学校代码：10286
分类号：TP311
密　级：公开
U D C：004.4
学　号：143465

东南大学

SOUTHEAST UNIVERSITY

工程硕士学位论文

基于社团的社交网络影响力传播分析

（学位论文形式：基础研究）

研究生姓名：张宗辉
导师姓名：何洁月　教授
王宏宇　高工

申请学位类别 工程硕士　学位授予单位 东南大学
工程领域名称 软件工程　论文答辩日期 2017年6月1日
研究方向 机器学习　学位授予日期 2017年　月　日
答辩委员会主席 姜　浩　评　阅　人 汪　鹏、王远斌

2017年6月1日

东南大学

硕士学位论文

基于社团的社交网络影响力传播分析

专 业 名 称：软件工程

研究生姓名：张宗辉

导 师 姓 名：何洁月

COMMUNITY-BASED ANALYSIS OF THE PROPAGATION OF INFLUENCE IN SOCIAL NETWORKS

A Thesis Submitted to

Southeast University

For the Professional Degree of Master of Engineering

BY

Zhang Zong-hui

Supervised by

Professor He Jie-yue

And

Senior Engineer Wang Hong-yu

College of Software Engineering

Southeast University

June 2017

东南大学学位论文独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

研究生签名：张宗辉 日期：2017年6月6日

东南大学学位论文使用授权声明

东南大学、中国科学技术信息研究所、国家图书馆有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括以电子信息形式刊登）论文的全部内容或中、英文摘要等部分内容。论文的公布（包括以电子信息形式刊登）授权东南大学研究生院办理。

研究生签名：张宗辉 导师签名 日期：2017年6月6日

摘要

从社交网络中发现具有广泛影响力的用户在很多方面具有重要作用，例如创新采用、社会舆论传播与导向、群体行为形成和发展等。通过社交网络中影响力的传播方式，商业营销能够以较低的费用将新产品推广到整个市场中，从而产生较大的社会影响力和商业价值。因此，从社交网络中发现具有影响力的 Top-K 节点集，对研究信息在网络上的快速传播具有重要意义。

所谓发现具有影响力的 Top-K 节点集，就是从网络中选取大小为 K 的用户集合，并以这个集合为触发集合，从而使信息在网络中的传播最大。然而影响力传播最大化是 NP-Hard 问题，经典的贪婪算法虽然能得到近似度较高的近似解，但是算法的计算代价太大。在网络中的社团结构具有社团内节点强连接，社团间节点弱连接的这一特性，信息的传播速度会受到社团的影响。因此，本文结合网络中社团的特性，在保证具有较高的近似解的前提条件下，采用网络拓扑信息和用户行为的综合分析方法，提升影响力算法的时间效率。论文的主要工作如下：

1、在单层网络中从研究节点的拓扑属性出发，基于 LeaderRank 的算法思想，在独立级联的传播模型下，在单层网中提出一种基于社团结构的 CLR（Community-based-LeaderRank）算法来发掘有影响力的 Top-K 节点集。CLR 算法通过结合 LeaderRank 算法的特点将社团的特性表现出来，然后计算得到的拓扑属性在网络中选择具有影响力的 Top-K 节点集合。

2、在多层网络中，利用多层网络中的网络融合分析和协同分析两种研究方法，提出多层网络中结合社团结构的 SMCLR（Super-Multi-Community-based-LeaderRank）算法，该算法将单层网络 CLR 算法延伸到多层网络中，从而解决了多层网络中具有影响力 Top-K 节点集问题。

在 Twitter 数据集上的实验结果表明，结合社团影响力的方法能够较好地发现网络中具有影响力的 Top-K 节点集。无论是单层网络中的 CLR 算法，还是多层网络中的 SMCLR 算法，在所选取的 Top-K 集合表现出的影响力上与传统的贪婪算法相差无几，但是从时间效率上来说，CLR 算法与 SMCLR 算法要远远优于贪婪算法。

关键词：社交网络，影响力，社团，信息传播

Abstract

Finding a wide range of influential users from the social network plays an important role in many areas, such as innovation adoption, diffusion and guidance of public opinion, formation and development of group behaviors. Through this transmission of influence in social networks, commercial marketing create greater social influence and commercial value, with promoting new products to the whole market at a lower cost. The discovery of the influential Top-K node set from the social network is of great significance to the research of rapid propagation of information on the network.

The discovery of the influential Top-K node set is selecting the set of users of size K from the network and using this set as the trigger set to maximize the spread of information in the network. However, the maximization of influence is NP-Hard problem. Although the classical greedy algorithm can get well approximate result of the optimum, the computational cost of the algorithm is too high. The community structure in the network has a feature that a strong connection between users within the communities and weak connection outside the communities, meanwhile the speed of information dissemination will be affected by the feature. Therefore, according to the characteristics of the community in the network, this article will import the time efficiency of the influence algorithm by using the comprehensive analysis method of network topology information and user behavior under the premise of guaranteeing the high approximate solution. The main work of the paper is as follows.

1、Because of the idea of LeaderRank algorithm, CLR(Community-based-LeaderRank) algorithm based on community structure is proposed to explore the influential Top-K node set in the single network. The idea of CLR algorithm is based on the characteristics of the LeaderRank algorithm to show the characteristics of the community, and then select the influential Top-K node set by calculating the topological properties of the network.

2、In the multi-network, the SMCLR(Super-Multi-Community-based-LeaderRank) algorithm combining the community structure is proposed, extending the CLR algorithm of single network to the multi-network and using the network integration analysis method and the collaborative analysis method, research propagation of the influence of multi-network.

In this paper, the results of experiment on Twitter datasets show that the performance of combination of community influence is better to find the influential Top-K node set than traditional algorithm. Whether it is CLR algorithm in single network or SMCLR algorithm in multi-network, the influence of selecting Top-K set is almost the same as that of the traditional greedy algorithm, the CLR algorithm and SMCLR

algorithm is far superior to greedy algorithm in the time efficiency.

Keywords: social networks, influence, community, propagation of information

目录

第一章 绪论

1.1 研究背景

在社交网络中，用户的行为和思想等信息在网络中传播，从而影响到网络中的其他用户，信息在传播过程中可能会被大部分用户所接受，也可能会逐渐消失。为了使更多的用户获取到某条信息，假设将这条信息通过触发用户传播给其他用户，并且使信息能够在网络中相邻的用户之间不停地传下去，如何选取这样的触发用户成为值得研究的问题。社交网络中影响力传播在社会生活和决策制定等方面发挥重要的作用，影响力分析在许多领域得到广泛应用，例如推荐系统、社交网络信息传播、链路预测、病毒式营销、公共健康、专家发现、突发事件检测和广告投放等[1]。

人们发现具有广泛影响力的用户，他们在创新采用、社会舆论传播和导向、群体行为形成和发展等方面具有重要作用，而且通过口口相传的影响力传播方式，商业营销能以较低费用将新产品推广到整个社交网络，从而产生较大的社会影响力和商业价值。为了探索用户的影响力及其在营销网络中的传播规律，计算机学界投入了大量的工作到相关问题的研究之中。

随着网络技术的迅速发展，在社交网络中海量的数据信息在用户与用户之间传播，从而研究人员所能够获得的数据急剧增多，使得在海量的数据中研究影响力的传播问题变得极具挑战。同时由于网络中的影响力最大化传播被证明是 NP-Hard 问题，所以利用传统的方法来研究社交网络中用户的影响力问题，不仅存在精度低的问题，而且算法的时间效率也不高。

通过结合网络中的拓扑属性与社团结构特性来标记具有影响力的节点，对网络中影响力传播问题的研究具有十分重要的意义。在网络中普遍存在的社团结构具有这样一个特性，即社团内部节点之间的连接比较紧密，而社团之间节点的连接相对稀疏。在研究影响力的传播问题时，社团内部节点之间的信息传播的速度和数量都远远大于社团外部的节点间信息传播。因此社团结构的研究，不仅可以将复杂庞大的网络以分而治之的方式来简化问题规模，而且还有助于研究网络中影响力的传播问题。

于此同时，由于不同的社团结构对在网络中的影响力差异很大，因此研究社团中节点的影响力也会受到这种差异的影响，通过研究以社团为单位的影响力传播，使得研究节点间的相互影响力更加准确。如何将网络中节点的拓扑属性与社团结构相结合，从而研究节点的影响力传播问题，是一个值得深入研究的课题。

然而在现今的互联网环境下，社交网络不仅仅是指具有单一性质的简单网络，在实际的社交网络中，用户之间的相关关系具有多样性。这种多样性使得我们不得不从多个角度来观测用户之间的关系，从而更准确地挖掘出用户之间的本质联系。对每种特定的用户关系来建立该关系下的简单网

络，从而将存在多种关系的社交网络就转变成多层社交网络。在这种现实存在多样的用户关系的多层网络中，研究同类与不同类用户关系的相互影响力传播问题将更具应用价值和挑战。

1.2 国内外研究现状

首先就社交网络影响力在相关文献中的定义进行阐述，然后介绍国内外影响力分析的相关研究现状，主要是从单层网络（single network）中影响力研究现状和多层网络（multi-network）中影响力研究现状两方面来介绍。

1.2.1 社交网络影响力

在社交网络中所涉及的主体是人，人在社交网络中的表现多种多样，人们的思想和行为都通过相互之间的影响而发生巨大的变化，同时人又是独立的个体，人与人之间有着很大的差异，所以这就给社交网络影响力的分析带来了挑战，以下文献中给出了社交网络影响力的不同定义。

在文献[1]中提到社交网络的影响力度量的方法可以分为：基于网络的拓扑结构、基于用户的行为和基于交互信息三种度量方式。其中基于网络的拓扑结构是从节点在网络中位置属性的角度，例如节点是否处在网络的中心，或者节点是否在网络中具有类似于桥梁的作用，或者节点是否拥有较多的关注者等；其次，基于用户的行为则是以用户的某种行为对网络中其他用户的影响来说的，例如该用户的行为对整个网络中用户的影响范围，以及影响的传播速度等，影响的范围越大或者影响的传播速度越快该用户的影响力就越高；最后，交互信息的影响力是以某种特定信息内容为前提的影响力评价，主要原因是考虑社交网络中用户对不同的信息的敏感程度不同，信息的传播范围和速度也不同，例如基于某个话题或者某个特定信息的影响力度量。

拓扑结构主要是根据节点在整个网络中所处的关键位置来衡量节点的重要程度，其中可以表示网络的拓扑属性有，节点的出入度描述与节点有直接关系的节点数量，节点的紧密中心度和介数中心度是从节点之间的最短路径的大小来描述节点之间的间接关系，特征向量中心度从网络中节点的地位或声望角度的考虑将节点的声望看成是所有其他节点声望的组合[2]。PageRank 度量[3]是特征向量的一种变型，在特征向量的基础上根据随机游走的思想添加了一个逃脱因子。韩忠明等人[4]在 PageRank 算法的思想上，通过度量节点活跃度（节点的权值）的机制和节点关系强度（边的权值）的加权网络，来分析计算社交网络中节点的影响力。Lü L 等人[5]又在 PageRank 和特征向量的基础上提出 LeaderRank 算法来研究节点的影响力。根据系统的核与核度理论，将删除度量节点（集）后对网络连接的破坏程度来定义其重要性，其主要的研究基础是核与核度理论。Kitsak[6]等人研究了 K-核分解在判断节点传播能力中的应用．K-核的定义是网络中所有度值不小于 k 的节点组成的连通片．属于 K-核同时不属于(K+1)-核的所有节点就是 K-shell 中的节点。

由于在社交网络中的研究中主体是用户，所以研究更关注于用户的信息动态地在网络中传播过程，即从用户行为的角度，研究在不同的信息传播模型下的影响力传播。用户的行为有微博中用户之间的相互关注，信息的转发和评论，以及添加好友关系等，在这些用户的行为中用户之间产生相互影响力，最终以在某种特定条件下用户所能影响的范围来评价用户的影响力。在文献[7]中是以某一特定的动作在一段时间内的传播频率作为评价用户的影响力依据，同时用动作的传播范围作为动作本身的评价依据。在文献[8,9,10,11,12]中认为无论信息是以什么方式，只要在经过一段时间后，同一信息在用户之间进行传播，以网络中所能接受该信息的用户数量，作为该用户的影响力。

总之，在对社交网络的结构、交互信息以及用户行为等特征的量化计算和分析时，可以把用户对信息传播过程或者其他用户行为所产生的影响统一作为用户的社交影响力。本文主要研究的影响力是指，在特定的信息传播的模型下，通过用户之间的相互影响，在经过一定时间后所能影响到的用户数量。

1.2.2 单层网络中的影响力研究

影响力的研究一般是选取网络中有影响力的节点集作为信息传播的触发集，使信息在网络中的传播最大化。目前的相关研究大致分为两类：（1）基于两种信息的传播模型，并且依据信息在单个网络或者多个网络中的传播过程来发掘网络中有影响力的节点，（2）从节点的属性和网络的拓扑结构来发掘网络中有影响力的节点。

Kempe D 等人[8]通过选取网络中包含 k 个用户的触发集合，并通过这个集合作为信息传播的初始集合，不断的把信息传播给相邻的用户，以此使信息能传播给更多的用户。如何去选取最优 k 个用户的触发集合，最优的解决方案被证明是 NP-hard，所以 Kempe D 等人提出一种近似求解的方法-贪婪算法。根据子模块的性质，这种近似算法是可以达到最优解的（1-1/e）的精度[13]。

在大型的网络中，由于贪婪算法的时间效率比较低，所以文献[14]中提出一种基于社团结构的贪婪算法，该算法以动态规划的方式来减少计算。在文献中利用了社团结构的特性，注重社团内部节点之间紧密联系的信息传播，而忽略社团之间节点的稀疏连接关系，将整个网络划分为多个社团，以分而治之的思想来提高时间效率，最后证明该算法能够得到一定比例的近似解。除此之外，也有一些在贪婪算法的基础上，通过启发式的方法来降低时间效率[10]，将网络中不可能作为集合的节点去除，只从一部分节点中选取最优集合，还有一些是通过降低网络计算的规模来提高时间效率[9]。

以上的研究方法中，无论是贪婪算法[8]，还是结合社团结构的贪婪算法[14]在时间效率上都比较低，同时选择所选取的节点在网络中不具有实际的意义。本文中通过结合社团结构与节点的拓扑属性来选取 Top-K 节点集，使得算法不仅在时间效率上接近于线性，同时还在实验效果上具有很高的近似度。

1.2.3 多层网络中的影响力研究

现实生活中的社交网络并不是那些简单的独立网络，而是复杂的异质网络。异质网络包括节点的异质性与边的异质性[15]，节点的异质性是指网络中的节点归属于不同的类别，例如包含用户、日志和位置的网络中，用户和日志之间为原创和转发关系，用户与用户之间为好友关系，日志与位置之间为位置关系；边的异质性则表现在网络中用户的关系在现实生活中所表现不用的实际意义，例如社交网络中用户之间有“好友”、“评论”和“关注”等关系，或者从平台的角度来说，用户之间的关系分别来自与“人人”、“微博”和“QQ”等社交平台。

在研究多层网络时，很多研究是将多层网络中的问题转化为单网络问题来研究。Berlingerio M等人[16]提出了两种通用的多层网络聚合方法：只要两节点之间有任意维的边存在则该边存在，不考虑边的不同属性；网络中两节点之间存在不同属性的边的个数记为聚合后的单层网络中边的权重，只考虑边的个数，没有考虑不同层的网络之间的联系和相互作用。Zhu G等人[17]主要针对多层网络聚合的两个缺点：网络聚合会丢失一部分网络信息；每一个子网络的聚合只考虑节点的邻居节点的信息，不能从全局上反映节点的信息。通过计算节点在两层网络中的相似性，来得到两层网络之间的相似性，最后通过各层之间的相似性系数计算得到每层网络在整个多层网络中的权值比重,将各层的邻接矩阵与每层的权值相结合得到聚合后的网络。

然而，由于异质网络的复杂性，早期对异质网络的研究方法是将异质网络解析成多个同质的单个网络，然后分别对单个网络逐个分析，这种分析方法忽略了每层网络之间的关联关系。在有异质节点的异构网络中，可通过节点之间的路径关系，将网络解析成为节点同质的简单网络。对于多层同质网络，也就是边异质的网络，Nguyen D T等人[11]提出LCI（Least Cost Influence）算法，算法是在线性阈值的传播模型下，在多层网络中添加中间层节点，然后利用这些中间层节点，在不丢失信息的前提下将网络聚合成一层大型网络，最后依据子模块性质的贪婪算法[8]来使得影响力传播最大化。Zhan Q等人[12]通过标签路径将异质网络解析为同质网络，如同user-Tweet-user标签路径。文中定义了两种信息传播的路径，分别是网络内的信息传播和网络间的信息传播。因为主要研究的是用户，根据用户之间的影响力（通过不同的标签路径）的传播路径，将多个类别的节点与边进行了聚合，变成了单层用户与用户之间的关系网络。在信息传播的过程中，分别定义了网络内和网络间的节点相互影响的概率，将网络内的信息传播和网络间的信息传播一致看待，最后利用贪婪算法的思想来选择影响力传播最大的节点集合。

与单层网络中的研究相同，在多层网络中同样是算法时间效率低的问题，因此，本文将在单层网络中关于社团的影响力研究的基础上，结合多层网络中相关的研究方法以及多层网络中的社团发现算法，从网络的拓扑属性的角度，来解决如何高效选取影响力最大化传播节点集。

1.3 论文研究目标和内容

1.3.1 研究目标

为了使得社交网络中信息的最大化传播，本文结合用户在社交网络中的拓扑属性与行为，选择大小为 K 的用户集合，并以这个集合为触发集合，从而使信息在网络中的传播最大。同时，在保证具有相对较高的近似解的前提条件下，利用网络拓扑信息和用户行为的综合分析方法，提升算法的时间效率。

1.3.2 研究内容

本文将从单层网络和多层网络两方面研究影响力，通过结合社团结构与网络拓扑属性来选取 Top-K 节点集，研究单层和多层网络中的相关社团划分算法，社团与网络拓扑属性的结合算法，以及基于相关的信息传播模型研究社交网络的影响力传播。

1）单层网络中的相关研究

在单层网络中主要的研究内容包括，如何结合网络的拓扑属性来评价节点的影响力、网络的信息的传播方式和社团的划分算法，以及如何将社团结构与拓扑属性相结合等几个方面。

(1) 网络的拓扑属性描述了节点在网络中局部或全局的位置，在一定程度上能够表现出节点的信息，与贪婪算法计算出来的节点集合相比，不仅能表现出节点本身在网络中的实际含义，更能在时间效率上有很大的提升。为了提升算法在庞大社交网络中的时间效率，考虑用网络拓扑属性来研究问题，同时结合网络的其它属性来提高实际运算效果。

(2) 网络中信息的传播方式和节点集合的影响力评价标准。在社交网络中，相邻的用户通过相互打分来确定用户在整个网络中的影响力，通过启发式的方法选择触发集合。分别在线性阈值模型和独立级联模型中计算触发集合 A 的影响力函数，然而计算影响力函数的方法各有不同，本文研究的计算方法是以集合 A 为初始激活的节点集，由激活的节点影响未激活节点，通过激活节点集合不停地迭代计算最终的被激活节点的数量作为影响力函数的结果。

(3) 针对 LeaderRank 算法的结果仅受到相邻节点的影响，具有局部片面的缺点，本文提出将网络中的节点分类，即根据网络社团的发现算法将节点标记为不同类别，从而降低这种局部属性对最终结果的影响。由于不同的社团算法也可能对最终选取的具有影响力的 Top-K 节点集合有影响，尝试找出一种合适的社团发现算法，并将其与 LeaderRank 结合以达到更好的实验效果。其中研究工作所涉及的社团发现算法有，改进的标签传播社团划分算法，随机游走社团划分算法，模块度贪婪社团划分算法等多种社团划分算法。

(4) 为了证明单层网络中结合社团结构的影响力算法的可行性，将其与传统的贪婪算法在所选取节点集合的影响力效果上，以及算法的时间效率上相比较。贪婪算法的基本思路是每次

选择对上一时刻算法产生的触发集合增益最大的节点，并将该节点添加到触发节点集中，依次选取 Top-K 的节点集合。

2）多层网络中的相关研究

将网络中的局部和全局属性进行综合，使单层网络社团发现算法和 LeaderRank 算法延伸到多层网络中，继而研究多层网络中的影响力传播问题。利用多层网络中网络融合分析和协同分析两种研究方法，结合异质网络的具体环境，研究如何将上述的方法延伸到异质网络中。主要的研究内容包括：多层网络中的不同社团划分算法，网络的抽样算法以及利用多层网络的几种相关研究方法结合社团结构的多层网络社团影响力算法。

(1) 首先，利用多层网络相关问题的研究方法，将多层网络中的社团划分算法大致分为两种，第一种是将多层网络的问题按照一定的策略转化为单层网络的问题，然后利用单层网络的社团发现算法，具体分为融合后划分和划分后融合两种；第二种是利用目前的方法直接对多层网络进行协同处理。

(2) 其次，就不同的多层网络社团划分算法，结合 LeaderRank 算法，并依据多层网络的研究方法，提出三种不同的结合策略，第一种是将多层网络融合后结合策略，第二种是分层先结合 LeaderRank 算法再融合的策略，最后一种是综合考虑多层网络的结构协同结合策略。

(3) 最后，因为相关的多层社交网络数据集中数据量太大，依据现有的网络抽样算法得到抽样数据集合，再结合多层网络中信息的传播方式，来研究多层网络中基于社团的具有影响力的 Top-K 节点集选取算法。

1.4 论文组织结构

本文共分为五章，各章的主要内容具体如下：

第一章为绪论，主要包含本文研究的背景，国内外相关研究现状，以及本文研究的主要内容和意义；

第二章为相关技术，主要介绍相关研究的基础技术知识，以及相关研究的方法和方向；

第三章是单层网络下基于结合社团结构的影响力分析，主要介绍单层网络中的 CLR 算法的相关研究，以及相关实验结果与分析；

第四章是多层网络下基于结合社团结构的影响力分析，主要介绍单层网络中的 CLR 算法在多层网络下的应用，同时利用多层网络的研究方法形成 SMCLR 算法，以及相关实验的结果与分析；

第五章是对本文相关研究问题的总结和展望。

第二章 相关技术

本章主要介绍社交网络中影响力研究的相关技术，首先介绍求解影响力传播最大化问题经典的近似求解算法即贪婪算法；然后介绍相关文献中对单层网络和多层网络中信息的传播问题的相关技术；最后是介绍多层网络中的基础问题的相关研究方法。

2.1 贪婪算法

在研究影响力传播的相关问题时，无论是在单层网络还是多层网络中都是直接或者间接地利用贪婪算法来求解问题。在文献[8,11]中是直接利用贪婪算法来处理单层和多层网络，而文献[10,12,14]是利用启发式方法，通过不同的方式缩小候选集合的大小以提升时间效率，继而间接地利用贪婪算法来处理单层和多层网络。

从网络中选取大小为 k 的节点集作为触发集合，从触发集合中的节点开始模拟网络中信息的传播过程，使信息在网络中最大化传播，即在信息传播结束时，网络中被激活的节点数最大。这个求最优解问题已经被证明是一个 NP 难问题，在文献[8]中提到通过贪婪的方式来选取节点，使得对当前已经选择的节点集的影响力贡献最大，依次进行直到得到大小为 k 的节点集合。

贪婪算法利用子模块定理的性质，能够得到相对较高的近似值。无论在哪种传播模型下，求解 Top-K 的最优解问题都满足子模块性质，即满足公式：

$$f(S\cup\{v\})-f(S)\geq f(T\cup\{v\})-f(T),\ \mathrm{S}\subseteq\mathrm{T} \tag{2.1}$$

$f(\cdot)$表示以集合 S 作为触发集合的影响力函数，不等式 2.1 表示，当选取同一个节点添加到集合 S 和 T 中作为触发集合时，影响力函数增益值集合 S 至少要比集合 T 大。根据这个性质利用贪婪算法来选择 Top-K 节点集，能够得到相对比较高的近似解。文献[13]中已经证明，假设集合S^*是函数$f(\cdot)$取得最大值的解，同时集合 S 是通过贪婪算法选取的节点集，那么：

$$\begin{aligned} f(S) &\geq (1-(\frac{K-1}{K})^K) f(S^*) \\ &\geq (1-1/e) f(S^*) \end{aligned} \tag{2.2}$$

不等式 2.2 可以说明当所选取的节点集合大小 K 趋向于无穷大时，贪婪算法至少可以达到最优结果的 63%以上[13]，而当选取的节点集合大小 K 越小时，贪婪算法所取得的近似比例越大，当 K 为 1 时，贪婪算法得到的结果就是最优解。

2.2 信息传播

在独立的网络中信息的传播有多种方式和模型，而在多层网络中信息的传播变得更加复杂，因为多层网络中又存在每层网络之间的信息传播路径，以下将从单层网络与多层网络两种环境下分别介绍信息的传播方式。

2.2.1 单层网络中的信息传播

信息的传播方式表示，在独立的网络中信息在传播的过程中节点的状态转变过程，节点的信息传播方式有 IA (Inactive -Active)[8,9,10]，SIS(Susceptible-Infected-Susceptible)[18,19]，SIR(Susceptible-Infected-Recovered)[19,46]等。而信息的传播模型表示，在满足什么样的条件下，节点才能从上一个状态转变到下一个状态，传播模型大致分为线性阈值模型和独立级联模型两种。

将社交网络中信息的传播过程类比于疾病传播过程，不同的疾病在人群中传播感染的方式不同。有的疾病在被感染后，一旦被感染的人被治愈，在以后的一生中都不会再次被该疾病感染，也就是 SIR 传播方式。而有的疾病在感染后，即使被治愈，在以后也会再次感染该疾病，即为 SIS 传播方式。在社交网络中的信息传播过程中，网络中的用户也有很多的状态，例如对于微博上的博文来说，用户就有已读状态，未读状态，评论状态，转发状态等。本文中为了简化社交网络中信息传播方式，只考虑两种状态，已知状态和未知状态，也就是 IA 传播方式。在网络中节点只有两种状态，即 A（Active）和 I（Inactive），初始时节点都处在 I 状态，节点的状态转移过程只倾向于从 I 变成 A，反之则不成立。

线性阈值模型。在该模型中，当节点的邻居节点有较多处在 A 状态时，那么该节点将会更容易从 I 状态转变为 A 状态。初始情况时，网络中的节点全部处在 I 状态下，并且为每个节点随机分配一个阈值θ_v，当节点 v 的邻居节点中处于状态 A 的节点与节点 v 的边的权值之和大于θ_v时，则节点 v 从 I 状态转变成 A。那么从实际意义上来说阈值θ_v表示，节点 v 能从 I 变成 A 的一个临界阈值，公式表示为：

$$\sum_{u \in N_v,\, S_u = A} w_{uv} \geq \theta_v \qquad 2.3$$

其中N_v表示节点 v 的邻居节点集合，S_u表示节点 u 的当前状态。当节点 v 满足不等式时，将节点 v 的状态置为 A，即 $S_v = A$.

独立级联模型。与线性阈值模型相比不同之处在于，用户之间的信息传播是相互独立的过程，在状态转变的过程中只与两节点之间的边权值有关，与其它节点无关。在某一时刻 t 节点 v 由状态 I 转变为状态 A，那么在下一个时刻，该节点 v 有且仅有一次机会去激活它处于 I 状态的邻居节点 u，激活的概率大小由他们之间边上的权值决定，权值越大，概率越大。若用户 u 在 t+1 时刻被激活，变成 A 状态，继而会以同样的方式激活它的邻居节点。若用户 u 在 t+1 时刻未被激活那么在以后的时刻该用户也不会被用户 v 激活，也就是说在独立级联模型下，节点之间的边有且仅被利用一次。

在文献[8]中已经证明上述的两种传播模型在求解本问题时，可以相互之间进行转化，所以在本文中只考虑其中的一种模型，即独立级联模型。在独立级联模型中，我们根据网络中边的权值来决定节点的状态转变的概率，给定触发集合 S，以集合中的元素作为初始的活跃节点，经过有限次数的状态转移以后，整个网络中处于 A 状态下的节点个数是一个稳定的数值，并依据这个数值作为评价这个触发集合 S 优劣的标准。因为所采用的是独立级联的传播模型，所以在计算最终结果的时候会存在一定的随机性，所以在本文中采用多次计算的平均值作为最后的结果。

2.2.2 多层网络中的信息传播

相比单层网络中的传播过程来说，多层网络要更为复杂。在文献[20]对多层网络中信息的动态传播研究中，分别定义了同层网络与不同层网络中节点的传播系数，然后通过设置这两种大小不同的传播系数来对多层网络进行研究。其研究结果表明，网络中具有多层性的网络结构能够加速信息的传播，同时得出结论，网络层之间的信息传播速度比每层网络中的信息传播速度快得多。

在多层网络中信息的传播模型也分有独立级联模型和线性阀值模型，与单层网络不同之处在于，在多层网络中网络层之间节点还存在信息的传播路径，由于网络层之间的节点的传播速度要高于同层网络中节点的传播速度，因此定义了以下两种层间节点传播：1）第一种如文献[11]中的描述，多层网络中锚节点之间的信息传播概率为 1，当时刻 t 时有一层中节点的状态被激活后，同一时刻其它层网络中对应的锚节点也变成激活状态；2）第二种如文献[12,19]中描述，多层网络中都利用 jaccard 相似系数来表示网络层间节点之间的传播速度，然后将存在的这两种传播方式统一对待处理。

多层网络中信息的传播过程如图 2-1 所示，图中描述在多层网络中少数几个节点同时存在不同的网络中，图中从左到右表示从最初的状态到传播结束时，网络中多个节点被激活的过程。从图中可以看出信息的传播在两个网络中穿插进行，信息从网络 A 传播到网络 B 中，最后又回到 A 网络，这种信息的传播过程在一定程度上说明了多层网络中信息的传播的复杂性和网络层间信息传播在整个信息的传播过程中的重要性。

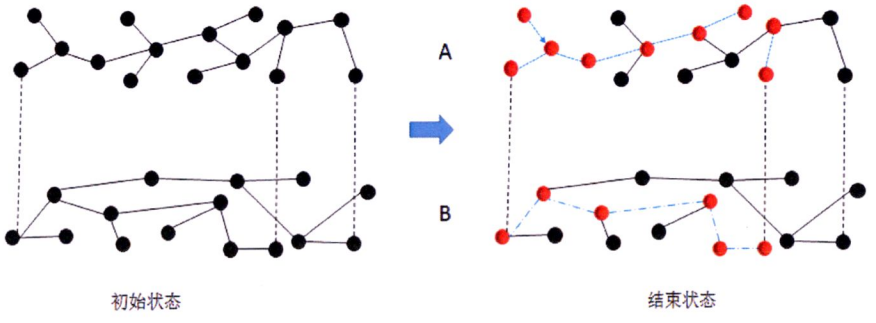


图 2-1 多层网络中信息的传播过程

如果将处于 A（Active）状态的节点激活该节点相邻节点中处于 I（Inactive）状态的节点这一过程所需要的时间记录为一个时间周期，并且经过 T 周期后网络中已经没有节点可以被激活，这时将网络中处于 A 状态的节点个数记作最终的影响力。在小型网络中计算节点集合的影响力时，可以完全等待信息传播 T 周期后统计网络的影响力，然而在大型的网络中，网络的平均半径较大，因此所需要的传播周期 T 也较大，如果通过信息完整的传播路径来计算节点的影响力将会十分缓慢。所以在文献[11]中提出在一定的传播周期后，网络中处于 A 状态的节点个数作为节点集的影响力。与原来的影响力计算方式不同的是，这里只添加了时间周期的因素，那么这种影响力的评价方式不仅可以在一定程度上体现 Top-K 节点集的影响力，同时也体现了 Top-K 节点集的影响力传播速度。

多层网络中层间节点的信息传播方式采用的是文献[11]中的层间传播，即当时刻 t 时有一层网络中节点的状态被激活后，同一时刻其它层网络中该节点所对应的锚节点也变成激活状态，同时不仅采用了整个传播周期 T 的网络的影响力的评价方法，也采用在规定的传播周期后网络的影响力评价方法。多层网络同样也采用的是独立级联的传播模型，同时为了保证结果的准确性，也采取多次计算求平均值的方式。

2.3 多层网络研究基础

在复杂的异质网络中存在两种异质性，一种是节点的异质性，一种是边的异质性。异质节点表示节点在网络中代表不同类别的个体，而边的异质表示节点之间的不同的关系，边的异质性不仅体现在异质节点之间，也体现在同质节点之间。例如在文献信息网络中，包括多种类型的节点，如作者、文献、会议或期刊，多种类型的关系，比如作者与文献之间的发表，文献与文献之间的引用关系，作者与作者之间的合作关系等。在社交网络中，因为主要的研究对象是人，所以研究的节点为同质，但是人与人之间的关系比较复杂，所以研究的重点是网络中的异质边，因此本文中所讨论的多层网络是指具有同质节点的异质关系的多层网络。

2.3.1 多层网络相关定义

多层网络的本质是多个简单网络的集合[21]，其中每一个简单网络都表示着节点间的某一种关系[22]。Cantador 等人[23]将多层网络中节点之间的不同关系定义为复边，在此基础之上，Newman[24]将多层网络定义为两个节点之间存在不止一种关系的网络。在对多层网络进行数学描述时，现在的研究一般采用向量定义、图论定义以及超矩阵定义等三种定义方法。

定义 2.1（多层网络的向量定义）在多层网络中，由于节点之间存在不同维度的关系，而这些关系可以通过不同的网络来表示，因此多层网络的实质可看成是在 N 个节点和 m 种不同关系下，所形成的不同网络图的集合。其中，多层网络下的网络图集合可用 $\{g_1, g_2, \ldots, g_m\}$ 表示，对于每一种网

络图 g_k 可采用图论中的邻接矩阵$A^k = \{a_{ij}^k\}$ 的形式来表示，且 a_{ij}^k 代表节点 i 和节点 j 在第 k 层网络中存在的边。因此一个多层网络就可以用向量 $A = [A^1, A^2, \ldots A^m]$ 表示。

定义 2.2（多层网络的图论定义）文献[25,31]中给出了多层网络的图论定义方式：$G = \{V, E\}$，其中 G 代表多层网络，V 代表多层网络的节点集合；L 代表组成多层网络的每个简单网络的集合，$E = \langle x, y, l \rangle$ 代表多层网络的复边，$x, y \in V$ 为多层网络中的节点，$l \in L$为该多层网络下的某一个简单网络。在图论的多层网络定义中，节点的个数为 N 表示总共的节点数，而以复边的方式来描述节点之间的多层关系。

定义 2.3（多层网络的超矩阵定义法）类比单层网络的邻接矩阵定义方式，文献[26]引入超级邻接矩阵（Supra-adjacency Matrix）的概念，试图采用一个矩阵来对多层网络的各网络内部节点间的连接关系以及网络间同名节点的连接关系进行描述。同邻接矩阵一样，在同一个网络中若节点之间有边相连则令该项为 1，否则为 0；在不同网络之间，只有当处于两层网络中的两节点表示为同一个节点时，两节点之间才会存在连边，并且连接两节点之间的边权重也有几种不同的度量方式。

$$w_{i,kl} = \frac{\left|N_{i,k} \cap N_{i,l}\right|}{\left|N_{i,k} \cup N_{i,l}\right|} \tag{2.4}$$

$$\mathrm{W}_{i,j} = diag(w_{1,ij}, w_{2,ij}, \cdots, w_{n,ij},) \tag{2.5}$$

$$\mathrm{M}=\begin{pmatrix} A_1 & W_{1,2} & \cdots & W_{1,L} \\ W_{2,1} & A_2 & \cdots & W_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ W_{L,1} & W_{L,2} & \cdots & A_L \end{pmatrix}_{NL \times NL} \tag{2.6}$$

如公式 2.4 所示，其中 $N_{i,k}$ 表示节点 i 在第 k 层网络中的邻居节点集合，同时对边的权重进行归一化处理。在定义了网络间节点之间边的权值的基础之上，公式 2.6 给出了超邻接矩阵的表示，其中超邻接矩阵的主对角线上的元素A_L表示第 L 层网络的邻接矩阵，而非主对角线上的元素$W_{i,j}$如公式 2.5 中所示的对角矩阵，表示第 i 层网络与第 j 网络之间节点的关系。

2.3.2 多层网络基本问题研究策略

多层网络分析研究方法的发展实质上是对多个网络处理方法的变化和发展，多层网络中相关问题的研究主要分为三个研究方向：1）多层网络下单个网络逐一分析；2）多层网络融合成简单的单层网络分析；3）多层网络协同分析。

早期的多层网络研究方法是对单个网络的逐个分析。这种分析方法主要存在以下几种缺点，首先，忽略了每层网络之间的相互关联关系，将节点之间的多种关系独立地处理；其次，将多种节点关系统一对待，认为多层网络中每层网络中的权重相同，但事实上并不相同，在文献[27]中就认为不同层网络在整体的多层网络中所占权重不相同，权重大的网络会影响权重小的网络；最后，当信息

在网络中传播的时候，这种逐个分析的方法忽略了信息在不同节点关系中的传播，例如在社交网络中，用户的微博信息可能会转发到该用户微信朋友圈中。

$$a_{ij}=\begin{cases}1 & if\ \exists k: a_{ij}^{k}=1\\ 0 & \text{otherwise}\end{cases} \tag{2.7}$$

$$a_{ij}=\sum_{k=1} a_{ij}^{k} \tag{2.8}$$

随后还有将所有单个网络进行聚合，生成新的聚合后的简单网络。在文献[28]中将多层网络中的边整合为单层，并提出了三种聚合的方法：1）只要两节点之间有任意维的边存在则该边存在，不考虑边的不同属性，如公式 2.7 所示；2）网络中两节点之间存在不同属性的边的个数，记为聚合后单层网络中边的权重，如公式 2.8 所示；3）根据具体的研究问题，将网络中关键属性抽取出来进行聚合，在文献[29]中主要针对多层网络聚合的两个缺点：网络聚合丢失一部分网络信息与每一个子网络的聚合只考虑节点的邻居节点的信息，不能完全反应节点的信息。通过计算节点在两层网络中的相似性，如公式 2.10 所示，其中$C_{i,sr}$表示节点 i 在网络 s 与网络 r 中的相似度，a_{ij}^{r}表示 r 层网络中节点 i 与节点 j 的关系，当两个节点之间有边则为 1，否则为 0，公式如下：

$$C_{i,sr}=\frac{\sum_{j} a_{ij}^{s}a_{ij}^{r}}{\sum_{j} a_{ij}^{s}+\sum_{j} a_{ij}^{r}-\sum_{j} a_{ij}^{s}a_{ij}^{r}} \tag{2.10}$$

由此，通过同一个节点在不同网络中的相似度，来得到两层网络之间的相似性，如公式 2.11 所示，其中 n 表示多层网络中不同节点的总数，公式如下：

$$C_{sr}=\frac{1}{n}\sum_{i} C_{i,sr} \tag{2.11}$$

然后，通过各层之间的相似性系数计算得到各层网络在整个多层网络中的权值比重，如公式 2.12 所示：

$$I_{s}=\frac{\sum_{r} C_{rs}}{\sum_{s}\sum_{r} C_{rs}} \tag{2.12}$$

最后，将各层的邻接矩阵与每层的权值结合得到聚合后的网络中，如公式 2.13 所示，其中A^{i}表示第 i 层网络的邻接矩阵，m 表示多层网络中网络的层数，公式如下：

$$L=\sum_{i=1}^{m} I_{i}A^{(i)} \tag{2.13}$$

通过定义多层网络中节点的相似度，来进行多层网络中节点相似度的聚合，然后研究多层网络中社团划分的问题。

相比单层网络逐个分析方法来说，将网络聚合的后的网络虽然能够得到原先不存在的信息通路，

能够相应地体现信息的传播，但是这种聚合的方法仍然存在以下几个缺点：1）这种聚合的方法仍然没有解决节点之间连接关系的独立性问题，忽略了两个节点之间可以同时存在多条边的事实；2）多层网络聚合时，将节点之间的关系看成同等关系，但事实上并不相同，不能将边的异质性体现出来。

多层网络的协同分析，在综合考虑多层网络中每种节点关系后，需要考虑具体的适用场景和应用环境，计算出能够描述多层网络多种关系的特征属性，并以此对所研究的问题进行协同分析。在文献中就提出基于多层网络中各层网络的不同拓扑特性，构造能够描述全部网络特性的超级矩阵，利用随机游走的思想对目标网络进行协同分析。

2.4 本章小结

本章主要介绍了相关研究的技术背景，首先介绍了贪婪算法，该算法是选取影响力 Top-K 节点集的近似算法；然后说明了信息的传播方式和模型，分别从单层网络和多层网络中介绍信息的传播；最后，介绍了现今多层网络中基本问题的研究方向，多层网络的几种定义以及各种研究方法的优缺点。

第三章 单层网络中基于社团结构的影响力分析

本章主要介绍在单层网络中利用社团的结构属性来选取影响力最大的 Top-K 节点集合，分别从以下几个方面展开：网络的拓扑属性，包括网络的基本属性，PageRank 与 LeaderRank 属性来衡量节点的影响力；相关的社团发现算法，描述了几种经典高效的社团划分算法；研究提出结合网络拓扑属性与社团特性的 Community-based-LeaderRank 算法；最后在不同的数据集上对比经典的贪婪算法，从实验效果和效率上证明本算法的可行性。

3.1 网络拓扑属性

通过分析网络中节点的拓扑属性来研究节点的影响力，分别从网络中的基本属性和相关的 PageRank 算法与 LeaderRank 算法来分析节点的影响力。

3.1.1 基本属性

网络节点的基本属性度量包括入度、出度、紧密中心度(closeness)、介数中心度(betweenness)、特征向量中心度(eigenvector)、PageRank[3]等。这些度量可以表示节点在网络拓扑结构中的局部和全局属性，相关属性定义如下表：

属性名	定义	描述
出度	$d_i^{out} = \sum_{j=1}^{n} A_{ij}$	A 表示网络的邻接矩阵
入度	$d_i^{in} = \sum_{j=1}^{n} A_{ji}$	同上
介数中心度	$BC_i = \sum_{i \neq j \neq r \in V} \frac{k_{jr}}{k_{jr}(i)}$	k_{jr}表示从节点 i 到 r 的最短路径个数，$k_{jr}(i)$表示经过节点 i 的最短路径个数。
紧密中心度	$CC_i = \frac{N-1}{\sum_{j=1}^{N} l_{ij}}$	l_{ij}表示节点 i 到 j 最短路径长度
PageRank	$PR_i = \alpha \sum_{j} A_{ij} \frac{x_j}{g_j} + (1-\alpha)\frac{1}{N}$	x_j表示节点 i 的邻居节点 j 的 PageRank 值，g_j表示 j 的出度。
特征向量中心度	$\lambda x_i = \sum_{j=1}^{n} \alpha_{ij} x_j$	λ表示邻接矩阵最大特征值

表 3.1 节点度量方法

节点的出度和入度表示该节点对其邻居节点的直接影响，当一个节点的出度比较大时，说明在网络中当前节点的追随者比较多，那么其所能直接影响的节点也比较多，具有一定的直接影响力。但是仅仅考虑了对邻居的影响力，忽略了节点在网络中的位置和对网络中其他节点的间接影响力，因此只能反应节点在网络中的局部特征。紧密中心度描述节点到网络中其他节点的最短路径之和的

倒数，其所表现的物理意义是紧密中心度值越大，节点到其他节点的平均距离最短，那么节点可能处在网络相对比较中心的位置。介数中心度表示经过节点的最短路径的数量，因此它刻画了在信息的传播过程中节点的繁忙程度。但实际上，信息在传播的过程中也不是只通过最短路径传播，所以只能在一定程度上描述节点的在信息传播时的桥梁作用。

除此之外，特征向量中心度从网络中节点的地位和声望角度的考虑，将节点的声望看成是所有其他节点声望的组合；还有根据系统的核与核度理论，将删除小于 K 度的节点后，利用删除该节点对网络连接的破坏程度来衡量节点重要程度的依据，同时 K-核分解在判断节点的传播能力中也有应用．K-核的定义是网络中所有度值不小于 k 的节点组成的连通片．属于 K-核又不属于(K+1)-核的所有节点就是 K-shell 中的节点。在信息的传播应用中，K-核能够得到关系特别紧密的节点结合，但是忽略了网络中绝大多数度小于 K 的节点在信息传播中的作用。

综上所述，基于节点属性的影响力分析的方法各有不同，不同的属性描述的侧重点也不同，所以得到节点影响力也不同。有一种多属性决策的节点重要性评估方法：TOPSIS（Technique for Order Preference by Similarity to an Ideal Solution）[20,21,32]，将不同方法所计算得到的节点影响力指标作为节点的一个属性，然后评估各个属性之间的相对重要程度，以此建立多属性的决策矩阵。

3.1.2 PageRank 与 LeaderRank 算法

Google 著名的网页排名算法 PageRank，根据网页的 PageRank 值的大小对网页的重要性进行排序。算法的大致思想来自于用户浏览网页操作，用户在浏览网页时，会根据网页中的链接不停地在网页之间跳转，在网页跳转的过程中会根据上一个网页的 PageRank 来给下一个网页打分，这样在这样网页不停的迭代，每个网页最终会得到一个稳定的值，这个值越大说明用户浏览的次数越多，相对网页的重要性就越大。

利用 PageRank 排名算法来研究社交网络影响力时，具体情况略微不同。在有向网络中，PageRank 算法侧重于表现网页链接所指向的网页，即其它大部分网页直接或者间接链接指向的网页，相反，而从社交网络信息传播的角度来看，网络中有相当一部分节点的信息来源与一个节点时，该节点在信息传播过程中表现的作用就越大，从而相应的节点评分也就越大。

与 PageRank 相似 LeaderRank 算法也是相邻的节点之间进行打分，经过不停迭代分值较大节点的重要性越大。但是不同之处在于，PageRank 算法在相邻节点互相打分的同时，需要添加一个逃脱因子，主要是因为在网络中会存在节点的入度为 0 的节点，再下一个迭代后，由于没有相邻的节点对其打分，所以他的值为 0；网络中也同样存在节点的出度为 0 的节点，那么在下一个迭代后，该节点的邻居节点会依据自己的分值对其重新打分，节点在上一个迭代得到的结果将被替换掉，这样整个网络中节点的分值就会存在丢失。当经过 n 次迭代后，最后的节点的 PageRank 值会越来越下，

直到最为都为 0.

在 LeaderRank 算法中，在所有网络中添加一个 Leader 节点，并且在该 Leader 节点与其他所有的节点之间添加双向的连接，这样就不需要像 PageRank 那样添加逃脱因子。其好处在于，在相同的迭代过程中，不需要调整跳脱因子的参数，因此所得到的结果表现出网络真实准确的拓扑属性。

3.2 相关的社团发现算法

在现实世界复杂的网络中，存在一群相互之间有较大的相似性的节点，这组节点之间的关系比较紧密，而与其它组节点之间的关系比较稀疏，这样的一组节点称为社团。寻找社团结构并对其进行研究和分析，对研究网络的各种组织结构具有重要的意义和价值，社团结构的研究在生物学，计算机科学以及社会学等领域都有管饭的应用。在信息的传播领域中，由于社团内部节点相似性比较高和节点之间的联系比较紧密，因此信息在社团内部的传播速度和范围要远远超过社团外。为了研究社交网络中的影响力传播问题，以下介绍了几种具有代表性的几种社团发现算法。

现今的社团发现算法大都是通过定义节点之间的相似距离来划分社团，在划分时有的算法是依据社团的模块度系数来决定社团的最终划分结果，还有的算法是选择使模块度最大的划分方式贪婪地进行节点的社团划分。对社团算法的研究目的是发现网络中具有影响力的节点集合，为了方便节点的影响力分析，在本章节中所介绍的社团划分算法都为非重叠社团划分算法，即所发现的社团结构不会出现节点重叠的现象，也就是一个节点只能属于一个社团。

3.2.1 社团的相关定义

社团发现算法致力于发现网络中联系相对紧密的节点集，表现这组节点集的聚集程度的相关参数大致分为两种，一种是相对局部的聚集度，另一种是全局的聚集度。

社团模块度[33]是描述社团特性强弱的重要指标，作为社团划分算法优劣的评价标准，其定义如下：

$$Q=\frac{1}{2m}\sum_{ij}\left(A_{ij}-\frac{k_i k_j}{2m}\right)\delta(C_i,C_j) \qquad 3.1$$

其中，k 表示节点的度值，m 表示网络中边的个数，而函数 $\delta(C_i,C_j)$ 表示节点 i，j 如果在同一个社团则为 1，否则为 0. 在实际意义上公式描述的事实是，网络中连接社区结构内部顶点的边所占的比例，与这样的社团结构下网络中任意连接这两个节点的比例的差值的期望。这个值的范围在 0~1 之间，越接近 1 说明社团划分的效果越好。

社团聚集系数[34]表现的是社团的局部聚集参数，它描述的是节点与它的邻居节点所形成的三角形的个数与该节点所有可能形成的所有三角形个数的比值，在一定程度上描述了节点的邻居节点之

间的连接紧密性。

3.2.2 随机游走算法

文献[35]中提出随机游走的思想来计算网络中节点之间的相似性距离，并以此距离进行社团划分。在给定的网络中进行随机游走的过程是，当时刻 t 时，从节点 i 开始随机地游走到该节点众多邻居节点中的一个节点 j，然后下一个 t+1 时刻，再从节点 j 开始继续在网络中随机游走的过程。在游走的过程中构建一个节点的概率转移矩阵Pro^t，其中的元素Pro_{ij}^t表示节点 i 在经过 t 次随机游走后到达节点 j 的概率。

当节点 i 和 j 属于同一个社团时，他们之间的随机游走概率Pro_{ij}^t的值将会比较大，但是反过来当随机游走概率Pro_{ij}^t的值比较大时，两个节点不一定属于同一个社团。当两个节点属于同一个社团时，由于社团之间的关系比较紧密，它们在经过多次随机游走后会很大可能地困在当前的社团结构中，那么以此判断它们到网络中其它节点的路径会趋于一致，因此定义了两节点之间的相似性距离γ_{ij}如下：

$$\gamma_{ij}(t)=\sqrt{\sum_{k=1}^{n}\frac{(Pro_{ik}^t-Pro_{jk}^t)^2}{d(k)}}=D^{-\frac{1}{2}}Pro_{i\bullet}^t-D^{-\frac{1}{2}}Pro_{j\bullet}^t \qquad 3.2$$

其中 d(k)表示节点 k 的度。为了方便将节点进行聚类，同样，通过节点间的随机游走概率Pro_{ij}^t，可以得到社团到节点的随机游走概率$Pro_{C\bullet}^t$，因此可定义社团与社团的距离$\gamma_{C_1C_2}$如下：

$$\gamma_{C_1C_2}(t)=\sqrt{\sum_{k=1}^{n}\frac{(Pro_{C_1k}^t-Pro_{C_2k}^t)^2}{d(k)}} \qquad 3.3$$

定义过节点之间距离和社团之间的距离以后，接下来就是将社团进行聚集的问题，聚集时采用 Ward[36]方法来定义距离。社团聚集的初始状态为$P_0=\{\{v\},v\in V\}$，表示在初始时每个节点都是独立的社团，采取从底向上的方式进行聚合，依次选取最优的合并方式，即选择合并后差异变化最小的两个社团将其合并。经过 m 次合并后将会产生 m 个不同的划分策略，那么选择哪种划分作为最终的社团划分结果，一般来说，会根据社团划分的质量评估标准模块度来确定，在这里采取另外的策略，即在进行合并的过程中，如果选取的最优的两个社团的合并前后 Ward 距离的差异较大时算法停止。

3.2.3 标签传播算法

与随机游走社团发现算法相同，标签传播算法[37]同样是考虑节点之间的相似性，不同之处在于它更加侧重于节点的邻居节点相似性的比较，因此整个网络中只表现出节点之间相对局部的拓扑信息。标签传播的算法思想是，将网络中的每个节点都标记上一个标签，这个标签标识当前节点所属的社团，网络中的每个节点都会从它的邻居节点中选择一个与之连接最紧密的节点，使其标签保持一致。那么随着标签在节点之间中的传播，连接比较稠密的一组节点会趋向于拥有相同的标签，直

到网络中没有需要改变标签的节点为止，然后将拥有相同标签的节点划分为一个社团结构。

本文根据信息的传播方式，定义了一种新的方法来选取节点的最近的邻居节点。当节点的多个邻居节点都趋向于同一个标签 L 时，那么当前节点的标签被修改为 L 的概率越大，因为本章中所研究的网络为有向，带权值网络，节点之间边的权值表示的是节点之间相互影响的概率，所以我们的将标签传播相似性度量为：

$$S_i(l) = 1 - \prod_{j \in N_i, L(j)=l} 1 - w_{ij} \tag{3.4}$$

公式3.4中 N_i 表示节点i的邻居节点集合，L(j) = l表示节点j的当前标签为l，公式中的 $S_i(l)$ 表示节点 i 上的标签与标签 l 的相似度。在标签传播的过程中，节点 i 的标签将不停地迭代和更新，节点 i 的标签为 $L(i) = \arg\max S_i$，直到网络中每个节点的标签值都不需要更正为止。

3.2.4 其他社团划分算法以及对比

除了上述的几种社团划分算法以外，还存在其它的社团划分算法。例如基于边介数的社团发现算法[38]，该算法认为社团与社团之间的连接相比社团内部节点之间的连接比较稀疏，那么在整个网络中如果连个节点之间的边介数比较大，说明通过这两个节点的节点对数比较多，那么这条边就应该是属于社团之间稀疏的边，那么这两个节点应该属于不同的社团，从而进行社团划分。在文献[39]提出的基于模块度最大化的贪婪算法，算法大致分成连个阶段，第一个阶段是把每个节点划分到该节点的邻近社团中，使社团的模块度不断增大，第二个阶段将社团替换为一个超级节点，然后根据第一步的社团结构重新构造网络。两个过程不停的迭代，直到网络结构不在变化为止。

以上几种社团划分算法各有优缺点。随机游走社团划分算法从整个网络的拓扑结构上来计算两个节点之间的相似性，划分出来的社团更加精确，但是计算的时间复杂度太高为$O(n^2 \log n)$，其中 n 表示网络中节点的个数，在大型的社交网络中算法的计算量太大；标签传播社团划分算法，在本文中利用信息传播的思想，将节点如何选取相似度最高的标签进行改进，并且算法时间复杂度为接近于线性为O(m)，其中 m 代表网络中边的个数；基于边介数的社团划分算法和基于最大模块度的贪婪社团划分算法都已经比较成熟，但是在研究社团的信息传播方面却不太适用。比如边介数的社团划分算法中，虽然边介数能够反映信息经过该边的繁忙程度，但是在实际的信息传播中社团之间的信息量却很小，同时边介数的计算量很大，算法的时间复杂度也很高，不适合大型网络的社团划分；基于最大模块度贪婪的社团划分算法，虽然算法的时间复杂度只有O(n)，算法的时间效率比较高，但是因为是基于模块度的划分，模块度系数本身也还存在内部缺陷[40]，也就是社团划分的分辨率低的问题，在不能发现很多实际存在的社团，会遗失网络的一些小社团。

3.3 基于社团与 LeaderRank 的影响力分析

在求解选取 top-K 节点集使影响力传播最大化问题中，计算全局最优解被证明是一个 NP-Hard 问题，文献[8]中所提到的贪婪算法可以计算出具有全局最优解$1-{}^{1}/_{e}$比例的近似解，但是在超大型的网络中，如果使用贪婪算法，每次选取当前最优节点时都需要遍历全部节点，然后分别计算每个节点的影响力函数，这样的计算效率就很低。

如果通过计算网络的拓扑属性来选取具有较大影响力的 top-k 节点集。但是网络的拓扑属性只能在一定程度上描述节点在信息传播中的重要性。其中效果相对比较好的 PageRank 属性虽然也能体现出节点的全局属性，但是从实际的算法上来讲，节点的 PageRank 数值主要受它的邻居节点影响，所以本文提出利用社团结构来发现网络中传播信息能力较强的节点。信息在网络中的传播过程是从当前节点传递到该节点的邻居节点，然而在社团结构中节点相互联系比较紧密，信息在社团节点中的传播速度和范围比较大，并且在研究庞大网络的时候，将其拆分为一个个相对较小的社团可以将所研究的问题分而治之。因此，研究社团结构对节点的影响力传播分析具有重要意义。

3.3.1 Community-LeaderRank（CLR）算法

社团结构在社交网络中表现的尤为明显，在社会关系中关系紧密的朋友之间往往组成连接紧密的团体，在团体内用户的连接相对稠密，而在团体之间用户的连接相对稀疏，正是由于社团的这一特性，社团内的信息传播相对比较快，而社团间的信息传播相对比较慢，并且不同的社团结构对信息传播的影响程度不同。

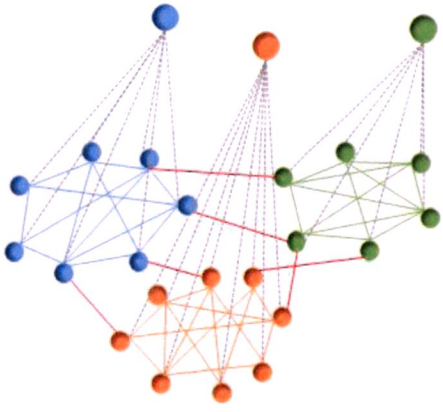


图 3-1 结合社团结构的 Leader 网络

在图 3-1 中描述了结合社团结构的 Leader 网络构建后的网络拓扑示例，图中节点之间的实线表示网络中单向关系，而虚线则表示节点之间的双向关系，在图中具有相同颜色的节点表示属于同一个社团中的节点，在网络上层的 3 个节点表示在社团中添加的 Leader 节点。在这个有向的网络中，根据 PageRank 的算法思想，如果网络中较多的节点都受到同一个节点的直接或间接的影响，那么这个节点将会成为信息传播中相对重要的节点，而 LeaderRank 算法是对 PageRank 算法的变形，因此

将 LeaderRank 算法与社团结构相结合不仅能够表现社团内节点的紧密性，也能在拓扑属性上表现出节点的特征。

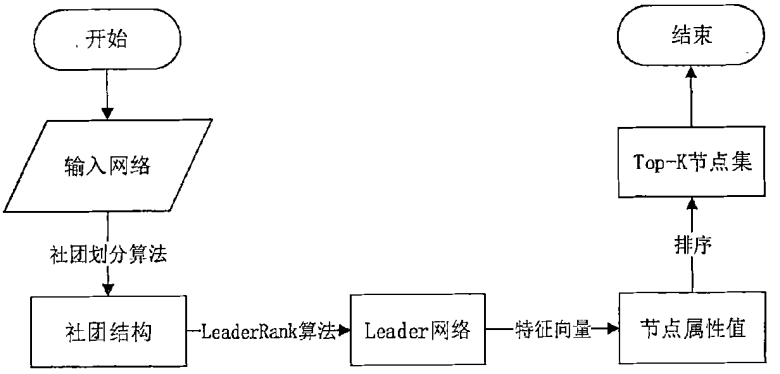


图 3-2 Community-LeaderRank 算法流程

结合社团结构的 CLR 算法的核心是，利用 LeaderRank 算法在每个社团结构中都添加一个 Leader 节点，使得本来就连接紧密的社团内节点变得更加紧密，于此同时也能保证在计算节点的特征时能够更加准确。再利用 LeaderRank 算法思想迭代生成节点的特征时，不仅能体现出节点的整体拓扑特性，也能体现出节点所在的社团特性，有利于研究信息的传播过程中影响力的传播。在图 3-2 中给出 CLR 算法流程，其中对 Top-K 节点集的选取有影响的两个步骤是社团划分算法和节点值排序，因为对于 LeaderRank 算法而言，在给定的社团划分结果下，最终得到的节点属性值为确定的值。关键的两步流程接下来会详细介绍，关于算法的具体步骤如算法 3-1 所示：

输入： 网络$\mathcal{G}=(\mathcal{V},\mathcal{E},\mathcal{W})$
输出： 节点集合$\mathcal{V}$权值列表

1. $\mathcal{C} \leftarrow$ 社团发现算法$(\mathcal{G})$
2. for k = 1 to $|\mathcal{C}|$ do
3. $\quad \mathcal{V} \leftarrow \mathcal{V} \cup g_k$
4. $\quad \mathcal{E} \leftarrow \mathcal{E} \cup \{(g_k, v_i), (v_i, g_k) | v_i \in \mathcal{C}_k\}$
5. end for
6. $f^1 \leftarrow \sum_{i=1} f_i^1$ and $f_i^1 \leftarrow \frac{1}{|\mathcal{V}|}$
7. while $\Delta d \neq 0$ do
8. $\quad$ for n = 1 to $|\mathcal{V}|$ do
9. $\quad\quad f_n^{t+1} \leftarrow \sum_{j=1}^{|\mathcal{V}|} \frac{w_{jn}}{k_j^{out}} f_n^t$
10. $\quad$ end for
11. $\quad \Delta d \leftarrow \sum_{i=1}^{|\mathcal{V}|} f_i^{t+1} - f_i^t$
12. $\quad t \leftarrow t+1$
13. end while
14. $f_i = f_i^t + \frac{f_{g_k}^t}{|\mathcal{C}_k|}$
15. return f^t

算法 3-1 CLR 算法

在给定的网络中，假设网络中总共含有 n 个节点和 m 条边，通过给定的社团发现算法划分出 k

个社团 $\{C_1, C_2 \ldots C_k\}$，接着在网络中添加 k 个 Community-Leader 节点 $g_1, g_2 \cdots g_k$，然后分别在节点g_l与社团C_l中各节点之间添加双向连接。添加节点和边后，在原有网络的基础上构建了一个新的网络，其中一共包含 n+k 个节点和 m+2n 条边。随后计算各节点的 Rank 值时，首先初始化令$f_i^1 = 1/N$，表示在时间 t=1 时，各节点的 Rank 值为网络中节点总数的倒数，N 表示新网络中节点总数。在求解时，由于在网络中添加 2n 条边的缘故，网络中不存在节点出度或入度为 0 的情况，因此新建的网络的邻接矩阵的行列式值不为 0，所以其有唯一的特征向量和特征值，计算得出节点的 Rank 值 f_i^t 对于任何一个节点都将趋向于一个稳定的唯一值。

对大型网络进行 CLR 算法计算时，由于网络的邻接矩阵有唯一的特征向量，因此算法能够在线性的时间复杂度 $O(n)$ 内得出各节点的特征属性值，相比贪婪算法的 $O(knm)$ 时间复杂度有很大的提升。在通过 CLR 算法得到最后的各节点的 Rank 值以后，接下来就是如何依据节点的 Rank 值来得到可以使影响力最大化传播的 Top-K 节点集的问题。

3.3.2 影响力社团评价策略

因为社团结构中节点之间的联系比较紧密，在社交网络中信息在社团内部的传播速度要比在整个网络中的快，所以提出依据社团影响力的大小来选取具有影响力的 Top-k 节点集。如何通过社团结构来选择能使信息最大化传播的 Top-K 节点集的问题，在这里转化为如何定义社团的影响力的问题。以下为社团影响力定义的三种策略：

(1) Size-Oriented(SO)：由于社团内的传播相对与社团间的传播要快，所以社团本身所包含节点的数量可以作为社团影响力的指标，即社团越大，表明可传播或可影响的用户越多。

(2) PageRank-Oriented(PO)：社团中的节点在整个网络中的拓扑属性也可作为判断社团影响力的指标。比如节点的度中心性、介数、接近度等指标，由于实验中节点的 PageRank 属性的表现效果要优于其他指标，所以将社团中所有节点的 PageRank 属性值之和作为评价社团影响力的指标，即社团中节点的属性之和越大，则说明该社团的影响力越大。

(3) LeaderRank-Oriented(LO)：依据 LeaderRank 算法的思想，在每个社团中都添加一个社团节点，并且在这个社团节点与当前社团中的用户之间添加双向的连接。原始的图包含 N 个节点和 E 条有向边，添加社团节点后变成了包含 $N+C$ 个节点和 $E+2N$ 有向边的网络，其中 C 表示网络中社团的个数。计算新生成网络节点的 LeaderRank，通过各个社团中的社团节点的 LeaderRank 数值作为评价社团影响力的指标，即社团节点的 LeaderRank 的数值越大，则说明该社团的影响力越大。

通过社团的影响力评价指标得到社团序列后，依据节点的 Rank 值选取每个社团中最具影响力的节点，直至筛选出 k 个大小的用户集，最后得到具有影响力的 Top-k 节点集合。

3.4 实验结果与分析

本文所用的实验数据为公开的网络数据集。数据集名称、数据规模、实际网络的相关描述以及来源如表 1 所示。

数据集名称	节点数	边数	描述
Twitter①	23,370	33,101	Twitter 中用户之间的信息转发关系的有向网络。
NYCM②	94,574	213,754	Twitter 中用户就呼吁全世界关注天气这一事件的转发关系的有向网络。

表 3-1 数据集来源及描述

注：①: http://snap.stanford.edu/data/egonets-Twitter.html
②: http://deim.urv.cat/~manlio.dedomenico/data.php

为了证明算法的有效性，本文分别从实验效果和效率两个方面，与经典的贪婪算法和 Leader 算法进行对比。由于本文中的实验是在社团发现算法的基础上进行的，因此不同的社团发现算法会对实验结果产生一定的影响，本文就不同的社团发现算法给出相应的分析，同时对于不同的社团影响力评价策略也进行了对比试验，并给出分析。

相关的实验环境，硬件环境为：i7-4790 CPU 3.60GHz，8GB，软件环境为：Linux 64 位操作系统，Python2.7 编程语言，PyCharm IDE。

3.4.1 CLR 算法 Top-K 影响力对比与分析

如图 3-3 表示不同的算法在两个数据集中的不同表现。图中的横坐标表示所选取的节点集大小，纵坐标表示节点集所能影响网络中的节点总数。虽然理论上已经证明贪婪算法能够得到比较高的最优解近似值，但是在利用贪婪算法求解 Top-K 节点集时，每次选取表现最优的节点添加到结果集中，然而所选取的最优节点在计算节点的影响力值的时候存在误差，所以从图 3-3 的 Twitter 实验数据显示贪婪算法表现出来的效果并不比 CLR 算法好。同时，也不难看出表现较差的分别依据节点的介数和接近度属性排名方法得到的 Top-K 节点集，而 LeaderRank 方法的效果却介于 CLR 与接近度之间。从而进一步说明了本章中的 CLR 算法的有效性，再结合社团结构的性质后表现出来的效果明显优于原先的 LeaderRank 算法，同时也证明了 CLR 算法拥有与贪婪算法相近的近似解比率。

在图 3-4 中的 HYCM 的实验数据显示，相对于该数据的节点数量来说，众多算法的总体效果差不多，CLR 算法计算得到的影响力指标略低于贪婪算法和 LeaderRank 算法，而优于剩余两种算法。同时在拥有将近 10 万节点数的网络数据中，包括贪婪算法在内的所有算法的 Top-100 节点集合的影响力不到整个网络的 5%，相比 Twitter 数据集中的 34%来说小了很多，其主要原因在于，在 NYCM

数据集中社团的结构不明显，可以从表 3-3 中的两个数据集中的模块度对比看出差别。因为前面介绍过，模块度表示的实际意义是描述网络中社团内存在的边的总数在整个网络中边的总数中所占的比例，比例越大说明社团结构越明显，比例越小说明社团结构越模糊。在表 3-4 中 NYCM 数据集中的模块度远小于 Twitter 数据集，因此影响力传播的比例 NYCM 中的 5%也远小于 Twitter 中的 34%。

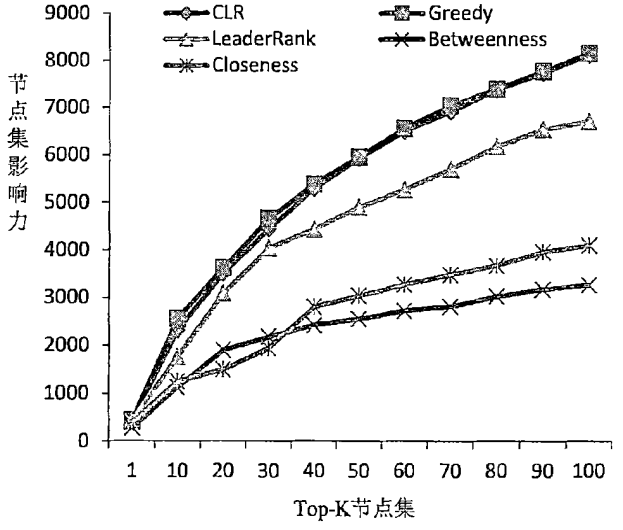


图 3-3 数据集 Twitter 的 Top-K 影响力

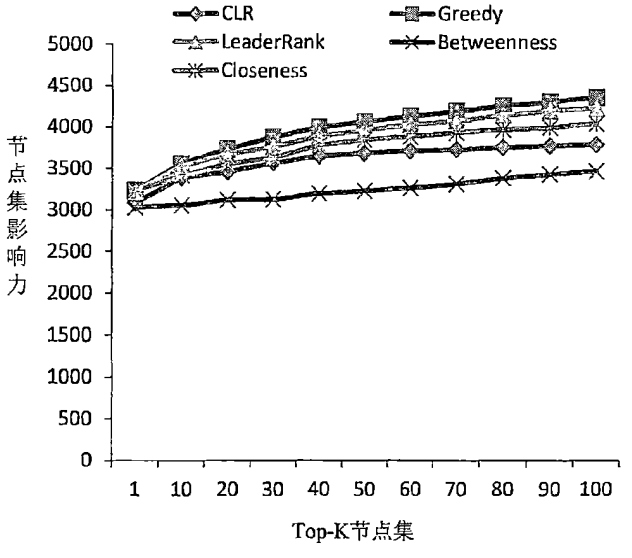


图 3-4 数据集 NYCM 的 Top-K 影响力

数据集	模块度
Twitter	0.84
NYCM	0.47

表 3-2 数据集中模块度对比

从实验效率上来说，贪婪算法的时间复杂度比较高，其复杂度接近于O(knm)，其中 n，m 分别表示网络中节点数和边数，k 表示所需的节点集合大小；而本文中的 CLR 算法中的特征向量的计算已经被证明可以通过有限次数的迭代而得到，所以其时间复杂度主要来自于社团发现算法，本文中的社团发现算法使用的是 LPA 标签传播算法，LPA 算法具有近似于线性时间复杂度的优点。因此从时间效率上来比较，近似于线性时间的 CLR 算法要明显优于传统的贪婪算法，两个程序运行的时间对比结果如下表 3-3 所示：

CLR 算法是在 LeaderRank 算法的基础上添加了节点的社团属性，由于本章节选取的算法是接近线性的标签传播算法，无论是 LeaderRank 算法还是 CLR 算法在时间效率上都接近线性，因此在时间效率上不做两种算法的比较。

数据	算法	Top-10	Top-30	Top-50	Top-100
Twitter	Greedy	17.7	67.6	122.5	370.4
	CLR	**2.7**	**2.7**	**2.7**	**2.7**
NYCM	Greedy	9303	9503	9708	10237
	CLR	**3.0**	**3.0**	**3.0**	**3.0**

表 3-3 运行时间对比

从表 3-3 中可以看出，随着选取节点集合大小的增加，贪婪算法的运行时间也在增加，而本章所提出的 CLR 算法却保持不变，并且从时间上比贪婪算法提升了很大。从两个数据集中的算法比较来看，随着网络数据集的增大贪婪算法的运行时间急剧增加，而 CLR 算法却继续维持在很小的时间内，由此可以看出 CLR 算法能够解决较大数据集的 Top-k 影响力节点选取问题，而贪婪算法则不能。因此随着网络数据集的增大，CLR 算法在时间效率上的提升也在不断增大。

3.4.2 不同社团影响力策略

由于社团内部节点之间的连接比较紧密，因此信息在社团内的传播速度和范围要比社团外大得多，所以如何定义某个社团的影响力就变得至关重要。以下实验为对比三种不同的社团影响力策略对选取具有影响力节点的影响，三种社团影响力策略分别是，Size-Oriented(SO)、PageRank-Oriented (PO)与 LeaderRank-Oriented(LO)。

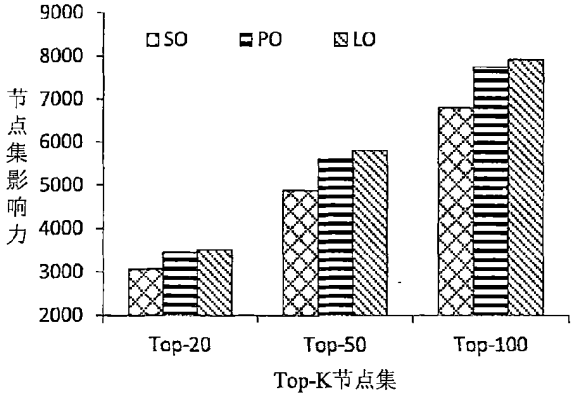


图 3-5 不同算法社团算法 Top-K 影响力

图 3-5 中 SO、PO、LO 三种方法分别对应上述所提到的三种社团影响力定义策略，从图中可以看出 LO 策略所表现出来的效果要优于其他两种方法。SO 策略和 PO 策略着重体现的是社团所包含的局部属性，而 LO 策略是依据所添加的社团节点的 LeaderRank 值，因此计算出来的结果具有全局属性。SO 策略是将社团中节点的数量作为社团的影响力，实际情况下如果社团的个数比较大，但是社团中信息传播速率比较低，那么实际上社团中的节点能够受到某个节点的影响力的数量也不大，因此比较片面；从图中还可以看出，PO 策略与 LO 策略的影响力相差不大，但是利用 LeaderRank 属性值衡量社团影响力大小要略微要好，主要原因是 LeaderRank 得到的是节点的特征向量精确值，而 PageRank 却是近似值。

3.4.3 不同的社团发现算法对比实验

由于 CLR 算法是借助社团结构的算法，因此不同的社团划分算法对实验的影响也会不同。本实验选取几种不同的社团划分算法进行对比，选取最有效的一种社团划分算法，实验的对比结构如图 3-6 所示，而不同社团划分算法得到社团结构的模块度系数如表 3-4 所示：

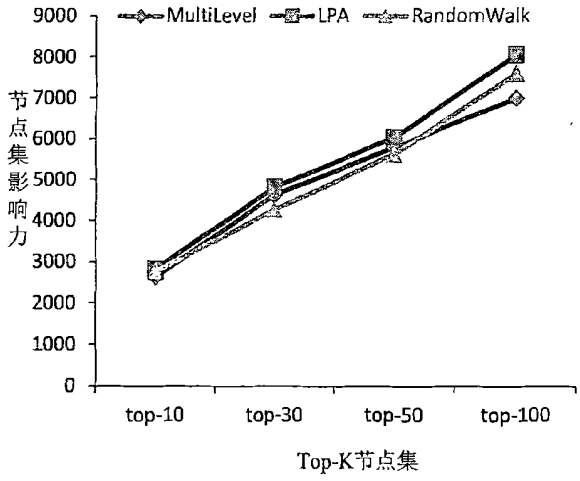


图 3-6 不同算法社团算法 Top-K 影响力

社团发现算法	Twitter	NYCM
Multi-Level	0.90	0.59
LPA	0.84	0.47
Random-Walk	0.86	0.53

表 3-4 数据集中社团划分算法模块度对比

从图 3-6 中可以看出 LPA 算法得到的影响力值略胜于其他两种算法，而随机游走算法和基于贪婪策略的 Multi-Level 算法的实验结果相近。LPA 社团划分算法是基于信息传播的思想来定义节点与某一标签的相似性距离如公式 3-4 所示，因此 LPA 社团划分算法所得到的社团结构内部信息的传播速度很高，在所选取的 Top-K 节点集的影响力也就大。从时间效率上考虑，LPA 算法的时间效率接近线性，所以无论从时间效率上还是最终的实验效果上来说，LPA 算法是最好的选择。

然而从社团发现算法模块度的数据来分析，从表 3-4 中的数据显示，模块度较高的算法却没有模块度低的算法表现的效果优越。模块度作为评价社团发现算法优劣的标准，模块度度量所描述的实际含义是，社团划分后社团内部中边的数量在整个网络中边的数量中所占的比例。同样实验在一定程度上也说明了，模块度在描述网络中社团的划分结果时存在一定的局限性。

3.4.4 实验数据可视化

图 3-7 和图 3-8 分别表示了贪婪算法和 CLR 算法在 Twitter 数据集上选取的具有影响力的 Top-10 节点集在整个网络中的示意图。所采用的画图工具为 Gephi-0.91，布局为 ForceAtlas2，劝阻 Hubs（画图时，沿输出的边分布吸引力，Hub 吸引较少，因此会被推倒边界）。

对比两个图中的节点选取结果，贪婪算法所选取的节点杂乱无章，而 CLR 算法选取的节点位于网络的边缘部分，由于画图的行为方式是劝阻 Hubs，所以网络中的 Hubs 节点大都分布在网络的边缘部分，因此 CLR 算法选择的节点大都为 Hubs 节点，相比贪婪算法来说更有实际意义。

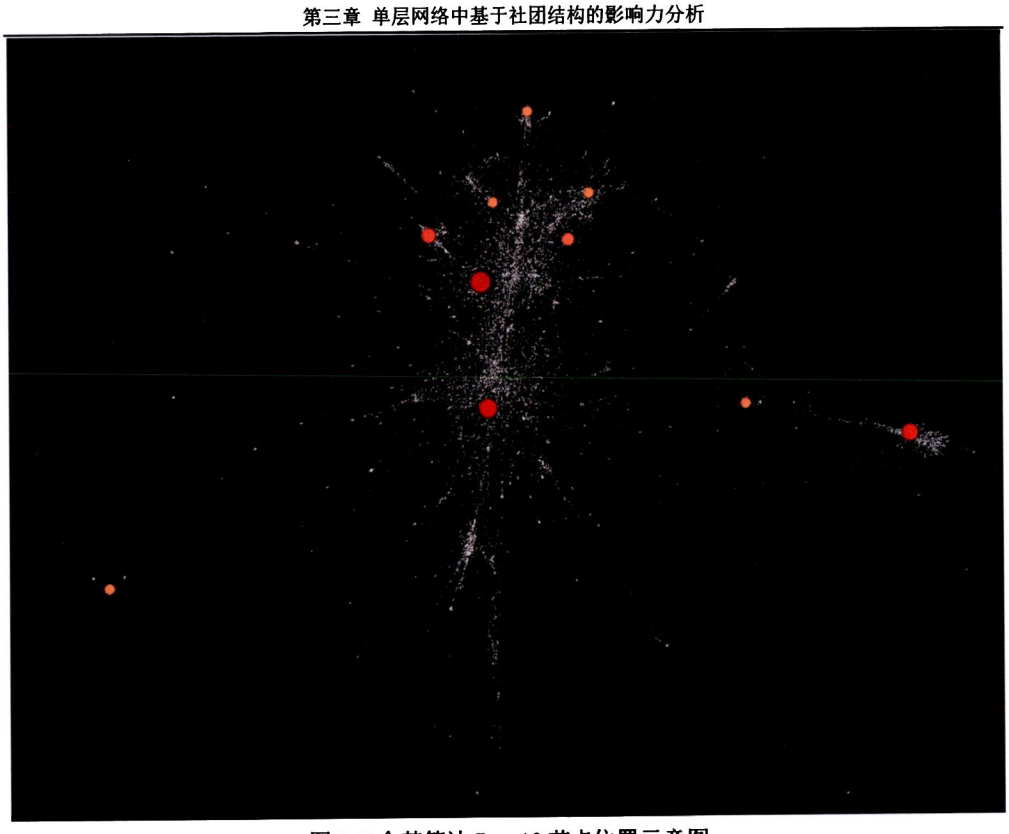


图 3.7 贪婪算法 Top-10 节点位置示意图

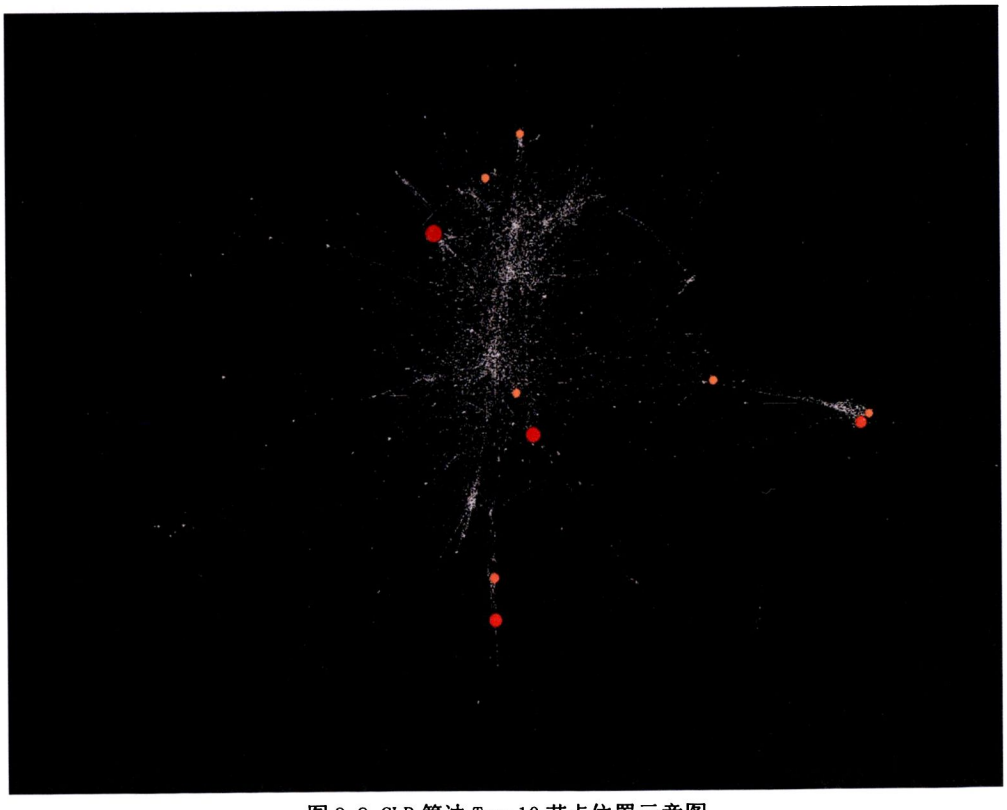


图 3.8 CLR 算法 Top-10 节点位置示意图

3.5 本章小结

本章节主要介绍了单个网络中通过结合社团结构和节点的拓扑属性来选择节点集，使得信息在网络中得到最大传播。从实验的对比结果来看，本章中的 CLR 算法与传统的贪婪算法所选取的 Top-K 集合表现出的影响力相差无几，但是从时间效率上来看，CLR 算法的表现远远优于贪婪算法。同时，对于不同的数据集，若网络数据中节点的社团划分越明显，CLR 算法选取的 Top-K 节点集合的影响力函数值越大。

第四章 多层网络中基于社团结构的影响力分析

由于多层网络中贪婪算法的时间效率太低，以及多层网络的数据集较大，因此将数据进行抽样处理。本章主要探究在多层网络中如何结合社团结构进行节点的影响力分析，因此分别从不同的多层网络社团划分策略，多层网络的抽样算法，以及多层网络下社团结合方法三个方面介绍，最后给出相关的实验结果与分析。

4.1 多层网络社团发现算法

总体上来看，与单层网络下的社团发现算法相同，多层网络中的社团发现算法也是利用节点在多层网络中的关系属性来划分连接紧密的节点集合，不同点在于不仅需要分析单层网络内的节点关系，而且需要考虑层与层之间的节点关系。多层网络中的社团发现算法大致分为两种[41]，第一种是将多层网络的问题按照一定的策略转化为单层网络的问题，然后在利用单层网络的社团发现算法来做；第二种是利用目前的方法直接对多层网络进行处理。

4.1.1 单层网络社团划分转化问题

将多层网络的社团划分问题转化为单层网络社团划分问题，主要分为两种情况，一种是将多层网络进行聚合转化为单网络，然后利用现有的单层网络社团发现算法来解决问题，另外一种是先利用单层网络的社团发现算法分别对多层网络中的每层进行操作，然后再将多层所得到的结构进行合并。其中第一种方法中涉及到多层网络的合并与单层网络的社团划分，单层网络的社团划分这里不再赘述，网络的融合如公式 4-1 所示：

$$G(i,j)=1-\prod_{g^r}(1-g_{ij}^r) \tag{4.1}$$

公式 4-1 中的 g_{ij}^r 表示多层网络中第 r 层网络节点 i 与节点 j 之间的边权值。

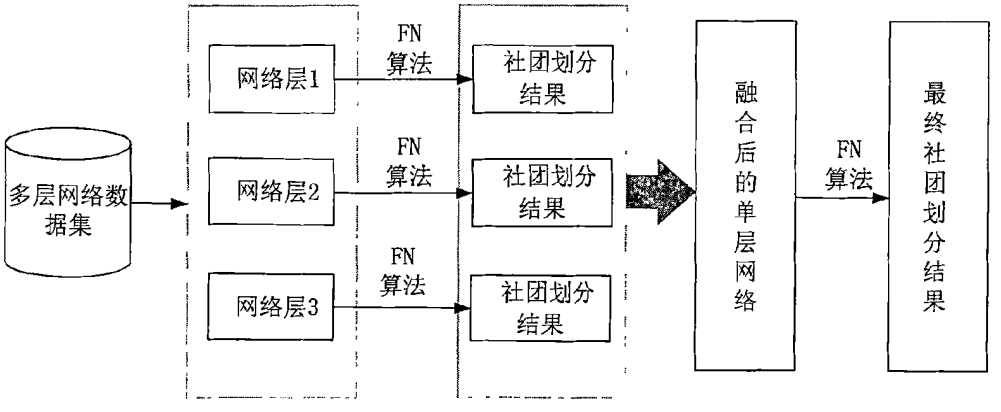


图 4-1 分层社团划分后融合算法流程图

第二种方法中对每层中社团划分的结果进行的合并问题。文献[42]中将多层网络的每层网络分别用 FN 算法得到相应的社团划分结果，然后将每层的社团划分结果通过公式 4-2 融合成一层网络，最后利用 FN 算法划分社团，具体的算法流程如图 4-1 所示：

$$G(i,j)=\sum_{r} g_r(i,j) \tag{4.2}$$

公式 4-2 中 $g_r(i,j)$ 表示第 r 层网络的社团划分结果，即在 r 层网络中如果节点 i 与节点 j 属于同一个社团则 $g_r(i,j)=1$，否则为 0. 从中可以看出，在融合多层社团划分结果时，只考虑了每层网络的社团中的连接，完全忽略了社团之间节点的连接，这种融合方式得到的结果会有很大的信息丢失，最后的社团划分结果也会失真。

4.1.2 多层网络社团发现算法

为了保证原数据信息的完整性，将多层网络用超级矩阵表示，然后利用现有的单层网络处理方法进行社团划分，以下为基于多层网络随机游走的社团划分算法[16]。

与单层网络中的随机游走算法思路类似，不同之处在于算法涉及到多层网络节点之间的随机游走，然后依据随机游走得到概率矩阵，以此来计算节点之间的距离，最后依据距离来进行社团的聚集。算法在多层网络中定义了四种游走方式，分别是同层网络中游走到邻居节点和停留在当前节点，不同层网络中游走到其它层网络对应的锚节点和这个锚节点所在层的邻居节点。

为了保留多层网络完整的特征，利用超级矩阵 A 的来存储网络的拓扑特征。在矩阵 A 的对角线的位置上有 L 个 N×N 的矩阵，每个矩阵分别是每层网络对应的邻接矩阵，然后在矩阵 A 的反对角线上对应的是每层网络之间锚节点对应的边，假如节点 i 在 k 层与 l 层网络中同时出现，则 $A_{(i,k)(i,l)}=\left|N_{i,k}\cap N_{i,l}\right|$，否则$A_{(i,k)(i,l)}=0$，其中$N_{i,k}$与$N_{i,l}$分别表示节点 i 在 k 层与 l 层的邻居节点集。多层网络四种随机游走方式用公式表示为：

$$\begin{aligned} \mathcal{P}_{(i,k)(i,k)}&=\frac{A_{(i,k)(i,k)}}{K_{i,k}} \quad \mathcal{P}_{(i,k)(j,k)}=\frac{A_{(i,k)(j,k)}}{K_{i,k}} \\ \mathcal{P}_{(i,k)(i,l)}&=\frac{A_{(i,k)(i,l)}}{K_{i,k}} \quad \mathcal{P}_{(i,k)(j,l)}=0 \end{aligned} \tag{4.3}$$

其中

$$K_{i,k}=\sum_{j,l} A_{(i,k)(j,l)} \tag{4.4}$$

经过 t 次随机游走后，随机游走概率矩阵$\mathcal{P}(t)$：

$$\mathcal{P}(t)=\mathcal{P}(t-1)\mathcal{P}=\mathcal{P}^{t} \tag{4.5}$$

与单层网络中的随机游走算法思路相同，通过已知的概率矩阵$\mathcal{P}(t)$求解多层网络中各个节点之间的距离，即相似性矩阵 S。在计算相似性矩阵时，分别定义了同层节点的相似性与不同层节点的相似性：

$$S(t)_{(i,k)(j,k)}=\sqrt{\sum_{h=1}^{N}\sum_{m=1}^{L}\frac{(\mathcal{P}^{t}_{(i,k)(h,m)}-\mathcal{P}^{t}_{(j,k)(h,m)})^{2}}{K_{h,m}}} \tag{4.6}$$

$$S(t)_{(i,k)(j,l)}=\sqrt{s_1+s_2+s_3} \tag{4.7}$$

其中

$$s_1=\sum_{h=1}^{N}(\frac{\mathcal{P}^{t}_{(i,k)(h,k)}}{\sqrt{K_{h,k}}}-\frac{\mathcal{P}^{t}_{(j,l)(h,l)}}{\sqrt{K_{h,l}}})^{2} \tag{4.8}$$

$$s_2=\sum_{h=1}^{N}(\frac{\mathcal{P}^{t}_{(i,k)(h,l)}}{\sqrt{K_{h,l}}}-\frac{\mathcal{P}^{t}_{(j,l)(h,k)}}{\sqrt{K_{h,k}}})^{2} \tag{4.9}$$

$$s_3=\sum_{\substack{h=1\\ m\neq k,l}}^{N}\sum_{m=1}^{L}\frac{(\mathcal{P}^{t}_{(i,k)(h,m)}-\mathcal{P}^{t}_{(j,l)(h,m)})^{2}}{K_{h,m}} \tag{4.10}$$

从公式 4.6 定义节点在单层网络中的距离，如果同层网络的两个节点到它们所在层其它节点的概率之差越小，那么这两个节点在当前层网络中的距离越小。公式 4.7 定义多层网络的相似性，其中s_1表示当处在不同层中的两个节点到各自所在层的节点的概率大致相同时，s_1值就越小；s_2表示当两个节点分别到对方所在层中的节点概率大致相同时，s_2值就越小；s_3表示当两个节点到除了该节点的所在层的其它层的节点的概率大致相同时，s_3值就越小。得到节点的相似矩阵以后，通过相应的节点聚集方法得到最后的社团划分。

4.2 多层网络抽样算法

网络抽样是在给定一个大型的目标网络上，运用一定的策略和方法选取节点和边来创建一个新的小型网络。一般来说可以通过两种方式来考虑网络抽样[44]，一种是缩小网络规模，并且在一定程度上能与原先的大型网络匹配为目标的抽样；另外一种是针对时序网络的抽样，目标是能够回溯来匹配原先大型网络的时序演变。本文主要研究的不是时序网络，因此抽样的目标为缩小网络规模。一般来说对网络的抽样方法大致分为三种类型，第一种是随机选取节点的抽样；第二种是随机选取网络中边的抽样；第三种以扩展的方式模拟随机游走或病毒传播来选取网络中具有代表性的节点与边的集合。

一般的抽样算法是采用第三种抽样方法，主要的过程是，从网络中随机选择一个节点，然后以该节点为起始在网络中随机游走，并记录所有的游走路径，最后得到随机游走生成的网络，即为抽

样网络。在文献[45]中采用多中心的随机游走抽样，与普通的随机游走抽样不同之处在于，初始选取的不是一个节点，而是一个节点集合，其算法思路为：1）每次随机游走时，便从节点集合 S 中随机选择一个节点 u；2）以该节点为起始节点随机选择该节点的邻居节点 v；3）在集合 S 中用新的节点 v 代替原先的 u，并将边（u，v）添加到抽样的结果集中；4）重复步骤 1，2，3，直至满足抽样的要求。

对于多层网络抽样算法来说唯一不同之处在于，当对每层网络进行随机游走抽样时，选取同一个节点集合作为起始节点集。具体的算法流程如算法 4-1 所示：

输入：$G = \{g_1, g_2, ..., g_r\}$,B = $|V|/10$, m=10, c=1

输出：抽样结果图g^s

1. U ← $V_1 \cup V_2 \cup ... \cup V_r$
2. L ← $\{v_1, v_2, .., v_m\}$ // 从集合 U 中随机抽样大小为 m 的节点列表
3. for k=1 to r do
4. n ← 0 // n 表示随机游走的步数
5. $l_k \leftarrow L$
6. $g_k^s \leftarrow \emptyset$ // 初始化第 k 层网络的抽样网络
7. while n ≥ B − mc
8. 选取节点u ∈ l_1，选取的概率$p_u = {\deg(u)}/{\sum_{v \in l_k} \deg(v)}$
9. 随机选取节点 u 的邻居节点 v，$g_k^s = g_k^s + \{(u, v)\}$
10. l_k节点集中以节点 v 替换 u
11. n ← n + 1
12. end while
13. $g^s = g^s + \{g_k^s\}$ //将抽样得到的第 k 层抽样网络添加到结果中
14. end for
15. return g^s

算法 4-1 多层网络 Frontier 抽样算法

4.3 多层网络中基于社团的 LeaderRank 影响力分析

与单层网络中的算法思路类似，通过社团发现算法得到社团划分结果，构建包含 Leader 节点的新网络，然后再利用节点的 Rank 属性值选择 Top-K 节点集合。在构造多层网络中的 Leader 网络时，分别用三种方法来实现，一种方法是分层构建，然后将结果融合；第二种方法将多层网络进行融合再利用单层网络中的 CLR 算法来完成；第三种则是构建包含 Leader 节点的超级网络。

第二种方法是利用公式 4.1 对多层网络进行融合生成单层网络后，再利用单层网络中的 CLR 算法来完成，CLR 算法在第三章中已经介绍，这里不在赘述。

4.3.1 Multi-Community-LeaderRank（MCLR）分层构建算法

多层网络中分层构建的思路类似于单层网络，是将多层网络拆分成独立的多个单层网络，利用

CLR 算法分别处理，然后将结果进行合并。结果合并的过程，是节点在不同网络中的 Rank 值进行累加的过程，算法的具体流程如算法 4-2 所示：

输入： 网络$\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \dots, \mathcal{G}_m\}$
输出： 节点集合$\mathcal{V}$权值列表

```
1.  d ← ∅
2.  for r=1 to m
3.      d^r ← CLR(G_r) // 调用单层网络中的 CLR 算法
4.      d ← d + {d^r} // 对每层网络计算的结果进行累加
5.  end for
6.  return d
```

算法 4-2 MCLR 分层构建算法

4.3.2 Super-Multi-Community-LeaderRank（SMCLR）超级网络构建算法

超级网络构建算法的算法思想是将多层网络构造成一个超级矩阵，超级矩阵的对角线表示的是各个层网络的邻接矩阵，反对角线则表示各层之间节点的对应关系，表示是否互为锚节点。随后对网络进行跨层的社团划分，然后利用 LeaderRank 的算法思想，构建 Leader 多层网络，最后得到节点的 Rank 值。算法的具体流程如算法 4-3 所示：

输入： 网络$\mathcal{G} = \{g_1, g_2, g_3, \dots, g_m\}$
输出： 节点集合$\mathcal{V}$权值列表

```
1.  super_matrix ← mix(g_1, g_2, g_3, .. , g_m) // 构建超级矩阵
2.  C = Multi_Random_walk(G)  // 多层网络随机游走算法进
    行社团划分
3.  super_graph ← super_matrix // 将超级矩阵用一层超级网
    络表示
4.  f_super = CLR(super_graph)  // 调用单层网络 CLR 算法
5.  U ← V_1 ∪ V_2 ∪ .. ∪ V_m        // 得到多层网络中节点全集 U
6.  f_U ← ∅
7.  for v in U do
8.      f_U = max(f_super(v))  // 统计不同层中节点最大值
9.  end for
10. return f_U
```

算法 4-3 SMCLR 超级网络构建算法

算法第一步构建超级矩阵的具体过程，同多层网络中随机游走社团划分算法构建的超级矩阵相同，算法的核心思想是将超级矩阵看作一个大型独立的网络来处理，把网络中存在的相同实体看成不同的节点，与不同实体之间关系不同之处体现在边的权值上。算法第二步是用多层网络随机游走社团划分算法得到的社团划分结果。算法第三步是将构建的超级矩阵转变成超级网络，然后利用单

层网络中的 CLR 算法计算多层网络中所有节点的 Rank 值，最后综合统计得出最终的节点 Rank 值。

4.4 实验结果与分析

本章实验的主要内容是，通过结合多层网络的研究方法，将单层网络中的 CLR 算法应用于多层网络。实验的主要目的是，保证所选取的 Top-K 节点集合具有较高的影响力值的同时，在算法的时间效率上也有要求。

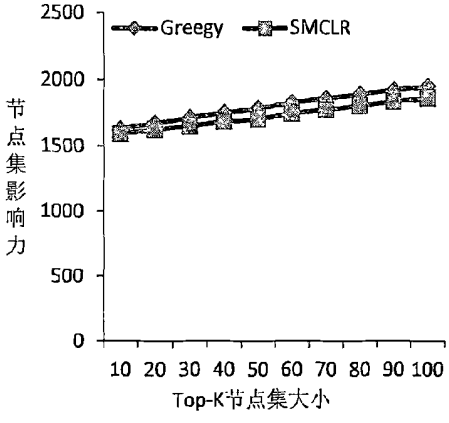
相关实验的环境，硬件环境为：i7-4790 CPU 3.60GHz，8GB，软件环境为：Linux 64 位操作系统，Python2.7 编程语言，PyCharm IDE。

4.4.1 抽样数据集的实验

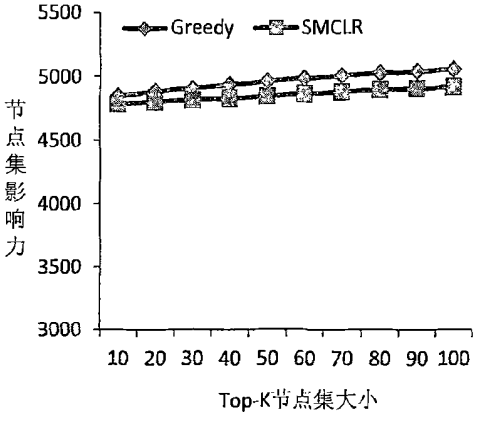
Higgs Twitter 数据集是对 Twitter 上用户对希格斯粒子新特性发布后，用户对该信息的转发以及评论和提及所形成的有向带有权值的多层网络。Higgs 抽样数据是对原数据中的转发关系，提及关系以及回复关系的三层网络抽取的节点全映射多层网络。

数据集	层数	节点数	边数	关系描述
Higgs Twitter	第一层	456,626	14,855,875	好友关系
	第二层	425,008	733,647	转发
	第三层	302,975	449,827	提及
	第四层	37,366	30,836	回复
Higgs 全映射抽样数据	第二层	8,610	29,012	转发
	第三层	8,610	25,124	提及
	第四层	8,610	10,199	回复
Higgs 非全映射抽样数据	第二层	6,505	16,637	转发
	第三层	6,575	16,743	提及

表 4-1 Higgs Twitter 数据集



非全映射抽样数据集实验



全映射抽样数据集实验

图 4-2 抽样数据的影响力曲线

在不考虑时间效率的情况下，解决当前问题最好的方法为贪婪算法，因为贪婪算法是穷尽所有可能来选取的 Top-K 节点集，因此贪婪算法能够表现出局部最优效果。根据图 4.2 中所示的曲线来说，多层网络中 SMCLR 算法在实验效果上与贪婪算法相似，但从算法的时间效率上来说，本章所提出的多层网络 SMCLR 算法大大优于贪婪算法，Top-K 节点集选取实验的运行时间如下表 4.2 所示：

实验	算法	Top-10	Top-30	Top-50	Top-100
Higgs 非全映射抽样数据	Greedy	8,698	8,793	8,885	9,106
	SMCLR	**3**	**3**	**3**	**3**
Higgs 全映射抽样数据	Greedy	468	494	522	591
	SMCLR	**1**	**1**	**1**	**1**

表 4-2 抽样数据的运行时间（秒）

对于不同的抽样数据集，在贪婪算法的节点集选取中，随着 K 值的增大程序的运行时间也在增加，而在 SMCLR 算法所需要的时间基本是固定的；同时，在 Higgs 非全映射抽样数据中，多层网络 SMCLR 算法在时间效率上有很大提升，而在另外一个抽样数据中也提升了将近 500 倍，多层网络结构越复杂，算法的时间效率提升的越明显。

4.4.2 完整数据集实验

为了排除由于抽样算法对原数据集中产生的误差，在这里选择相对小的完整数据集合来证明算法的可行性。由于贪婪算法的运行效率很低，这里采取非完整传播路径的影响力评价方法来选取节点集，同样对多层网络的 SMCLR 算法也采用非完整传播的影响力评价方法，在以下的实验中将信息的传播周期 T=5。

数据集	节点数	边数	描述
NYCM	94,574	213,754	好友
	50,054	131,679	转发

表 4-3 NYCM 完整数据集

实验的数据集为多层网络数据集中两层完整的网络数据，数据集 NYCM 是社交网络 Twitter 上用户关于纽约的气候游行这一事件的相关评论和转发网络，如表 4-3 所示：

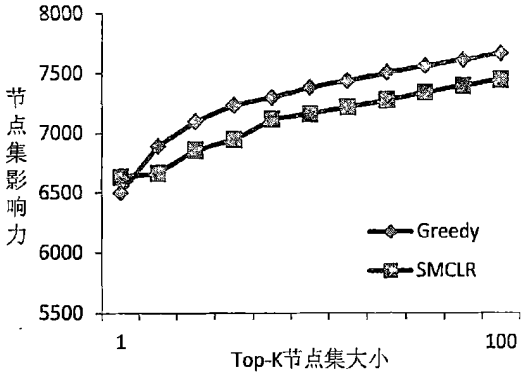


图 4-3 Step=5 NYCM 影响力曲线

比值	Top-1	Top-20	Top-40	Top-60	Top-80	Top-100
SMCLR/Greedy	1.019849	0.966295	0.975466	0.970284	0.970638	0.971425

表 4-4 Step=5 Greedy 与 SMCLR 算法的影响力比例

图 4-3 中所示的曲线是在完整的数据集 NYCM 上 SMCLR 算法和贪婪算法的 Top-K 节点集影响力曲线，同时在表 4-4 中给出两种算法影响力的比值。从表 4-4 和图 4-3 中都可以看出，在选取 Top-1 节点时，贪婪算法所选取的节点所表现出来的影响力效果并没有 SMCLR 算法好，主要原因是贪婪算法是在所有节点中利用穷举的方法，计算每个节点的影响力，然后依次选择其中影响力最大的节点，由于节点之间信息传播是以边上的权值随机的过程，而在计算节点影响力的时，只计算一次随机传播的结果作为当前节点的影响力，所以贪婪算法在选取的节点具有一定程度上的误差，在最后多次计算节点的影响力并求均值时，Top-1 节点选取时贪婪算法的计算误差就更大，因此 Top-1 时 SMCLR 算法比贪婪算法高是合理的，同时也说明了本章中的 SMCLR 算法的合理性。

从表 4-4 中可以看出，除了 Top-1 节点在 SMCLR 算法影响力比贪婪算法好，其它节点集 SMCLR 算法所选取节点集的影响力能够达到贪婪算法的 96%以上。

算法	Top-10	Top-30	Top-50	Top-100
Greedy	5,773	9,461	13,008	22,195
SMCLR	**13**	**13**	**13**	**13**

表 4-5 Step=5 运行时间（秒）

在表 4-5 表示两种算法选取不同节点集合大小的运行时间，从运行时间上的对比可以看出，在信息传播周期 T=5 时，本章中的 SMCLR 算法在时间上提升了 400 至 1700 多倍，并且贪婪算法程序的运行时间随着节点集合大小的增加而增加，而 SMCLR 算法保持稳定。

4.4.3 几种社团发现算法的对比

在多层网络中 MCLR 算法与 SMCLR 算法是结合社团结构而构建 Leader 网络，然后计算节点权值的算法，因此算法的关键是社团划分算法。在多层网络中社团划分策略不同，构建 Leader 网络的策略也就不同。三种社团划分算法的结合策略所选取的节点集影响力曲线，如图 4-4 所示：

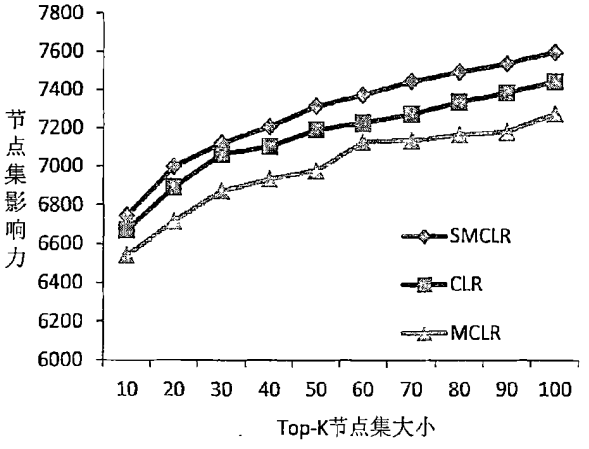


图 4-4 不同社团划分结合策略对比

在图 4-4 中的三条曲线分别表示：SMCLR-超级网络构建算法曲线，多层融合后 CLR 算法曲线和 MCLR-分层构建算法曲线，从图中可以看出在三种算法中超级网络构建算法表现的效果最好，而分层构建算法表现的效果最差，多层融合构建算法居于两种算法之间。其中主要原因在于，分层构建算法是将多层网络中每层网络隔离处理，最后将每层网络得到的节点权值综合统计，最后得到多层网络的综合结果，与超级网络构建算法相比，该算法在处理的过程中遗漏掉各层网络中节点的相互关系，因此超级网络构建算法比分层构建算法要好很多。而多层融合构建算法是将多层网络依据公式 *4.1* 融合得到的单层网络，继而利用单层网络中的 CLR 算法得到的结果，在一定程度上保留了多层网络中的部分信息，所以介于其它两种算法之间。

由于三种算法在整体的时间效率都比较高，因此不同的社团发现策略在时间效率上差别并不大，主要的差别只是体现在实验效果上。

4.5 本章小结

在本章中介绍了多层网络中社团的几种划分算法，并且就不同的社团划分算法提出了三种不同的多层网络社团结构的结合算法。从分析的结果来看，无论是网络抽样算法所得到的网络数据，还是完整的小型的网络数据，SMCLR 算法所选取节点集的影响力都有很好的表现。同时，相比贪婪算法来说，SMCLR 算法的时间效率也大大的提升。

第五章 总结与展望

5.1 总结

在社交网络中发现具有广泛影响力的用户在很多领域都具有重要作用，例如创新采用、社会舆论传播和导向、群体行为形成和发展等。通过社交网络中影响力的传播方式，商业营销能够以较低费用将新产品推广到整个市场，从而产生较大的社会影响力和商业价值。从社交网络中发现具有影响力的 Top-K 节点集，对研究信息在网络上的快速传播具有重要意义。

从网络中选取大小为 K 的用户集合，并以这个集合为触发集合，从而使信息在网络中的传播最大。然而影响力传播最大化是 NP-Hard 问题，经典的贪婪算法虽然能得到近似度较高的近似解，但是算法的计算代价太大。本文中提出的算法结合网络中社团的特性，在保证具有较高的近似解的前提条件下，利用网络拓扑信息和用户行为的综合分析方法，提升算法的时间效率。

网络中的社团结构具有社团内节点强连接，与社团间节点弱连接的这一特性，信息的传播速度会受到社团的影响。本文从研究节点的拓扑属性出发，基于 LeaderRank 的算法思想，在独立级联的传播模型下，在单层网中提出一种基于社团结构的 CLR 算法来发掘有影响力的 Top-K 节点集，并结合多层网络的研究策略将 CLR 算法思想运用于多层网络中产生多层网络 SMCLR 算法。

本文完成的主要工作分为两个方面，单层网络中的影响力研究和多层网络中的影响力研究：

1）在单层网络中的研究主要包括：

(1) 网络的拓扑属性描述了节点在网络中局部或全局的位置，在一定程度上能够表现出节点的信息，不仅能表现出节点本身在网络中的实际含义，更能在时间效率上有很大的提升。为了提升算法在庞大社交网络中的时间效率，考虑利用网络拓扑属性，同时结合网络的其它属性以提高实际运算效果。

(2) 本研究在 LeaderRank 算法的基础上提出改进，根据网络社团的发现算法将节点标记为不同类别，由于不同的社团算法也可能对最终选取的具有影响力的 Top-K 节点集合有影响，尝试找出一种合适的社团发现算法，并将其与 LeaderRank 结合达到更好的实验效果。

(3) 网络中信息的传播方式和节点集合的影响力评价标准。分别在线性阈值模型和独立级联模型中计算触发集合 A 的影响力函数，为了证明单层网络中结合社团结构的 CLR 算法的可行性，将其与传统的贪婪算法在所选取节点集合的影响力效果上，以及算法的时间效率上相比较。

2）在多层网络中的研究主要包括：

(1) 利用多层网络相关问题的研究方法，将多层网络中的社团划分算法大致分为两种，第一种

是将多层网络的问题按照一定的策略转化为单层网络的问题，然后利用单层网络的社团发现算法，具体分为融合后划分和划分后融合两种；第二种是利用目前的方法直接对多层网络进行协同处理。

(2) 就不同的多层网络社团划分算法，结合 LeaderRank 算法，然后依据多层网络的研究方法，提出三种不同的结合策略，第一种是将多层网络融合后结合策略，第二种是分层先结合 LeaderRank 算法再融合的策略，最后一种是综合考虑多层网络的结构协同结合策略。

(3) 相关的多层社交网络数据集中数据量太大，依据现有的网络抽样算法得到抽样数据集合，再结合多层网络中信息的传播方式，来研究多层网络中基于社团的具有影响力的 Top-K 节点集选取算法。

相关实验结果表明，结合社团影响力的方法能够较好地发现网络中具有影响力的 Top-K 节点。无论是单层网络中的 CLR 算法，还是多层网络的 SMCLR 算法，在所选取的 Top-K 集合表现出的影响力上与传统的贪婪算法相差无几，从时间效率上 CLR 算法与 SMCLR 算法都表现远远优于贪婪算法。

5.2 展望

本文中讨论了社交网络中如何利用网络的社团结构，来选取具有影响力的 Top-K 节点集的问题，提出在单层网络和多层网络中的相关算法，下一步的工作可以从以下几个方面展开：

1) 基于信息的传播模型，依据节点之间的信息传播速度来进行社团划分，保证社团内信息的传播速度，并依此得到社团影响力的计算策略，得到影响力 Top-K 社团，最后解决影响力最大化传播问题。

2) 在多层网络中，考虑各种节点之间关系的不同，导致不同层内的信息传播速度不同，以及各层网络之间的信息传播速度不同，通过协同分析的方式划分网络中有影响力的 Top-K 社团，最后解决多层网络中影响力最大化传播问题。

3) 在单层网络中，与 CLR 结合的标签传播算法要比随机游走算法要好，因此通过改进标签传播算法并将其应用在用超级矩阵表示的多层网络中，从而对多层网络进行社团划分，可能会提升多层网络中算法的效果。

致谢

在三年的硕士研究生学习和生活期间，我得到了许多老师、同学和亲人的关心和帮助。他们不仅帮助我在学业上取得进步，而且生活上也给予了我很多支持。从独立解决问题的能力到树立正确的人生观和价值观。在他们的帮助下，我各方面的综合能力得到提高，对于我以后的人生道路将是一笔宝贵的财富。

首先我要感谢我的导师何洁月教授，何老师治学严谨，平易近人。尤其是教我站在更高的角度审视问题，抓住问题的重点，让我能以更积极上进的态度对待学习和生活。何老师在我的课题研究和论文撰写方面提供了很多指导，她的谆谆教诲将激励我在以后的人生道路上奋勇前进，将是我以后学习的楷模。

同时，感谢实验室的兄弟姐妹们，融洽而团结的实验室氛围，使得原本枯燥无味的研究生生活变得轻松而愉悦。他们是罗浩、沈斌、张伟、王帅、郑承俊、王帅、胡平伍、周航。感谢他们在学习和生活中给予的指导和帮助。尤其感谢胡平伍、王帅、周航，在我实验遇到困难的时候，能及时耐心的帮我解决。正是有他们的陪伴，让我度过了一个快乐而有意义的研究生生活。能够在这样一个大家庭和和谐的氛围中成长，是我极大的荣幸。在此我向各位表示深深的感谢。

在此，我还要感谢我亲爱的父母和朋友，多年的求学生涯中他们对我的无条件的支持是我取之不竭的动力和信心的源泉。若是没有他们的支持和鼓励，我也无法一路克服重重困难完成学业。

除此之外，我要感谢东南大学对我的培养。时光荏苒，三年离乡求学，转眼毕业在即。回想在东大的三年，有过懵懂、迷茫，但更多的是对明天的笃定。东大，见证了我的青春岁月，见证了我学习上的拼搏、生活中逐渐的成熟以及感情上的归属，在这里，我对母校充满无限感激和留恋。谨在此预祝母校 115 周岁生日快乐。

最后向评阅本毕业设计论文的各位专家、学者致以崇高的敬意。

参考文献

[1] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752.

[2] 赵之滢, 于海, 朱志良, 等. 基于网络社团结构的节点传播影响力分析[J]. 计算机学报, 2014, 37(4): 753-766.

[3] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[4] 韩忠明, 苑丽玲, 杨伟杰, 等. 加权社会网络中重要节点发现算法[J]. 计算机应用, 2013, 33(6): 1553-1557.

[5] Lü L, Zhang Y C, Yeung C H, et al. Leaders in social networks, the delicious case[J]. PloS one, 2011, 6(6): e21202.

[6] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature physics, 2010, 6(11): 888-893.

[7] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in twitter: The million follower fallacy[J]. Icwsm, 2010, 10(10-17): 30.

[8] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003: 137-146

[9] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model[C]//Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 88-97

[10] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009: 199-208.

[11] Nguyen D T, Zhang H, Das S, et al. Least cost influence in multiplex social networks: Model representation and analysis[C]//Data Mining (ICDM), 2013 IEEE 13th International Conference on. IEEE, 2013: 567-576.

[12] Zhan Q, Zhang J, Wang S, et al. Influence maximization across partially aligned heterogenous social networks[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer International Publishing, 2015: 58-69.

[13] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of approximations for maximizing submodular set functions—I[J]. Mathematical Programming, 1978, 14(1): 265-294.

[14] Wang Y, Cong G, Song G, et al. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 1039-1048.

[15] Shi C, Li Y, Zhang J, et al. A survey of heterogeneous information network analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(1): 17-37.

[16] Pons P, Latapy M. Computing communities in large networks using random walks[C]//International Symposium on Computer and Information Sciences. Springer Berlin Heidelberg, 2005: 284-293.

[17] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

[18] Saumell-Mendiola A, Serrano M Á, Boguná M. Epidemic spreading on interconnected networks[J]. Physical Review E, 2012, 86(2): 026106

[19] Buono C, Alvarez-Zuzek L G, Macri P A, et al. Epidemics in partially overlapped multiplex networks[J]. PloS one, 2014, 9(3): e92200.

[20] Gomez S, Diaz-Guilera A, Gomez-Gardenes J, et al. Diffusion dynamics on multiplex networks[J]. Physical review letters, 2013, 110(2): 028701.

[21] Szell M, Lambiotte R, Thurner S. Multirelational organization of large-scale social networks in an online world[J]. Proceedings of the National Academy of Sciences, 2010, 107(31): 13636-13641.

[22] Mucha P J, Richardson T, Macon K, et al. Community structure in time-dependent, multiscale, and multiplex networks[J]. science, 2010, 328(5980): 876-878.

[23] Cantador I, Castells P. Multilayered semantic social network modeling by ontology-based user profiles clustering: application to collaborative filtering[C]//International Conference on Knowledge Engineering and Knowledge Management. Springer Berlin Heidelberg, 2006: 334-349.

[24] Newman M. Networks: an introduction[M]. Oxford University Press,2010.

[25] Kazienko P, Musial K, Kukla E, et al. Multidimensional social network: model and analysis[J]. Computational Collective Intelligence. Technologies and Applications, 2011: 378-387.

[26] Pons P, Latapy M. Computing communities in large networks using random walks[C]//International Symposium on Computer and Information Sciences. Springer Berlin Heidelberg, 2005: 284-293.

[27] Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks[J]. Physical Review E, 2014, 89(3): 032804.

[28] Berlingerio M, Coscia M, Giannotti F. Finding redundant and complementary communities in multidimensional networks[C]//Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011: 2181-2184.

[29] Zhu G, Li K. A unified model for community detection of multiplex networks[M]//Web Information Systems Engineering–WISE 2014. Springer International Publishing, 2014: 31-46.

[30] Liu Z, Jiang C, Wang J, et al. The node importance in actual complex networks based on a multi-attribute ranking method[J]. Knowledge-Based Systems, 2015, 84: 56-66.

[31] Zhang J, Xu X K, Li P, et al. Node importance for dynamical process on networks: A multiscale characterization[J]. Chaos: an interdisciplinary journal of nonlinear science, 2011, 21(1): 016107.

[32] 于会, 刘尊, 李勇军. 基于多属性决策的复杂网络节点重要性综合评价方法[J]. 物理学报, 2013, 62(2): 20204-020204.

[33] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Physical review E, 2004, 69(2): 026113.

[34] Watts D J, Strogatz S H. Collective dynamics of 'small-world'networks[J]. nature, 1998, 393(6684): 440-442.

[35] Pons P, Latapy M. Computing communities in large networks using random walks[C]//International Symposium on Computer and Information Sciences. Springer Berlin Heidelberg, 2005: 284-293.

[36] Ward Jr J H. Hierarchical grouping to optimize an objective function[J]. Journal of the American statistical association, 1963, 58(301): 236-244.

[37] Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical review E, 2007, 76(3): 036106.

[38] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.

[39] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): P10008.

[40] Fortunato S, Barthelemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

[41] Kanawati R. Multiplex network mining: a brief survey[J]. IEEE Intelligent Informatics Bulletin, 2015, 16: 24-28.

[42] Chen X, Han G, Yuan L, et al. Community Detection in Multi-dimensional Network[C]//Computational Intelligence and Design (ISCID), 2015 8th International Symposium on. IEEE, 2015, 1: 598-601.

[43] Ribeiro B, Towsley D. Estimating and sampling graphs with multidimensional random walks[C]//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement.

ACM, 2010: 390-403.

[44] Leskovec J, Faloutsos C. Sampling from large graphs[C]//Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006: 631-636.

[45] Ribeiro B, Towsley D. Estimating and sampling graphs with multidimensional random walks[C]//Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. ACM, 2010: 390-403.

[46] Valdez L D, Buono C, Macri P A, et al. Social distancing strategies against disease spreading[J]. Fractals, 2013, 21(03n04): 1350019.