天津财经大学

《大数据与互联网思维》课程报告

****

|  |  |
| --- | --- |
| **题目：** | 面向社交网络的社区发现算法研究 |
|  |  |

|  |  |
| --- | --- |
| **院系名称：** | **理工学院信息科学与技术系** |
| **专业班级：** | **网络1301班** |
| **学 号：** | **20161234567** |
| **姓 名：** | **陈 媛** |
| **指导教师：** | **饶 俊** |

**2017年 5月 4 日**

目录

[一、绪论 1](#_Toc510639134)

[（一）复杂网络研究背景及研究意义 1](#_Toc510639135)

[（二）社区发现技术现状 1](#_Toc510639136)

[（三）本文组织结构 2](#_Toc510639137)

[二、复杂网络理论概述 3](#_Toc510639138)

[（一）复杂网络及其特点 3](#_Toc510639139)

[（二）复杂网络模型 4](#_Toc510639140)

[（三）社交网络及其分析方法 5](#_Toc510639141)

[三、拉普拉斯矩阵的社区发现算法 5](#_Toc510639142)

[（一）拉普拉斯矩阵简介 5](#_Toc510639143)

[（二）图分割理论 7](#_Toc510639144)

[（三）基于代数连通性的社区发现算法 8](#_Toc510639145)

[（四）算法的优缺点 9](#_Toc510639146)

[四、基于谱聚类的社区发现算法 10](#_Toc510639147)

[（一）K-means算法 10](#_Toc510639148)

[（二）谱聚类社区发现算法 10](#_Toc510639149)

[五、基于模块度的社区发现算法 13](#_Toc510639150)

[（一）GN算法 13](#_Toc510639151)

[（二）基于模块度的谱优化社区发现算法 14](#_Toc510639152)

[（三）实证分析 16](#_Toc510639153)

[六、总结与展望 18](#_Toc510639154)

[（一）论文总结 18](#_Toc510639155)

[（二）提出问题与展望 19](#_Toc510639156)

# 一、绪论

## （一）复杂网络研究背景及研究意义

复杂网络是由许多个社区构成的，因此社区发现对于复杂网络的进一步探索具有重要意义。对于一个复杂的系统，我们可以将其看成是一个复杂网络，而这个系统中的个体或组织也就是复杂网络中的结点或社区，个体或组织之间的联系就可以用网络中邻接结点之间相连的边来形象地表示。在便于研究的基础上，我们把复杂网络表示成图，通过对图的描述来研究该网络的一些性质。

## （二）社区发现技术现状

在真实世界网络中，有时会有结点同时具有多个社区的特性，能够划分到不同社区。这时，重叠社区算法更适合我们的应用。

但是这两种社区发现算法相比较而言，非重叠算法已经比较成熟，主要包括基于模块度谱优化算法、谱聚类算法和影响力矩阵算法等，本文主要研究对象为非重叠社区发现算法。

谱聚类算法是传统聚类算法K-means的改进，主要是计算相似性矩阵，将高维度的数据，以特征向量的形式简洁表达，通过降维的方式能够更好的使用K-means算法进行聚类。

## （三）本文组织结构

第一章，绪论，首先说明了复杂网络的研究背景及研究意义，接着对社区发现算法的研究现状进行了分析，最后介绍了论文结构。

第六章，总结与展望。本章分为两个部分。主要针对全文做出总结，并提出需要改进和研究的方面。

# 二、复杂网络理论概述

## （一）复杂网络及其特点

复杂网络的定义：在真实世界中，我们把具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络称为复杂网络。

复杂网络具有以下三个特点：

（1）复杂网络是大量真实复杂系统的拓扑抽象；

（2）复杂网络的生成机制比规则网络和随机网络更复杂；

（3）复杂网络是大量复杂系统得以存在的拓扑基础[[1]](#footnote-1)。

为了方便研究，我们通常把复杂网络用图来描述，假设图有n个点，m条边，可以抽象表示为G=（V,E）,其中V表示顶点集且n=|V|,E表示边集且m=|E|。E中的每个元素都与V中的一组结点（i,j）相对。如果图中的每一组结点最多只有一条边与之相对应，那么该图为无向图；反之，则为有向图。若是在该网络图中，连接结点的边均与不同权值相对应，则为加权图。在本文中所研究的均为无向图。复杂网络一般具有以下特性：

1.小世界

小世界用简练语言形象地表示出了虽然真实世界的大部分网络的范畴庞大，然而网络图中的任何两结点间都有一条短路径。

从平常描述来看，它表现出来的是虽然网络图的结点之间相互关联的数目很少，但是却能够将整个世界连接起来的这样一种状态。就像在社交网络中，虽然每个人认识的人是有限的，这种认识关系也是有限的，可是我们却可以通过认识的人找到离我们遥远的没有关系的其他人，所以说世界变得越来越小。

2.集群

即集聚程度的概念。比方说，社交网络中每个人有亲朋好友，认识很多人，组建成自己的朋友圈。简而言之，集群是指社交网络中社区化的程度：这是一种网络的内聚倾向，在一定程度上反映了社区结构。连通集团指的是在社交网络中各个社区联通和分布的情况[[2]](#footnote-2)。比如，连通集团概念能够表现不同朋友圈之间的连接关系。

## （二）复杂网络模型

复杂网络可分为这四大类：随机网络、小世界网络、规则网络和无标度网络。在本文中，我们重点讨论随机网络和无标度网络。

1.随机网络

ER模型是早期学者最早提出来的一个随机图模型。在上个世纪40年代，匈牙利数学家Erdӧs 和Rényi建立了ER Model。在随机图中，任意的两个结点之间有一条边相连接的概率为P。在ER模型中，对于任意一个具有N个结点的随机图，边的最多数目为N(N-1)/2；当图的边数为N(N-1)/2时，该图为完全无向图。所以对于边的总数目X来说，服从二项分布X~b(N(N-1)/2,P)，则期望值为P\*[N(N-1)/2]。由数学推理可知，当N足够大时，X服从泊松分布，此时ER的度分布为图2-1所示：

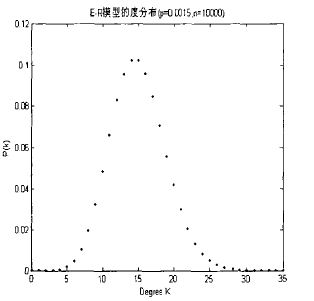


图2-1 E-R模型的度分布

## （三）社交网络及其分析方法

社交网络是由社会成员和社会关系组成的，比如，新浪微博，知乎，贴吧，这些都属于社交网络的一种。社交网络中的结点就是指社会成员，而网络是指成员之间关系的集合。通过这些社会关系，比如亲朋好友、团队等关系，甚至可以将相隔甚远彼此不认识的人或组织联系起来。

# 三、拉普拉斯矩阵的社区发现算法

## （一）拉普拉斯矩阵简介

1.拉普拉斯矩阵

拉普拉斯矩阵，又叫作导纳矩阵，是图的一种表现形式。将复杂网络描述成一个随机网络，记为图G=（V，E），有n个结点，m条边，且n=|V|，m=|E|。

介绍拉普拉斯矩阵之前，先要介绍与之相关的邻接矩阵W，邻接矩阵为对称矩阵，其元素用于描述复杂网络中结点的连接情况：

（3-1）

拉普拉斯矩阵的4个性质：

（1）L是对称半正定矩阵；

（2）L\*1=0\*1,即L的最小特征值是0，相应的特征向量为1。证明：

2.拉普拉斯矩阵的特征值

拉普拉斯矩阵的特征值在社区发现算法中具有重大的研究价值，在算法研究的很多方面都有举足轻重的作用。其中，最重要的特征值为第二小特征值，基于第二小特征值的社区发现算法一般用于将社交网络划分成2个社区。

根据上图能够观察得出，在代数连通度对应的特征向量中，元素值根据正负情况能够被划分到不同的社区。按照这样的划分情况，如果社交网络中的社区个数为奇数时，很容易把社区分割成两个独立的子社区中。

## （二）图分割理论

社区发现问题可以约简为在网络中发现最小割集的方法，即为图分割问题[[3]](#footnote-3)。

1. 在介绍图分割问题前我们需要用到一个简单定义，不同子图之间所有边的权值和，其中、是的子图，为边对应的权值。

## （三）基于代数连通性的社区发现算法

基于代数连通性的社区发现算法主要是建立在迭代二分法的基础上，它的主要思想是先把网络图划分成两个子图，再迭代，最后再得出需要划分的子图数。

## （四）算法的优缺点

1.优点

相比基于拉普拉斯矩阵的二分社区发现算法，采用迭代二分法来划分社交网络，可以划分成为多个社区。基于拉普拉斯矩阵的社区发现算法，在社交网络实际划分情况为二分时，根据费德勒向量的划分结果很理想。但是实际情况并不是每个网络都是而二分的，对于多分的网络，迭代二分法体现了一定的优势：

（1）适用于网络多分，在某些情况下，可以很快地得到比较好的社区发现结果。

（2） 社区分割时考虑到结构的“平衡”[[4]](#footnote-4)，而不是切割最少的边，划分结果比较理想。

2.缺点

这个算法虽然有改进，但是在一般的情况下，利用该方法得到的社区发现结果却不如人意，该算法的缺陷也很明显：

（1）并没有从本质上改变cut size对于社区结构体现的不足。

（2）分割的结果总是倾向于平分，但实际上这个不一定合理。

# 六、总结与展望

## （一）论文总结

本文所提出的算法均是对于多分社区的，在传统的社区划分算法基础上有所改进，相比之下具有一定的优势。本文重点分析了所提出的三种算法的优缺点，并对计算公式进行了较为详细的推导。图分割理论作为许多算法发现的理论基础，值得我们进一步的探索，提出更好的分割方法可以适用于多分社区，而不只是局限于二分社区。拉普拉斯矩阵也是我们研究复杂网络的一个很好的奠基石，它的特征值具有重要的意义，除了拉普拉斯半径和代数连通性，其他的还有待于我们进一步的发现其实用价值。

本文所介绍的算法都是社区发现研究中比较经典的算法。在研究中发现，拉普拉斯矩阵有三种不同的表现形式，根据拉普拉斯矩阵的不同，得到的特征值和特征向量会有很大差别，它们反映的特性也各不相同，只有选取合适的矩阵才能得到好的分析结果。拉普拉斯矩阵分为标准型和非标准型，本文所选均为非标准型。

## （二）提出问题与展望

在本文中，我们提到了基于代数连通性的社区发现算法、基于谱聚类的社区发现算法、基于模块度的社区发现算法，这些算法虽然在传统的社区发现技术上有所改进，但是仍存在许多缺陷。能否对算法进行改进，使得更准确地划分出多个社区，也值得我们探求。本文提出了这几种算法存在的优缺点，为后面学者研究出改进的社区发现算法提供一定的参考意义。

此外，社交网络社区结构分析对实际问题的解决还有着实用价值。因此，社区结构对分析解决实际问题的价值也需要进一步地探讨。

参考文献

1. 赵富强,张烁等.基于代数连通性的复杂网络割边模型研究[J].计算机工程与应用2014(11):135-138.
2. 程学旗，沈华伟.社会信息网络中的社区分析[J], 中国计算机学会通讯，第 7 卷 第 12 期 2011 年 12 月.
3. 张娜,复杂网络社区结构划分算法研究[D].硕士学位论文,大连理工大学,2009.
4. 李林荣．社交网络的特性及发展趋势[D].西南政法大学，2012.
5. 骆志刚，丁凡等.复杂网络社团发现算法研究新进展[D].国防科技大学,2011.
6. Girvan M, Newman M E J. Community structure in social and biological networks[J].Proceedings of the National Academy of Sciences of the United States of America, 2002,99(12):7821-7826.
7. Alex Arenas, Albert Diaz-Guilera, Conrad J. Perez-Vicente, Synchronization reveals topological scales in complex networks[J], Physical Review Letters, vol. 96, 114102, (2006).
8. Ingo Bürk. Spectral Clustering[D]. University of Stuttgart. 2012.Bachelor Thesis in partial requirement for the degree of the Bachelor of Science.
9. Zhou H J. Distance, dissimilarity index, and network community structure[ J ]. Phys Rev E, 2003, 67 (6) : 061901.
10. Y.Kim and M.Mesbahi,”On maximizing the second smallest eigenvalue of a state-dependent graph laplacian,”IEEE Transactions on Automatic Control,vol.51 no.1,January 2006,116-120.
11. 房产电商与中介利益之争，戳中各自O2O痛点[EB/OL].(2014-08-25). http://www.chinaz.com/start/2014/0825/364883.shtml.

1. 张娜,复杂网络社区结构划分算法研究[D].硕士学位论文,大连理工大学,2009. [↑](#footnote-ref-1)
2. 程学旗，沈华伟.社会信息网络中的社区分析[J], 中国计算机学会通讯，第 7 卷 第 12 期 2011 年 12 月. [↑](#footnote-ref-2)
3. Y.Kim and M.Mesbahi,”On maximizing the second smallest eigenvalue of a state-dependent graph laplacian,”IEEE Transactions on Automatic Control,vol.51 no.1,January 2006,116-120. [↑](#footnote-ref-3)
4. [Alex Arenas](http://arxiv.org/find/cond-mat/1/au:+Arenas_A/0/1/0/all/0/1), [Albert Diaz-Guilera](http://arxiv.org/find/cond-mat/1/au:+Diaz_Guilera_A/0/1/0/all/0/1), [Conrad J. Perez-Vicente](http://arxiv.org/find/cond-mat/1/au:+Perez_Vicente_C/0/1/0/all/0/1), Synchronization reveals topological scales in complex networks[J], Physical Review Letters, vol. 96, 114102, (2006). [↑](#footnote-ref-4)