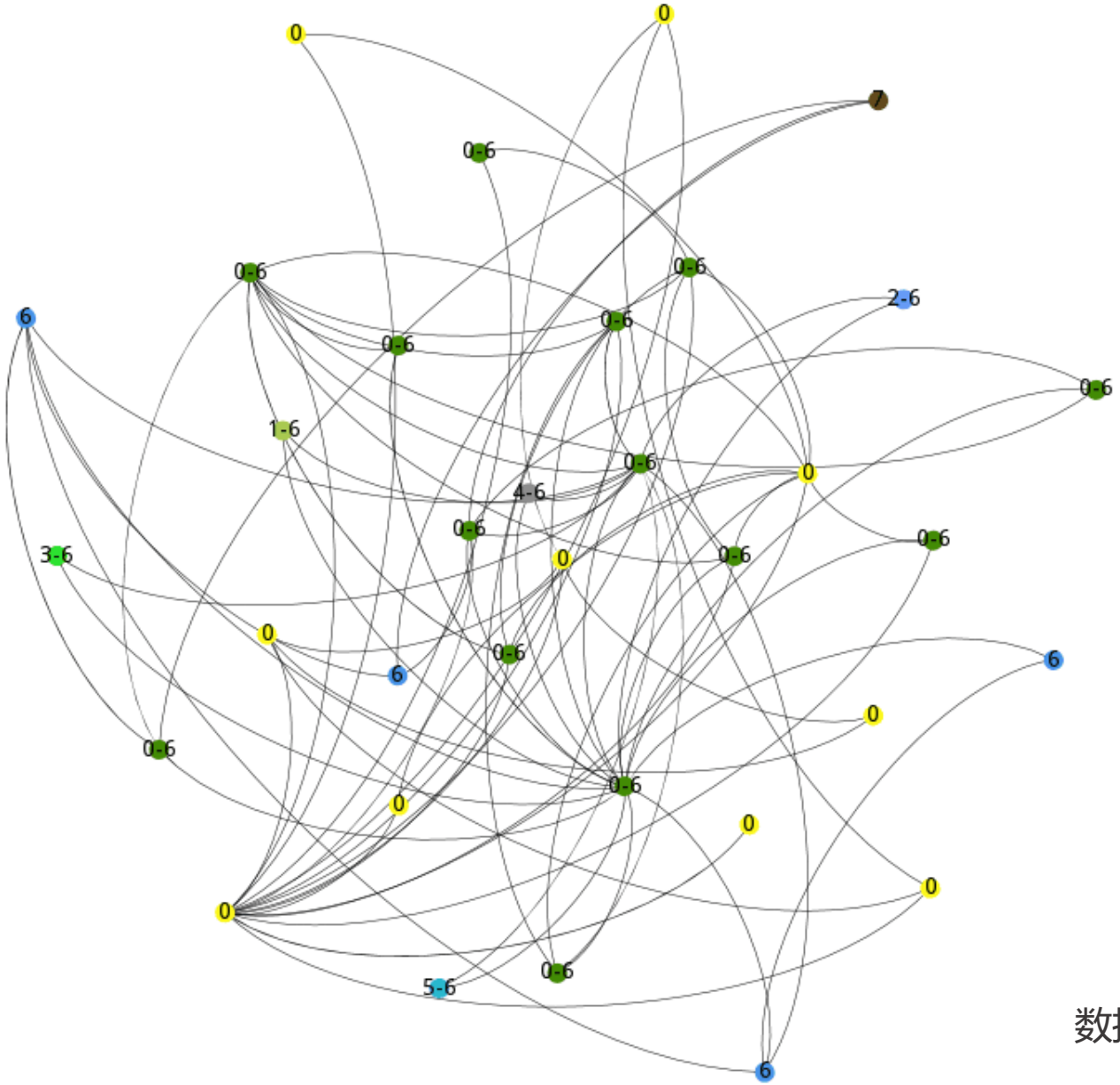
clique的划分太过于局限，每一个clique的节点数量固定相同，是强连通的。这样会导致潜在社区被隐藏。因此应该找一种发掘弱连通clique的方法，从而也会大大降低计算复杂度。

研究思路之第三步

03

判断两个clique的相似度是基于共享的k-1个节点数量。但是在图计算领域有非常多的计算相似度的方法，也许会得到更好的效果。

改进的标签传播算法
结果可视化



数据集：karate.gml

将集成学习思想应用于社区发现

社区发现本质是图的聚类

为了获得鲁棒性好且稳定的聚类性能, 人们提出了聚类集成算法. 聚类集成被认为在很多方面都能够超越单个聚类算法的性能, 如鲁棒性等。

如何实现

个体聚类的产生

为了节约计算成本，采用最简单的LPA算法生成*n*种社区划分方案。

个体聚类的选取

不能采用平等的方式对所有的个体聚类进行集成，所有我们采用CCChooser从多个社区划分方案种选出m种

个人聚类的集成

根据选取的各个体聚类所包含簇的关系建立簇相似图, 然后通过层次软聚类来实现聚类集成。

4、收获和总结

- 学会了什么

收获与总结

文献

知道如何去看科研文献，如何对文献进行分类，如何总结每篇文献中的创新点，知道了读文献时一定要注重摘要、实验方法和结论，以及最后的讨论部分（因为作者通常会在这里论述本实验本研究的不足以及改进方法，给未来的研究指明方向）。

编码

通过此次项目，知道了时间复杂度和空间复杂度对程序的性能有极大的影响，所以在编写程序时一定时注意代码的优化。

创新

大量的阅读别人的论文，了解到了该如何在已有的研究基础上进行创新，如何改进现有的算法使其拥有更好的稳定性以及更快的收敛。

合作

原来基本上都是一个人完成一整个项目，现在知道一组人如何去完成一个项目，该如何分工、如何配合、如何整合。

谢谢聆听

信息科学与技术学院

指导老师：陈红梅老师

答辩人：黎家昊