分类号：_____ 密级：_____

UDC：_____ 编号：_____

工学硕士学位论文

基于复杂网络的病毒营销传播机制研究

硕士研究生：周　颖

指 导 教 师：董宇欣　副教授

学科、专业：计算机科学与技术

论文主审人：董红斌　教授

哈尔滨工程大学

2015 年 3 月

哈尔滨工程大学
学位论文原创性声明

本人郑重声明：本论文的所有工作，是在导师的指导下，由作者本人独立完成的。有关观点、方法、数据和文献的引用已在文中指出，并与参考文献相对应。除文中已注明引用的内容外，本论文不包含任何其他个人或集体已经公开发表的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者（签字）：周毅

日期：2015 年 3 月 16 日

哈尔滨工程大学
学位论文授权使用声明

本人完全了解学校保护知识产权的有关规定，即研究生在校攻读学位期间论文工作的知识产权属于哈尔滨工程大学。哈尔滨工程大学有权保留并向国家有关部门或机构送交论文的复印件。本人允许哈尔滨工程大学将论文的部分或全部内容编入有关数据库进行检索，可采用影印、缩印或扫描等复制手段保存和汇编本学位论文，可以公布论文的全部内容。同时本人保证毕业后结合学位论文研究课题再撰写的论文一律注明作者第一署名单位为哈尔滨工程大学。涉密学位论文待解密后适用本声明。

本论文（☑在授予学位后即可　□在授予学位12个月后　□解密后）由哈尔滨工程大学送交有关部门进行保存、汇编等。

作者（签字）：周毅　　导师（签字）：[illegible]

日期：2015年3月16日　　2015 年3 月16日

分类号：_____ 密级：_____

U D C：_____ 编号：_____

工学硕士学位论文

基于复杂网络的病毒营销传播机制研究

硕 士 研 究 生：周　颖

指　导　教　师：董宇欣　副教授

学　位　级　别：工学硕士

学 科 、 专 业：计算机科学与技术

所　在　单　位：计算机科学与技术学院

论文提交日期：2014 年 12 月

论文答辩日期：2015 年 3 月

学位授予单位：哈尔滨工程大学

Classified Index：

U.D.C：

A Dissertation for the Degree of M. Eng

Research of Viral Marketing Propagation Mechanism Based on Complex Network

Candidate: Zhou Ying

Supervisor: A.Prof. Dong Yuxin

Academic Degree Applied for: Master of Engineering

Specialty: Computer Science and Technology

Date of Submission: Dec,2014

Date of Oral Examination: Mar,2015

University: Harbin Engineering University

摘　要

随着对复杂网络的深入研究，病毒营销作为新型的营销模式已经成为研究的热点。病毒营销又叫做口碑营销，主要是通过网络中的用户自发进行传播营销信息实现的。在病毒营销过程中，首先商家将产品对个别用户进行推广，当个别用户获得较好的口碑后，就会将商品推荐给其好友，获得口碑的新个体会以同样的方法将商品信息传播给他人，从而实现营销信息的传播范围的扩大。对复杂网络的病毒营销传播机制研究，是在真实的复杂网络的基础上，通过网络的开放特性，使得营销信息从几个节点传播给其邻居节点，最终目的是使复杂网络中的所有节点都能够受到营销信息的影响。

对复杂网络的病毒营销传播机制研究主要包含两个部分，一是对病毒营销传播过程进行建模，主要是用来模拟营销信息在复杂网络上的传播过程。另一部分是对种子节点选择算法进行研究，即确定复杂网络中的若干个节点作为初始传播节点，使其作为传播营销信息的起点，能够使得受到营销信息影响的节点数目达到最大值，也被称为影响力最大化问题。

本文首先对网络中的节点之间的影响关系进行分析，考虑到节点之间存在互不相同的影响力，提出了对节点之间影响力的度量方法，能够区分不同节点对同一节点存在不同的影响值。针对现有模型存在的缺陷，本文提出了基于阈值的元胞自动机模型用来模拟营销信息的传播过程，该模型在充分考虑节点局部拓扑结构的情况下，对节点的个性阈值进行计算，使得模型更加合理。实验证明，该模型在没有随机概率影响的情况下能够获得较快的传播速度，能够促进营销信息的传播。然后，本文针对基于网络拓扑结构的种子节点选择算法的缺陷进行研究，提出了关于节点重要性新的度量方法，进而实现了基于社区结构的种子节点选择算法，并与其他种子节点选择算法进行对比。通过实验表明，基于社区结构的种子节点选择算法具有较好的影响范围，且能够挖掘出最具影响力的节点，同时具有较低的时间复杂度。

关键词：复杂网络；病毒营销；传播模型；影响力最大化

Abstract

With the in-depth study of complex networks, viral marketing as a new marketing model has become the focus of attention. Viral marketing also known as word of mouth marketing, mainly achieve through the users of network spontaneously spread marketing messages to others. In the process of viral marketing, first of all, businesses will promote the product for individual users , when individual users get a better reputation and will recommend these merchandise to their friends, new individuals will get word of mouth and spread the product information to others in the same way in order to achieve marketing information expand the scope of dissemination. Viral marketing propagation mechanism research of complex network, mainly through the open nature of the network, making marketing message spread from a few nodes to its neighbor nodes on the basis of really complex network, the ultimate goal is to make all the complex networks node can be affected by marketing information.

The virus marketing propagation mechanism of complex network research mainly contains two parts, one is for viral marketing propagation modeling, is mainly used to simulate the propagation process of marketing information on the complex networks. The other part is the research of seed node selection algorithm, namely to determine the number of nodes in complex networks as the initial transmission nodes, make these nodes as the starting point of marketing information, to maximize the number of nodes are affected by marketing messages, also known as the influence maximization problem.

Firstly, this article analyzes the influence of the relationship between the nodes in the network, considering the influence of the node between each other is different, proposing the measure of influence between nodes, distinguishing between different nodes on the same node exist different influence values. Aiming at the flaws of the existing models, this paper presents a cellular automaton model based on threshold is used to simulate the marketing information dissemination process, the model fully consider the node topology and calculate personality of node threshold value, in this case that makes the model more reasonable. Experiments show that the model without the influence of random probability can get faster speed, could promote the spread of marketing information. Then, aiming at the defects of the seed node selection algorithm based on network topology propose the importance of node a new measurement method, and implements the seed node selection algorithm based on the structure of community,

and compared with other seed node selection algorithm.The experiments show that the seed node selection algorithm based on community structure has good influence scope, and able find the most influential nodes, and has lower time complexity.

Key words: complex networks; viral marketing; propagation model; maximum impact

目　录

第1章 绪 论

1.1 研究背景及意义

复杂网络在人们日常生活中广泛存在，社会中的每一个人都可以被看作是复杂网络中的一个节点，与其他人之间的关系看作边，组成人际关系网络；互联网则是把电脑和路由器看作网络中的节点，把网络连接看作边。目前，对复杂网络的研究已经涉及到多个领域，如：社会学、生命科学、计算机等不同领域，常见的复杂网络有社会网络，电力网络[1]、交通网络[2]、食物链网络[3-4]、神经网络、长途电话网络等。社交网络(SNS)是在 Web2.0 环境下的技术框架，即人们建立的社会性网络的互联网应用服务[5]，如 Facebook，人人网，MySpace 等社交网络逐渐改变了人们的交互方式和信息获取及传播渠道。

随着互联网的蓬勃发展，使得传统的营销模式发生了转变，越来越多的信息途径不再只专注于报纸，电视等传播媒体，人们在网络上获取的信息量逐渐增加且超过了传统媒体的信息量。在互联网上人们可以找到有购买经验且具有相同兴趣爱好的人进行讨论评价该物品或者信息，并且打破了时间和空间的限制，如在论坛上进行评论，转载、转帖等用户行为就会逐渐扩大商品或品牌的口碑传播范围，信息通过快速复制的方式进行传播和扩散，从而为更多的人成为信息的“感染者”，因此互联网上的信息传播具有病毒式的传播效应。

病毒营销是由 Steve Jurvetson 和 Tim Draper 首次提出的，并定义为“基于网络的口碑传播[6]”。病毒营销的实践最早起源于电子邮件。病毒营销用于商品，服务的推广。病毒营销是利用大家感兴趣的信息，使人们自愿加入到传播过程之中，例如当用户在某网站获得感兴趣的信息时，该用户就会把链接复制给其好友，若其好友接受其网站信息，则其好友就成为信息的“感染者”。病毒营销问题实质上是复杂网络传播动力学的问题，其目的是使社会网络中接受营销信息的用户实现最大化。

病毒营销传播机制研究主要是利用社会网络中个体之间的关系及互相影响的特点来研究信息传播问题，在实际中有极广泛的应用前景。不仅可以对营销信息的传播过程进行分析，还可以探索如何加快信息传播，例如，品牌在网络上的推广，使得该品牌商品可以获得较好的口碑，它对企业提高知名度及经济效益有非常重要的意义。另一方面，对传播机制的研究可以对危机信息得到有效的控制，例如谣言以及危机信息的传播等，分析其传播规律，可以有效地控制传播源头，对控制不良信息的传播同样具有很重要的

意义。

1.2 国内外研究现状

1.2.1 病毒营销理论研究

病毒营销是通过用户的口碑实现营销信息的传播扩散，并且营销信息是以类似病毒的传播方式进行的。随着互联网的发展，社会网络已经成为重要的营销工具，病毒营销作为有效的网络营销方法在社会网络上得到了广泛应用。

病毒营销最早起源于电子邮件，在 1996 年，Sabeer Bhatia 和 Jack Smith 创建了一个免费邮件服务即 Hotmail.com。Hotmail 在邮件的末尾处加上免费邮件的广告，用户每次发送和接受邮件时，都会将收到和转发广告信息。Hotmail 投入使用在很短的时间内，用户数量以极快的速度增长，并且制作该邮箱的公司花费的营销费用极少。Hotmail 的快速成长过程是由于其采用了病毒营销这一营销模式，它使得用户成为产品广告信息传播的主体。

1997 年，Steve Jurvetson 等人对病毒营销这一新兴的营销模式做出了定义，“基于网络的口碑传播”。

国内学者对病毒营销研究主要是对营销模式的应用研究[7-13]，文献[14]分析了在病毒营销过程中用户的传播意愿的影响因素的分析，以及对传播策略进行了讨论。陆昊菁[15]从个体的传播意愿的角度出发对病毒营销进行研究，认为传播意愿是取决于信息以及人际关系的互动，在该文章中分析了病毒营销的传播控制系统以及传播用户的行为规律。

1.2.2 影响力最大化问题的研究

在计算机领域，近年来主要是在病毒营销传播模型以及影响力最大化这两个方面研究病毒营销问题。Domingos 和 Richardson[16]根据病毒营销在网络中的实施状况，对病毒营销的研究邻域进行了扩展，认为网络中存在影响力的传播，用户行为对其邻居用户具有影响力，从而推动病毒营销在网络中推广。并从数据挖掘的角度建立了概率模型，影响一批潜在用户，再通过口碑传播去影响其他用户，实现受到影响的用户最大化。

Kemple 和 Kleinberg 在文献[17]中，通过计算边际影响值并以该影响值大小为依据选择节点，认为边际影响值越大则节点影响力越大，反之越小。但是该算法每次都需要计算未激活节点的边际影响，导致该算法在节点数目多的网络中运行困难，存在一定的缺陷。文献[18]提出了对上述贪心算法的改进算法即 CELF 算法，利用次模特性使得大量节点的边际影响不需要重新计算，提高了算法的效率。

Forrest Stonedahl 等人[19]用智能算法分析病毒营销问题，结合节点的度、平均距离以及聚集系数等网络结构的物理特征，通过遗传算法寻找这些特征的权重，实现传播最大化。

Christine Kiss 等人[20]提出一种根据网络中结点的度中心性，介数以及接近中心性等度量进行种子结点的选择。实验结果表明，中心性度量和 SenderRank 选择的种子具有较好的营销效果。

田家堂等人[21]针对贪心算法的缺陷即时间复杂度过高的问题给出了 HPG 算法，该算法是结合了 LT 模型的影响力累积特征，在算法的初期阶段，挖掘节点的潜在影响，在后期阶段中再利用贪心算法找出具有实际影响力的节点，HPG 算法获得了很好的效果。

文献[22]对独立级联模型进行了改进，结合节点影响强度随时间的变化，给出了完全级联模型的算法。在影响力最大化算法的研究过程中，文献[23-24]对群体挖掘算法有了进一步的研究，考虑到网络的节点之间的关系来研究动态和静态的核心群体。

张伯雷等人[25]给出了 KCC 算法，该算法是在用户聚类分析的基础上找出代表性节点，并且可以影响更多用户，避免用户重复接收信息。KCC 算法运行效率较高，具有可扩展性。苏晓萍等人[26]提出了用户兴趣强度的计算方法，通过邻居信息的半局部中心性来发现最具影响力的节点，该算法综合考虑了准确率和效率，使得二者达到平衡。

1.2.3 病毒营销传播模型研究

Frank Bass[27]提出了 Bass 模型，该模型主要是在市场营销领域进行分析，对新产品的扩散趋势进行预测，在总体上反应出购买商品的用户数量的曲线，而该模型并没有复杂网络的网络结构。

Yongkun Li 等人[28]利用局部平均场分析在线社交网络中的节点的状态，分析多种影响机制并且提出了预测买家最终行为的框架。

Tsung-Hao Hsu 等人[29]提出了一个混合的分类模型，该模型通过混合多重关系排名区分一个未激活用户未来是否成为激活用户，通过解决节点的分类问题发现潜在的激活用户，把信息网络分为两层，分别代表激活用户和未激活用户的社会关系，并且使用 SVM 对潜在激活用户进行分类，取得了较好的影响效果。

Jianmei Yang 等人[30]扩展了复杂网络理论，并且对病毒营销进行建模，在该模型中计算每个节点的度值，从相应度值中选取任意节点，并引入服从正太分布的激活概率计算感染节点和覆盖节点的数目。但该模型缺点是比较明显的，即每个度值都要计算感染

节点和覆盖节点，导致模型运行十分耗时，对于大型社会网络该模型运行起来会十分困难。

张彦超[31]通过分析用户的自身的特征，如自主性以及信任值等因素，对网络的用户观点交流过程进行深入描述，提出个体观点交互模型，即 AI 模型，通过实验论证，该模型获得了较好的效果。

文献[32]采用博弈理论对传播动力学进行分析，通过建立多智能体博弈模型，即 MAS 模型，该模型对不同的情景进行分析，对用户的收益进行分析进而判断个体的传播行为，并和 SIR 模型和马尔科夫扩展模型等进行对比分析。

吴凯[33]对微博网络中的个体行为进行分析和预测，主要是通过在用户的过去的行为进行分析，在历史分析的基础上，判断个体未来的行为概率。在预测模型中，对用户进行特征提取计算用户的行为概率，提出 RTP 模型。

郭进时[34]等人根据计算局部影响力和社区之间的弱纽带节点的影响力，提出 InfG 模型，得出 k 种子节点，通过实验证明该算法的取得很好的效果，影响范围在 90%以上。

1.2.4 存在的问题与不足

综上所述，对于病毒营销问题的研究已经取得了一定的研究成果，但是现有的对病毒营销问题的研究仍然存在以下问题：

（1）大多数模型都没有明确节点之间影响力的计算方法，只有少数提及节点之间影响力的计算方法，但其计算方法并没有考虑单个节点的网络特征。

（2）传播最大化问题仍然是病毒营销研究中的一个热点问题，直接关系着病毒营销的影响效果，传播范围及传播模型的效率仍然需要进一步研究。

（3）已有的病毒营销传播模型没有考虑到网络实际的拓扑结构，不能准确刻画营销信息在现实现实网络中的传播过程，尤其在当前对复杂网络的研究逐渐深入，网络的结构特征逐渐体现出来，信息传播模型存在着更多的不确定性，因此现有的传播模型不能很好地解决当前的具体问题。

（4）现有传播模型都是在某一时刻的固定的网络拓扑上进行的，没有考虑拓扑结构随时间改变的动态过程。然而现实网络是总是在一直变化的,在不同的时间段，网络结构都有可能产生变化,这时，传播模型不能准确刻画节点的传播动向，因此，在传播模型应用到现实生活中具有一定的局限性。

（5）对病毒营销的模型研究也存在不足，以前的研究一般是结合相对较早的传播模型来研究病毒营销问题，不都适合现有的研究需要，须对以前的模型进行更新和改进。

1.3 论文的主要研究内容

本文的研究工作是在借鉴前人的基础上，针对目前工作存在的一些不足，提出病毒营销传播模型，以实现传播最大化。

通过阅读和学习相关文献以及进行相关的研究工作，本文的研究内容主要包括以下几个方面：

（1）阐述复杂网络的相关理论及两个比较经典的影响力传播模型，包括独立级联模型（IC）和线性阈值模型（LT）并重点研究几种主要的影响力最大化算法，包括贪心算法、Degree discount 算法、最大度算法并且分析各算法的优缺点。

（2）结合单个节点的度和聚集系数等网络特征，提出了节点之间影响力的计算方法。节点的度是衡量节点重要程度的指标，有研究认为若节点具有相同的度，其中在节点的邻域结构中具有较低的聚集系数的节点比其邻居节点的重要程度要高。

（3）病毒营销问题实质上是复杂网络传播最大化的问题。其目的是使社会网络中接受营销信息的用户数目实现最大化，促进品牌或网站的推广。在节点之间影响力估计的基础上，提出了基于阈值的元胞自动机模型，分析网络中用户的状态，分别是处于激活状态的用户和未激活状态的用户，该模型更好的模拟了病毒营销信息传播过程。

（4）种子节点的选择对病毒营销起着重要的作用，在以往的传播模型中，往往选择网络中任意 k 个节点作为种子节点集合。在社会网络中，通常认为具有较好传播能力的节点对社会网络的信息传播具有重要影响，选择 *top* k 个传播能力最大的节点作为种子节点的集合，使其实现加速病毒营销过程且扩大传播范围的目的。

1.4 论文的组织结构

本文主要是对病毒营销的传播机制进行研究，根据现有模型存在的缺陷，提出了一种传播模型，并讨论了种子节点对病毒营销的影响，提出了基于社区结构的种子节点选择算法。

本文的具体章节安排及内容简述如下：

第 1 章，绪论部分主要论述课题的背景及其研究意义，阐述了国内外学者对病毒营销的研究进展，同时对目前研究现状存在的问题与不足进行了分析总结，最后简要介绍了本文的研究工作及论文安排。

第 2 章，介绍了复杂网络相关理论及几种目前常用的信息传播模型，详细介绍了独立级联模型（IC）和线性阈值模型（LT）以及传染病模型，着重介绍了元胞自动机模型。同时对种子节点选择算法进行了详细介绍，如贪心算法，最大度算法等。

第 3 章，对营销信息的传播过程进行建模，即提出了基于阈值计算的元胞自动机模型，在分析个性阈值的基础上建立元胞自动机模型，该模型把网络中节点分为三种状态，即分为激活节点和未激活节点以及处于移除状态的节点，并且根据节点之间的影响力计算得出每个节点的阈值，进行传播过程的模拟。使得网络中的所有节点都能够受到信息感染，使信息得到最大范围的传播。

第 4 章，针对现有的种子节点存在的缺陷进行分析，提出了基于社区结构和网络局部特征的混合种子节点选择算法，以解决最大度算法中重叠邻接节点的缺陷，并通过实验证明其有效性。

第2章 复杂网络的相关理论

本章首先详细阐述了复杂网络的基本概念，从图论的角度进行综合介绍， 同时还详尽介绍了网络结构特征的相关度量以及现有的传播模型。

2.1 复杂网络的基本概念

复杂系统在人们的生活中随处可见，如电力系统，生态系统，交通系统等，这些系统都可以通过网络进行刻画和分析，而复杂网络则是对复杂系统研究的一种方法,随着对复杂网络的深入研究，人们逐渐发现其存在的特征。复杂网络[35]具有小世界和无标度特性以及自组织、自相似等特性。随着计算机技术的发展，在Web2.0时代背景下，社交网站得到了蓬勃的发展，社交网站在满足六度分隔原理的基础上建立的,因此社交网络是比较典型的复杂网络。

复杂网络的结构虽然复杂，但它可以用图论来刻画，把网络抽象为由节点和边组成的图$G(V,E)$。图中的节点集合V代表参与到网络中的个体元素的集合，该集合表示是网络中的所有成员，即包含所有节点。E中的元素代表网络中的边，代表个体元素之间的存在的联系。$N=|V|$代表节点的数目，$M=|E|$代表网络中边的条数。同时，也可以使用邻接矩阵$A=\{a_{ij}\}_{n\times n}$描述复杂网络，n代表节点的总数，当$a_{ij}=1$时说明节点v_i与v_j存在直接相连的边，当$a_{ij}=0$时说明节点v_i与v_j不存在直接相连的边。

2.2 复杂网络的结构特征

人们在对大量现实网络的研究过程中，发现网络显示出许多结构特性，例如：节点的度、聚集系数、平均路经长度、介数以及社区结构等。

1.度

度是指节点与其他节点存在直接相连的边的数目，在图 G 中，节点v_i的度记为$d(i)$，其邻接矩阵记为 A。在有向图中，节点的度分为出度和入度，入度是指以节点邻居为起点连接到该节点有向边的条数；出度是指以节点为起点向其邻接节点发出连接的边的条数。在复杂网络中，度用邻居节点的数量来描述节点的重要程度，节点的度越大，说明其影响其邻居节点的程度越大，即节点的重要程度越大，反之则节点的重要程度就越小。复杂网络中的平均度是即对网络中所有节点的度之和与节点总数之比，记为$\overline{d}$，平均度用来对网络中节点连接的密集程度进行度量。节点v_i的度与平均度的定义如下：

$$d(i)=\sum_{j}a_{ij} \tag{2-1}$$

式中 a_{ij} 是网络图的邻接矩阵 A 中 (i, j) 相对应的元素。

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d(i) \tag{2-2}$$

对于图 $G(V,E)$,度分布是用来描述复杂网络的最基本的结构特征，若节点度的值为 k ，即 $d = k$ ，则度分布记作 $P(k)$ 代表网络中度为 k 的节点在网络中所占的比率，同时也求得节点度的值等于 k 的概率。

2.聚集系数

Watts 和 Strogatz 在 1998 年首次提出了聚集系数的概念，聚集系数是用来刻画网络的特性，表示节点邻域内的密集程度，是描述复杂网络的集群化程度的一种度量，研究结果表明在大多数复杂社会网络中节点有集群化倾向，这种倾向的特征是连接相对紧密。对复杂社会网络来说，这种集群现象是网络中的朋友社交圈或者同一社团在网络结构中所表现出来的，并且集群中的人员互相熟悉。聚集系数表现在现实生活中，可以看作为你的朋友与其他朋友之间熟悉的可能性。例如，A 节点和 B 节点有连接，B 节点和 C 节点有连接，聚集系数越大，A 节点和 C 节点有连接的倾向越大，聚集系数分为全局和局部聚集系数。

全局聚集系数的计算是以三角形的数目统计为基础的，设在图 $G(V,E)$ 中，$V = \{v1, v2, ..., vn\}$ 表示在集合 V 中有 n 个节点，$E = \{e_{ij}, 1 \le i \le n, 1 \le j \le n\}$（ e_{ij} 表示节点 v_i 与 v_j 相连接的边），若节点之间是两两相连的，那么则在网络图中形成了三角形，即由三个节点构成的三元组，若其中对应的三个节点之间存在互相连接的边，则称为闭三元组，开三元组则被定义为三个节点之间有两条边，即为缺少一条边的未闭合的三角形，开三元组和闭三元组都可以被认为是连通三元组。全局聚集系数被定义为一个网络图中闭三元组的数目与连通三元组的数目之间的比值，如下所示：

$$C = \frac{3 \times S_1}{S} \tag{2-3}$$

其中， S_1 为闭三元组的个数， S 为连通三元组的个数， C 为聚集系数。

网络图中的局部聚集系数主要是衡量节点的邻居节点之间相互连接的紧密程度。假设在有向图 $G(V,E)$ 中，节点 v_i 有 $d(i)$ 条边，即节点 v_i 的出度和入度之和为 $d(i)$ ，设 N_i 是节点 v_i 的邻居节点的集合。在有向图中，e_{ij} 与 e_{ji} 是两条不同方向的边，易知节点 v_i 的邻居节点之间最多可以有 $d(i)(d(i)-1)$ 条边，此时，节点 v_i 的聚集系数 C_i 则被定义为 v_i 的邻居节点之间实际存在的边的条数与最多可能存在的条数 $d(i)(d(i)-1)/2$ 之比，即局部

聚集系数。如下所示：

$$C_i = \frac{|\{e_{jk}\}|}{d(i)(d(i)-1)} \tag{2-4}$$

其中 $e_{jk} \in E$，v_j，$v_k \in N_i$。

在无向图中，由于 e_{ij} 与 e_{ji} 表示相同的连接，因此在无向图中，节点 v_i 的邻居节点之间可能存在最多的边为 $d(i)(d(i)-1)/2$ 条边。故节点 v_i 的聚集系数如下所示：

$$C_i = \frac{2|\{e_{jk}\}|}{d(i)(d(i)-1)} \tag{2-5}$$

其中 $e_{jk} \in E$，v_j，$v_k \in N_i$。

平均聚集系数是以每个节点的聚集系数为基础衡量整个网络的聚集程度，如下所示：

$$C = \frac{1}{n}\sum_i C_i \tag{2-6}$$

其中 n 为网络中包含节点的个数，C_i 为节点 v_i 的聚集系数。

3.平均路径长度

平均路径长度是复杂网络中的又一重要度量，在现实的网络中有着十分重要的作用，该度量往往能够衡量网络的规模。通常用矩阵 D 来刻画整个复杂网络的所有节点之间的最短路径，其中 d_{ij} 代表节点 v_i 到节点 v_j 所需的最少跳数，即最短路径。平均路径长度是对复杂网络中实际存在的最短路径和任意节点之间可能存在的最多最短路径数目之比，平均路径长度记为 L，如下所示：

$$L = \frac{1}{n(n-1)}\sum d_{ij} \tag{2-7}$$

其中 n 为网络中包含的节点的个数。

4.介数

介数[36-38]分为边介数和节点介数，用来描述单个节点或某条边在网络结构中的重要程度和影响力，描述的是节点或边在网络中的繁忙程度。其中边介数被定义为网络中经过该边的最短路径数目所占的比例，节点介数是指在网络中经过该节点的最短路径数目所占的比例。介数还可以控制节点之间信息传播的能力，具有最大介数的节点在很大程度上对其他节点之间的信息传播具有决定性作用，若去掉该节点，则会使节点之间的最短路径变长，从而可以达到控制信息传播的作用。

5. 社区结构

许多研究表明，现实中的社交网络是由社区[39-44]所构成的，即由许多节点组构成的，

社区结构已经成为对复杂网络进行研究的重要度量参数。在社区内部，两个节点之间比不同社区的两个节点更有连接的倾向，因此社区内部的节点具有较高的聚集程度，社区之间的关系连接相对较少，描述了网络结构的局部特性及连接分布的不均匀程度。在实际生活中，按照职业，爱好，习惯等把人们分为不同的群体，即社区代表具有某个共同特点的团体。例如，在科学家合作网中，该网络描述了科学家之间的相互合作，在同一社区中表示研究方向或领域相同或者相近，即同一社区网络代表了相同研究领域的合作关系。社区结构这一度量对挖掘复杂网络中非正式团体有重要意义。

同时，复杂网络还存在着其他结构特征，如节点度相关性[45]、分层结构[46]、网络自相似性[47]等。

2.3 复杂网络上的传播模型

近些年，对信息在复杂网络上传播过程的研究主要是对传播模型进行研究。若把社交网络上的一个用户看作一个节点，则整个社交网络可以看作是一个复杂网络，信息在复杂网络上传播，可以是病毒的传播蔓延，营销信心在网络上的传播，创新产品在社会网络上的传播等。信息传播模型刻画了信息传播过程，即在首先在网络中选择若干个种子节点，种子节点是已经激活的节点，种子节点具有很强烈的传播倾向且作为传播的起点。其次是激活节点把信息传播给其邻居几点，例如在微博上，节点可以通过它所关注的其他节点获取信息，而人人网，脸谱网，每个节点都可以看到其他邻居节点的信息，如转发的信息，邻居编辑的信息或者广告信息等。

下面主要介绍几种重要的传播模型，如独立级联模型、线性阈值模型、元胞自动机模型、传染病模型以及其他模型等。

2.3.1 独立级联模型

首先将网络用有向图 $G(V,E)$来描述，E 表示有向图中边的集合，边代表节点之间的关系，如在人人网中朋友关系，新浪微博中的用户关注关系，科学家合作网的合作关系等。V 表示节点的集合，节点表示在网络中个体成员，如在互联网上的计算机，社交网络上的用户等都可以看作网络中的节点。

独立级联模型[48]是由 Goldenberg 等人提出来的，是以概率论为基础的为相互粒子系统建立的一个信息传播模型。在该模型中，首先给定种子节点集合 A_0，该集合是在网络中选择若干个节点作为集合 A_0 的元素，以 A_0 为传播的起点，网络中的传播信息都是由种子节点按照独立级联模型的传播规则进行的。假设 p_{uv} 代表节点能够成功激活其邻

居节点的概率。根据上述给定条件，独立级联模型主要分为以下几个步骤：

（1）选定种子节点，t=0 时刻时，网络中激活节点的个数即是种子节点集合 A_0 的元素个数。在 t=1 时刻时，种子节点以概率 p_{uv} 激活其邻居节点，即其邻居节点以概率 p_{uv} 转化为激活节点。

（2）在 IC 模型中，节点的激活顺序是随机的，当节点 A 为未激活节点，其邻居节点 B 和 C 则可采用任意顺序激活节点 A。若在 t 时间步时，如果节点 u 的邻居节点 v 被激活，则在 t+1 时刻，v 则成为激活节点。

（3）重复以上过程，直到没有新的节点被激活。在 IC 模型中，节点 u 激活节点 v 的 p_{uv} 与其他节点试图激活节点 v 的概率无关，即激活概率是独立的。

2.3.2 线性阈值模型

线性阈值模型[49]是由 Granovetter 等人提出的，该模型是以阈值理论为基础的。在有向图 $G(V,E)$中，在 LT 模型中，给定某一节点的所有邻居节点对该节点的影响力之和都小于等于 1，每个节点都有一个个性阈值θ，所有节点的阈值都从(0,1)中任意选取，认为在网络中节点对其每个邻居节点都存在一个影响力，当未激活节点受到其已经激活的邻居节点的累积影响超过个性阈值时，未激活节点就会变为激活节点，并在下一刻以同样的方式作用其邻居节点。假定有向图的邻居节点规定为入边邻居，节点 v 的邻居节点的集合记为 $N(v)$，其中 $N(v)$中激活节点的个数为 m，记 m=$|A(v)|$，$A(v)$是 v 的邻居节点中激活节点的集合。若节点 v 的入边邻居 u 对其有影响力 b_{uv}，若满足如下式（2-8）条件时，则节点 v 转变为激活节点，否则，节点 v 仍是未激活节点，但是节点 u 对节点 v 的影响力被累积起来，这样，b_{uv} 对节点 v 转化为激活节点起到了“影响力累积”的作用，使节点 v 成为激活节点的可能性大大增加。

$$\sum_{u\in A(v)} b_{uv} \geq \theta_v \tag{2-8}$$

线性阈值的传播过程如下：

（1）首先给定初始激活节点的集合 A_0 作为信息传播的起点，给定网络中节点的阈值 θ_v 以及节点 v 的邻居节点对其的影响力 b_{uv}。

（2）满足条件（2-8）时，节点 v 转化为激活节点，在时间步为 t-1 时，激活节点的集合为 $A_{t\text{-}1}$，在第 t 时间步时，在集合 $A_{t\text{-}1}$ 的基础上激活新的节点，得到集合 A_t。

（3）重复第（2）步，直到没有新的激活节点加入集合 A 中。传播过程结束。

2.3.3 元胞自动机模型

1.元胞自动机定义

元胞自动机，（又称为细胞自动机、点格自动机）简称 CA，最早由冯. 诺依曼依据自然界的自我复制原理的基础上提出的，本质上是一种空间和时间上离散的动力学系统，元胞自动机是用简单规则模拟复杂系统。目前元胞自动机可应用于多种领域，例如复杂系统科学，通信科学，社会学，数学等各个领域。元胞自动机最早的应用模型是“生命游戏”，该模型采用的是八邻域，描述的是单个元胞的生死由其邻居元胞的生死状态来决定，当邻居元胞中存活的元胞个数是 3 时，中心元胞为生，邻居元胞有 2 个元胞存活时，中心元胞生死状态不变。由“生命游戏”可以看出多个元胞可以演化出复杂的动态网络，依据相应的状态转换规则可以模拟出许多复杂行为。

2.元胞自动机的组成

最基本的元胞自动机包括四个部分，即包括元胞、元胞空间、邻居及规则。可以把元胞自动机看做是在元胞空间上的元胞依据一定的状态转换规则不断更新元胞状态的演化过程。

（1）元胞

元胞自动机的基本组成单位是元胞，是分布在欧几里得空间上的晶格上的，是 CA 模型中最基本的元素。若用元胞自动机模拟社交网络，则网络上的个人或者用户可作为一个元胞；在社会网络中，把个体成员看作是一个元胞；在化学中，可把原子或分子看作一个元胞。

（2）元胞空间

元胞空间是指元胞所在的空间网点集合，元胞空间具有多样性，可以呈现出多种形式[50]，如三角形、正方形、六边形的点格空间排列，图 2.1 和图 2.2 显示了一维和二维元胞的空间的的几种形式，其中一维元胞空间只有图 2.1 一种表现方式。

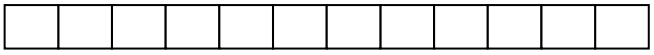
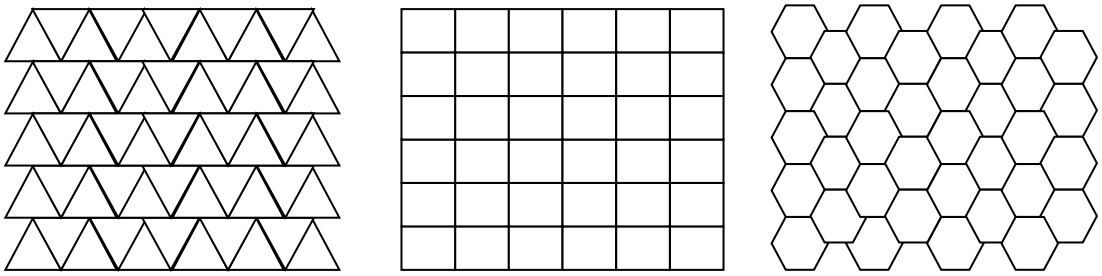


图 2.1 一维元胞空间

在二维元胞自动机中，这三种元胞空间中最适宜现代计算机描述的是四方形元胞空间，主要是由于其具有直观的优点；而三角点格不利于描述，应用最少；六边点格空间邻居节点相对较多，比前两种元胞空间描述的更加复杂，但是很好地对物理或化学中各向同性问题进行描述。



(a)三角点格　　(b)四方点格　　(c)六边点格

图 2.2 二维元胞空间的多种形式

由于在实际使用 CA 模型时，我们只能对有限的系统进行模拟，因此元胞的个数是有限的，这要求元胞空间的边界需要按照一定的规则进行设置。对于边界上的元胞，由于其所处位置的特殊性，应对边界元胞的演化行为制定区别于普通元胞的转换规则。元胞空间的边界[51-52]主要有以下几种：

① 定值型边界：在元胞自动机中所有的边界元胞均采用常量作为界值。

② 周期型边界：对于一维空间，周期型边界则表现为左右边界是相互连通的，形成了一个环；对于二维元胞自动机，周期型边界的元胞空间表现为上下相对的边界元胞相互连接，左右相对的边界元胞也是相互连接的。周期型边界使元胞空间是无限的，其中元胞不断重复。

③ 随机型边界：在元胞自动机运行时，都对处于边界的元胞给定一个随机值。

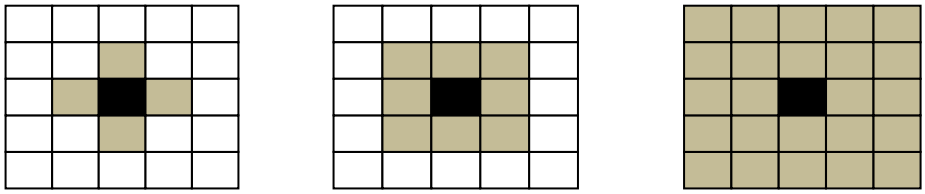
④ 反射型边界：在空元胞中复制边界里面的邻居元胞，相当于以边界元胞为对称轴，左右两侧的元胞的值是相同的，故也成为对称型边界。

（3）元胞状态

元胞的状态则根据实际情况划分，元胞自动机的离散状态集包含元胞的所有状态，即该集合是由有限个状态值组成。例如在生命游戏中，元胞的有两种状态，即生死状态。在其他模型中也可以采用包含多个状态的状态集，例如可采取 $\{s_0, s_1, \cdots, s_n\}$。

(4) 元胞的邻居

元胞的邻居则是状态改变时需要搜索的空间域。模型中的每个元胞都依据同一个状态转换规则，在局部范围内进行更新元胞的状态。简单地说，通过简单构造规则就能够实现复杂系统和非线性科学的动态演化，而不需要复杂的数学方程。若模型中包含以上几部分，则可认为是元胞自动机模型。元胞自动机通常二维元胞自动机通常有如下几种邻居：



（a）　　（b）　　（c）

图 2.3 元胞自动机的邻居

元胞自动机如图所示，在图 2.3（a）所描述的是冯·诺依曼型邻居，元胞总共有上下左右四个邻居。图 2.3（b）描述的是摩尔型邻居，元胞的邻居分别位于元胞的八个方向上，共有八个邻居，而图 2.3（c）描述的是扩展的摩尔邻居，共有 24 个邻居。以上三种邻居是常用的元胞自动机的邻居模型。

（5）规则

规则即指状态转换规则是元胞自动机的核心部分，主要是为元胞状态转换提供依据，即在任意元胞以及邻居元胞的当前状态的基础上确定该元胞下一时间步的状态。状态转换规则可用转移函数来表示，记为 f。由于其元胞状态取决于自身状态及邻居元胞状态，因此状态转换规则具有局部性。

（6）元胞自动机

元胞自动机对系统动态模拟是在时间离散且时间间隔相等的基础上运行的，假设时间间隔为$\triangle t$,令 t_0 为初始时间步，则 $t_1=t_0+\triangle t$，…，$t_m= t_0+m\triangle t$，其中 m 为整数。

由以上论述可知，CA 模型可用四元组来表示：$A=(C_d，S，N，f)$，其中 C_d 代表元胞空间，d 是维数；S 是元胞自动机的元胞状态集；N 代表所有邻域的元胞；f 代表状态转换规则。

标准的元胞自动机具有同质性，离散性包括时间离散及空间离散，同步计算和局部时空性等特征。

2.3.4 经典传染病模型

在最早研究复杂网络时，主要是针对流行病等疾病以及信息的传播过程或传播原理进行研究。应用比较广泛的且传染病模型包括 SI，SIS，SIR 这三种模型，下面主要是对 SIR 型进行简要介绍。

SIR 模型是由 Bailey N[53]提出的。在 SIR 模型中，考虑了感染个体恢复健康后拥有免疫能力，处于恢复状态的个体不能够把疾病传染给易感染状态的个体，同时也不会被感染个体传染。在 SIR 模型中，复杂网路中的节点分为三种状态：S 为易感染状态，I

为感染状态，R 为恢复状态。若易感染节点的邻居中存在感染节点，则处于易感染状态的节点将会依一定概率转化为感染节点；对于感染节点则会以一定概率转化为恢复状态的节点，即不具有传播疾病的能力且不会被疾病再次感染，但在感染节点成为恢复状态之前，感染节点则对其邻居节点传播疾病。在复杂网络中，可以用式（2-9）描述 SIR 模型：

$$\begin{cases} \dfrac{ds}{dt} = -\lambda < k > is \\ \dfrac{di}{dt} = \lambda < k > is - \gamma i \\ \dfrac{dr}{dt} = \gamma i \end{cases} \tag{2-9}$$

其中λ为感染节点与易感染节点之间的感染率，γ是感染节点的恢复率，s 是易感染节点在复杂网络中总节点所占的比率，i 是感染节点在复杂网络中占节点总数的比率，r 是处于恢复状态的节点在复杂网络中所占的比率。

SIR 模型克服了 SI 模型和 SIS 模型的缺点，对传染病的影响因素进行了进一步分析，节点的状态更加符合实际。在 SI 模型中，没有考虑处于易感染状态的节点能够治愈的状况，跟实际的流行病传播过程有较大差距，因此 SI 模型存在局限性。而在 SIS 模型中，虽然考虑到节点的治愈状况，但其直接恢复到易感状态，没有考虑到免疫状态。因此，相对 SI 模型和 SIS 模型，SIR 模型有一定的优越性。

在真实复杂网络上的流行病传播过程中，感染状态，易感状态，恢复状态，都精确地对网络中的节点进行了分类；各种状态的节点的分布和比率，以及不同状态的节点之间的联系程度，都是影响流行病传播范围的影响因素。随着对流行病传播模型逐渐深入研究，学者们提出许多相关的其他模型，如 SIRS 模型[54]，以及相关概念，如在文献[55]中提出了“潜伏节点”用来描述处于潜伏状态的节点。

2.3.5 其他传播模型

AggarwaL 等人提出了统计模型，其描述信息传播过程如下：给定复杂网络中的种子节点的集合 A_0, 即集合中的所有节点都为激活节点，所有信息都从种子节点集合中传播出去，其他节点的传播概率是由其邻居节点的传播概率决定的，节点的传播概率用式（2-10）来描述：

$$\pi(i) = \begin{cases} 1 & i \in A_0 \\ 1 - \Pi_j (1 - \pi(j) t_{ji}) & i \notin A_0 \end{cases} \tag{2-10}$$

其中 $\pi(i)$ 为节点 i 的信息传播概率，t_{ji} 是节点 j 与节点 i 的传导的可能程度。

此外，博弈论模型[56-57]主要针对用户根据自身的利益是否能够达到最大化来决定用户的策略，即用户是否接受邻居节点的传播信息，若接受邻居节点的信息能够使节点本身的效益达到最大，则该节点采取接受邻居节点的策略作为最佳选择，相反若是达不到最大效益则采取不接受邻居节点的传播信息。

因此，博弈论模型与其他模型不同，节点不需要与阈值进行对比，而是根据自身利益最大化为衡量标准。

此外，其他信息传播模型还有基于热力学的扩散模型[58-59]、BASS 模型、递减独立级联模型、完全级联模型、投票模型，以及 SIRS 模型等。

2.4 种子节点选择算法

2.4.1 贪心算法

在文献[15]中 Kempe 等人详细介绍了经典的爬山贪心算法，并证明该算法保证了（1-1/e）近似最优的。目前有很多学者提出了对其改进的算法，如 CELF 算法，HPG 算法等。该算法是在全局网络上寻找最具影响力的节点。假设复杂网络 *G(V,E)*, 找出最具有影响力的 Top k 个节点，S 表示种子节点的集合，初始化为空集，$I(S)$表示在种子节点集合为 S 的条件下，网络中受到影响节点的数目。

在算法 2.1 的描述中，可以看出贪心算法是每一次都是选择边际影响节点数目最大的节点加入种子节点集合 S 中，算法中 $\max\{I(S\cup v)-I(S)\}$表示最大的边际影响值，在贪心算法中，其缺点也很明显即不断计算边际影响，造成该算法运行起来十分耗时，在大型网络中不适用，贪心算法不具有可扩展性。

算法 2.1 贪心算法

算法 2.1 贪心算法(G,k)
输入：G,k
1:初始化：$S=\varnothing$
2:for i=1 to k
3:　for $v\in V\backslash S$
4:　　$v=\max\{I(S\cup v)-I(S)\}$
5:　end
6:　$S=S\cup v$
7: end
输出：S

对于贪心算法的耗时问题，文献[16]介绍了一种对贪心算法的改进算法，即 CELF

算法，该算法利用次模特性，大大减少了种子节点集合影响增量的计算量。对于次模特性的概念，假设集合 A 是集合 B 的子集，即 $A \subseteq B$，对于函数 $f(x)$, 存在 $f(A+e)-f(A) \geq f(B+e)-f(B)$，则成为函数 $f(x)$是具有次模特性的。在该算法中次模特性指的是，种子节点集合 S 中的数目越少，则加入节点 v 时，获得的影响边际值越大。在 CELF 算法中，假设 $m = I(S \cup v) - I(S)$，对网络中的节点按 m 值进行排序，只对 m 值大的节点进行计算，如果 m 最大值的节点更新后 m 仍然是最大，利用次模特性，其他节点重新计算的 m 值不会对 m 的最大值有所改变，所以对其他节点省略了重新计算 m 值的步骤，因此 CELF 算法大大提高了算法的运行时间效率。

对于影响力最大化问题，除了上述论述的贪心算法及其改进算法，还有对网络拓扑结构对网络中节点影响力的评估可以应用到影响力最大化问题中，主要是对网络的结构特征进行分析。

2.4.2 最大度算法

度是复杂网络拓扑结构中的重要概念，也是判断一个节点网络中是否具有重要性的常用指标。在复杂网络中，处于中心位置的节点往往具有较高的度数。最大度算法是基于网络拓扑特征的启发式算法。在该算法中，首先要计算出网络中所有节点的度，并找出前 k 个度值最大的节点作为初始激活节点，同时作为影响力最大化算法的一种，同样要制定一种传播模型，依照传播模型进行影响力的传播，最后统计被影响的节点数目。

最大度算法的伪代码如下：

算法 2.2 MaxDegree 算法

算法 2.2 MaxDegree(G,k)
输入：*G(V,E)*
1:初始化：*S=ø*
2:for *i*=1 to *k*
3: for each *v*∈*V**S*
4: v=maxDegree(*v*)
5: *S=S*∪*v*
6: *V=V-S*
7: *i=i*+1
8: end for
9: end for
输出：S

最大度算法的明显的优势是算法运行时间快，时间复杂度相对较小，该算法在较大的复杂网络也能取得较好的时间效率。该算法的缺点是没有考虑到网络的变化及传播过

程，因此不能获得理想的效果，通常具有不稳定性，不能确定最终影响结果是最优的。

2.4.3 DegreeDiscount 算法

在文献[60]中 Wei Chen 阐述了 DegreeDiscount 算法, 该算法属于启发式算法，与最大度算法不同的是，最大度算法中是以度值为依据进行种子节点的选择，在 DegreeDiscount 算法中是以贡献大小为依据进行种子节点的选择。贡献即是指节点能够激活的节点的数目多少，节点能够激活的节点数目越多则贡献值越大，反之，则越小。例如，一个激活节点中其邻居节点也为激活节点，则该节点能够激活的节点数目为 0，该节点贡献度为 0。在初始条件下，一个种子节点的邻居节点没有其他种子节点，则该种子节点的贡献值是种子节点的度值；若种子节点的邻居节点包含其他种子节点，则该种子节点的贡献值是度值减去其邻居中种子节点的数目。

在算法 2.3 中，t_v 表示节点 v 的邻居节点中激活节点的个数，d_v 为节点 v 的度，dd_v 为节点 v 的贡献值，p 为激活概率。由于该算法考虑到最大度算法的邻居重叠的情况，按贡献度大小选择种子节点在一定程度上解决了该问题，因此该算法获得比最大度算法较好的传播效果。算法的时间复杂度为 $O(k\log n+m)$，其时间复杂度远低于贪心算法。算法描述如下：

算法 2.3 Degree discount 算法

算法 2.3 DegreeDiscount(G,k)
输入: (G,k)
1: 初始化：$S=\varnothing$
2: for $v \in V$
3: 计算 d_v
4: $dd_v = d_v$
5: 初始化：$t_v=0$
6: end
7: for i=1 to k
8: u=argmax$\{dd_v \mid v \in V \setminus S\}$
9: $S=S \cup \{u\}$
10: for neighbor v of u and $v \in V \setminus S$
11: $t_v = t_v + 1$
12: $dd_v = d_v - 2t_v - (d_v - t_v)t_v p$
13: end
14: end
输出：S

2.4.4 其他算法

除上述的影响力最大化算法中，随机算法是最简单的算法，主要思想是随机地在网

络中选择 k 个节点作为种子节点，从这 k 个节点开始向其他节点传播信息。该算法并没有考虑节点在网络中的影响力，则该算法不能保证能够取得较好的传播效果。还有基于网络社区的影响力最大化算法，如 OASNET 算法[61]，该算法把种子节点分散到各个社区中，获得了较好的影响范围。此外，CGA 算法[62]基本思想与 OASNET 算法相同，都是把种子节点分散到不同社区，但是社区内部寻找种子节点是通过贪心算法实现的。

2.5 本章小结

在本章中主要是对复杂网络的结构特征进行介绍，以及在复杂网络上的几种主要的传播模型，在第一节主要阐述了复杂网络的概念，如复杂网络的小世界特性等，还说明了社会网络是复杂网络的理论依据，同时还介绍了用图论来描述复杂网络。第二节主要介绍了复杂网络的结构特征，如度、聚集系数、平均路径长度和介数等以及这些度量参数的衡量方法。第三节主要阐述了在复杂网络上的几种主要的传播模型及其传播过程，如独立级联模型，线性阈值模型，以及传染病模型等，在第三章中的对比实验中则是与独立级联模型对比。

第3章 复杂网络病毒营销传播模型

3.1 引言

传播模型是复杂网络的病毒营销问题研究的一个热点，传播模型主要是用来模拟信息在复杂网络上传播的过程。在病毒营销的传播过程中，从种子节点集合开始传播，其中网站或商品的微博，或者需要成本的网络营销节点等作为种子节点，在复杂网络中，把接受营销信息的节点状态认为是处于激活状态，未接受营销信息的节点状态认为是未激活状态。病毒营销的目的是把营销信息传播给网络中的所有节点，复杂网络中处于激活状态的节点试图把营销信息传播给其未激活的邻居节点，使其成为激活节点，随着传播范围的逐渐扩大，越来越多的节点成为激活节点，从而达到病毒营销的目的。

现有的传播模型主要存在以下问题，首先，在传播模型中只能反映整体网络的传播变化，而不能够刻画节点之间的交互过程。其次，在以往的传播模型中传播概率是给定的参数，具有很强的随机性，不能够描述实际的传播过程。第三，在许多传播模型中没有提出有效的节点之间的影响力的估计方法。本章主要是针对现有传播模型存在的以上问题进行了研究，提出了节点之间影响力的的计算方法，进而计算个性阈值，提出了基于阈值计算的元胞自动机模型，分析网络中节点的状态，分为激活节点和未激活节点以及处于移除状态的节点，根据节点之间的影响力计算得出每个节点的阈值，建立元胞自动机模型，进行传播过程的模拟。通过实验结果表明，本文提出的模型明显比现有模型具有优越性。

3.2 节点之间的影响力估计

目前，节点之间的影响力估计主要根据以下三个方面得出：

（1）设置（0.1）范围内的随机数。

（2）根据节点的度计算出 b_{ij}。

（3）根据边上的权值得出节点之间的影响力。

在以往线性阈值模型中, 节点之间的影响力通常用节点的度进行计算，通常用$1/d_j$计算节点i对节点j的影响力b_{ij}，b_{ij}即节点 i 对节点 j 的影响力，d_j代表节点 j 的度。$1/d_j$说明节点 j 的所有邻居节点对其的影响力均是相等的，忽视了其邻居节点的差异性，并不符合实际。b_{ij}通常用如下公式计算：

$$b_{ij}=\frac{1}{d_j}, j\in N(i) \qquad (3\text{-}1)$$

在本小节中，根据网络节点的局部拓扑结构得到更加合理的 b_{ij} 估计值，分别考虑有向图和无向图两种情况。

3.2.1 无向图的 b_{ij} 估计

节点的度是衡量节点重要程度的关键指标，我们认为相同节点的度，具有较低的聚集系数的节点则认为该节点对邻居节点具有更大的影响力，本文如下定义 b_{ij} 。

定义 1，b_{ij} ，b_{ij} 是节点 i 对邻居节点 j 的对其的影响力大小，见式（3-2）

$$b_{ij} = \frac{d(i)}{d(i)+d(j)}.(1-c(i)) \tag{3-2}$$

其中 $c(i)$ 为节点 i 的聚集系数。$d(i)$ 代表节点 i 的度。

图 3.1 给出一个社会网络的子图，灰色节点 5 和灰色节点 8 互为邻居，节点 5 的度为 3，节点 8 的度为 4，计算两个节点的聚集系数，根据式（3-2）计算得到 $b_{58}=0.429$ 。$b_{85}=0.475$ 。

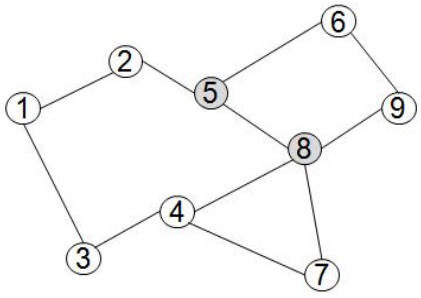


图 3.1 无向图的 b_{ij} 计算举例

3.2.2 有向图的 b_{ij} 估计

有向图的节点之间的影响是单向的，节点的度分为入度和出度，d_{outi} 表示节点 i 的出度，d_{inj} 表示节点 j 的入度。节点的出度越大，表示该节点影响其他节点的数目越多，越少则相反。节点的入度越大表示影响该节点的邻居节点就越多。聚集系数越大则越降低节点的中心性及影响程度，则有向图的节点之间的影响力是由节点的出度和入度及聚集程度决定的。

b_{ij} 的定义如下：

$$b_{ij} = \frac{d_{outi}}{d_{outi}+d_{inj}}(1-c(i)) \tag{3-3}$$

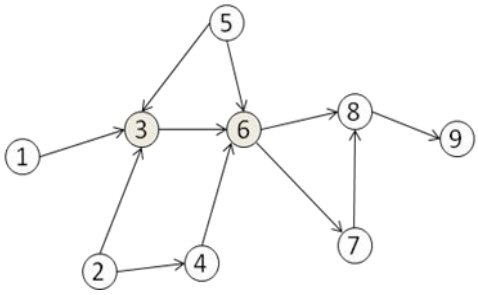


图 3.2 有向图的b_{ij}计算举例

图 3.2 中公 9 个节点组成了一张网络子图，节点 3 的出度为 1，节点 6 的入度为 3，计算节点 3 和节点 6 的聚集系数，根据式(3-3)得出 b_{36}=0.333。

3.3 基于阈值的元胞自动机模型

3.3.1 模型的提出

元胞自动机最大的优点是简单、直观，可以用简单规则对十分复杂的系统进行模拟。元胞自动机是一个离散的动力学系统，具有时间离散性，空间离散性以及状态离散性等特点，同时还具有局部性。其中元胞自动的局部性可以有效的描述节点与其邻域节点的交互过程，克服了基于平均场的微分方程的缺陷，即不能够很好的描述信息传播过程中节点之间的交互行为，因此用平均场方法描述病毒营销传播过程具有一定的局限性。元胞自动机能够实现使用简单规则描述复杂系统，是描述复杂网络传播动力学的有效方法。文献[63]表明采用元胞自动机模拟传染病模型能够得到很好的效果。

在线性阈值模型中，把每个节点的阈值设置成为同一个数值，没有考虑节点所在的具体的拓扑结构，因此具有其局限性。因此本小节在 3.2 节计算 b_{ij} 估计的基础上，结合不同节点的个性阈值进而提出了基于阈值的元胞自动机模型 CABT(The cellular automata model based on threshold)。

3.3.2 CABT 模型实现

在第 2 章中，详细介绍元胞自动机的组成，一般元胞自动机的组成主要由四个元素组成，即元胞空间，元胞邻居，状态集合以及转换规则。本文提出的基于阈值的元胞自动机模型由五个元素组成，除上述四个元素之外，增加了 T 集合，用来表示激活节点的集合，能够更加清晰地描述模型在一定时间步内有哪些节点被激活以及激活节点的数目。在 CABT 模型中，还重新对元胞的邻居进行了重新定义。由上述论述可知，CABT 模型可以看成如下五元组：

$$D=(C,S,N,T,f) \tag{3-4}$$

(1) C 代表元胞空间：复杂网络中的所有节点构成了元胞空间，每个节点对应一个元胞。

(2) S 代表元胞状态： S={-1，0，1}，$S_t(i)$表示节点 i 在 t 时刻的状态。在本文中，0 表示节点处于消息未覆盖的状态，即邻居节点中没有激活节点；1 表示节点处于激活状态，即节点获得消息并主动向其邻居节点发送消息；0 表示未激活节点，即该节点的邻居节点中有激活节点；-1 表示节点处于移除状态，即节点的邻居皆为激活节点，该节点此时不需要加入传播过程。本文定义 t 时刻代表第 t 次迭代。

(3) N 代表元胞的邻居：传统的 CA 模型把空间距离定义为半径来确定元胞的邻居，并且用网格来描述邻域，如冯. 诺依曼型及摩尔型邻居等。但在实际的复杂系统和复杂网络中，排列为标准网格的元胞空间很少，多与实际系统和网络结构的分布不符，对于实际的网络使用网格邻域进行模拟具有很大的局限性，因此，本文重新定义元胞自动机的邻居规则，即定义为若节点之间存在边的关系每个节点代表一个元胞，在网络中与节点 i 相连接的节点是 i 节点的邻居，即 $A(i.j)=1$。本文采用有向图来描述社会网络，$N(u)$ 节点 u 的入邻居集合，$O(u)$表示节点 u 的出邻居的集合。

(4) T 代表激活节点数目的集合。

(5) f 代表元胞自动机模型的状态转换规则：

$$S_{t+1}(j)=\begin{cases}0,\ S_t(j)=0,\sum\limits_{i\in N(j)}b_{ij}<\theta_j,S_t(i)=1\\ 1,\ S_t(j)=0,\sum\limits_{i\in N(j)}b_{ij}\geq\theta_j,S_t(i)=1\\ -1,\ S_t(j)=1,\sum\limits_{k\in O(j)}S_t(k)=d_{outj}\\ 1,\ S_t(j)=1,\sum\limits_{k\in O(j)}S_t(k)<d_{outj}\end{cases} \tag{3-5}$$

从状态转换规则(3-5)中，当节点 j 的状态为 0 时，$S_t(i)$=0,节点 j 的入邻居中处于激活状态的节点对其的影响力之和大于等于节点 j 的个性阈值时，节点 j 的状态在 t+1 时刻时转变为激活状态，即 $S_{t+1}(j)$=1。反之，则节点 j 的状态在 t+1 时刻保持不变。从式中还可以看出，节点 j 在 t 时刻的状态变化取决于其邻居对其的影响力和本身的个性阈值，其中，当 $S_t(j)$=1 时，即当节点 j 在 t 时刻的状态为 1 时，当节点 j 的出度为 k 时，当出邻居节点的状态和等于节点 j 的出度时，节点 j 不再影响其邻居节点，因此节点 j 此时处于移除状态。反之，则节点 j 在 t 时刻的状态不变。

在阈值模型中，各个节点的阈值都设置成同一个固定数值，即累积影响力大于这个数值时，未激活节点转化为激活节点，没有考虑到节点之间的差异，是不符合实际的，在本文中，θ_j 表示节点 j 的个性阈值，定义 θ_j 的值由邻居节点的影响力及节点 j 的入度来决定,即计算节点的邻居节点对其的平均影响力，若激活节点对该节点的影响力的值大于平均影响力的值，则该节点被激活，否则该节点保持原状态不变。

由上分析，θ_j 用如下公式表示：

$$\theta_j = \frac{\sum_{i \in N(j)} b_{ij}}{d_{inj}} \tag{3-6}$$

令复杂网络用图论来描述，即抽象为 $G(V,E)$, V 表示节点，E 表示节点之间的边，即节点之间的关系。

模型运行步骤：

输入：网络图 $G(V,E)$

输出：T

步骤 1：初始化网络，即 $t=0$ 时，选择任意 k 个节点作为初始的激活节点，设置这 k 个节点的状态为 1，其他网络节点的状态设置为 0，把这 k 个节点加入到集合 T 中。

步骤 2：分别计算节点 i 与其邻居节点 j 的影响力 b_{ij}，并计算 $\sum b_{ii}$。

步骤 3：根据式（3-5）和式（3-6）计算每个节点的状态，且把状态为 1 和-1 的节点加入到集合 T 中，状态转化为-1 的节点从集合 T 中移除。

步骤 4：重复 2、3 步。

步骤 5：若 $T_{t+1}=T_t$ 时，算法结束。

模型实现的伪代码如下：

算法 3.1 CABT 模型实现算法

Algorithm3.1 CABT 模型实现算法
Input:$G(V,E)$
1:Initialize：$t=0$; T
2:random a//任意选择 a 个种子节点
3: $\|T\|=a$
4: for $t=0$ to m
5: for each $i,j \in V$
6: $b_{ij}=(d(i)/(d(i)+d(j)))(1-C(i))$//计算节点 i 和 j 之间的影响力
7: $\theta_j=\sum b_{ij}/d_{inj}$ //计算节点 j 的的个性阈值
8: if $S_t(i)=1\&\&S_t(j)=0\&\&\sum b_{ij}<\theta_j$
9: $S_{t+1}(j)=0$
10: end

11: if $S_t(i)=1\&\&S_t(j)=0\&\&\sum b_{ij}\geqslant\theta_j$
12 $S_{t+1}(j)=1$
13: end if
14: if $S_t(j)=1\&\&\sum S_t(k)=d_{outj}$
15: $S_{t+1}(j)=-1$
16: end if
17: else
18: $S_{t+1}(j)=1$
19: end else
20: end for
21: end for
Output:T

3.4 实验

本文实验的硬件环境为：处理器：Intel(R) Core(TM) 2 Quad Q8400,主频为 2.66GHz，安装内存为 2.00GB，32 位操作系统。操作系统为 windows7 旗舰版 SP1，模拟软件为 matlab7.0.

3.4.1 实验数据集

本文选择如下四个数据集见表 3.1（Zachary's karate club，American College football，Polblogs，Email），表中显示了数据集名称，数据集规模，以及数据集的网络结构特征，用如下四个数据集来验证算法的有效性。

表 3.1 实验数据集

序号	数据集名称	节点数	边数	平均度	平均聚集系数
1	Zachary's karate club	34	78	4.5882	0.5706
2	American College football	115	616	10.6609	0.4032
3	Email	1133	10903	9.6231	0.2219
4	Polblogs	1490	19090	22.4383	0.2636

Zachary's karate club 数据集是社会学家 Zachary 是对美国一所大学的空手道俱乐部的两年时间观察结果,并且对 34 名成员以及这些成员间的关系进行分析。

American College football 数据集是由 Girvan 和 M. Newman 整理的 2000 年秋季的美国大学足球网络，该网络包含 115 个节点，每个节点代表足球队，边代表两个足球队参与比赛的关系。

Email 数据集,描述的是曾通信的联系人记录,节点代表通信主体,节点之间的关系代表节点之间的通信关系。

Polblogs 数据集, 该数据集是由 Lada Adamic 和 Natalie Glance 在 2005 年二月整理的政治博客圈，博客之间的连接是自动抓取的，在该数据集中，节点代表单一博客，边代表博客之间的连接。

3.4.2 实验结果

1. b_{ij} 对比实验

本文采用线性阈值模型对节点之间的影响力不同的计算方法进行对比实验，以传播效果为衡量不同计算方法优劣性的标准。其中 L1 表示网络中节点之间的影响力按照式（3-1）来计算节点之间的影响力，L2 表示本文提出的节点之间的影响力的计算方法, 即按照式(3-2)来确定节点之间的影响力。

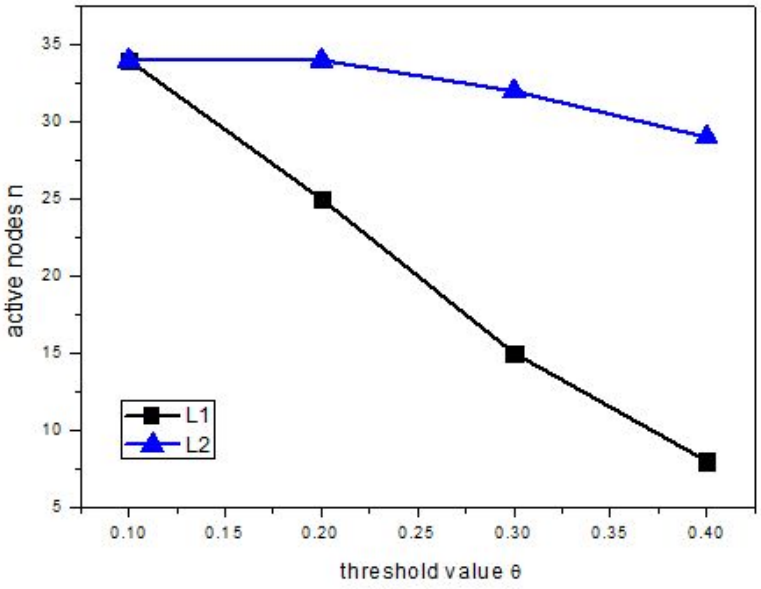


图 3.3 Zachary 网络不同阈值的影响范围

在 Zachary's karate club 中，任取三个节点作为初始的激活节点，由图 3.3 可以看出，在 LT 模型中的阈值θ分别取 0.1,0.2，0.3，0.4 时的激活节点的数目，可以看出 L1 在θ=1 时，节点的激活数目为 34，即网络中所有节点的数目都被激活。在θ=0.2 时，L1 获得的节点感染数目只有 25，而本文提出节点之间影响力的计算方法即 L2 能够激活的数目仍为 34。在θ取 0.2 和 0.1 时，L1 获得激活节点的数目急速下降，到θ=0.1 时，能够激活的数目只有 8 个节点。随着节点阈值的提高，则节点被激活的难度越来越大，因此本文在θ由高到低取值时，L2 激活的节点数目也有所下降，但激活节点的数目仍然接近

最大值。

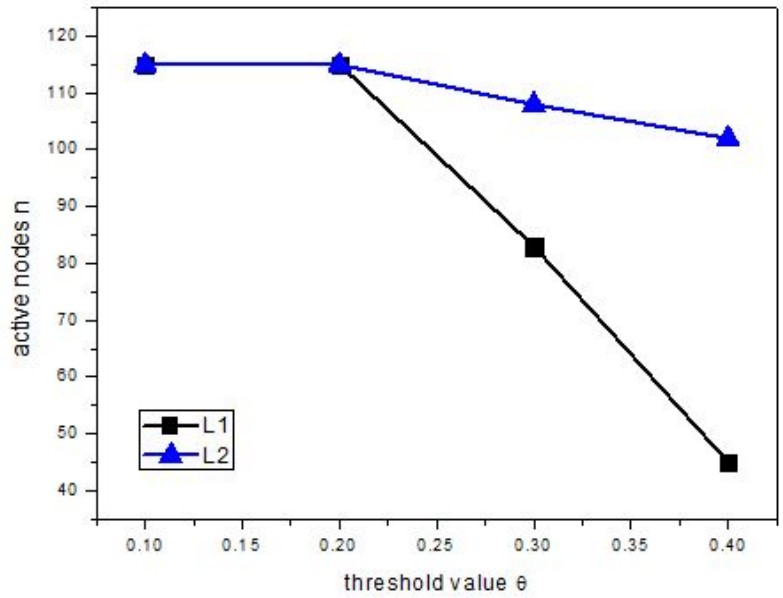


图 3.4 football 网络不同阈值的影响范围

在 football 网络中，如图 3.4 所示，选取 8 个初始激活节点，从图中可以看出，L1 具有同 Zachary's karate club 同样的趋势，阈值θ等于 0.1 时，具有较快的传播速度且能够实现网络中激活节点的最大值。θ等于 0.2 时，同样可以影响更多的节点，但是随着θ增加，激活节点的数目急剧减少。同时可以注意到，图 3.3 两条折线的差异比图 3.4 更加明显，这是由于在 Zachary's karate club 网络中，不同节点的度分布和聚集程度差异较大，而在 football 网络中则节点的度分布和聚集程度分布比较平均，节点之间的影响力区分度较小。

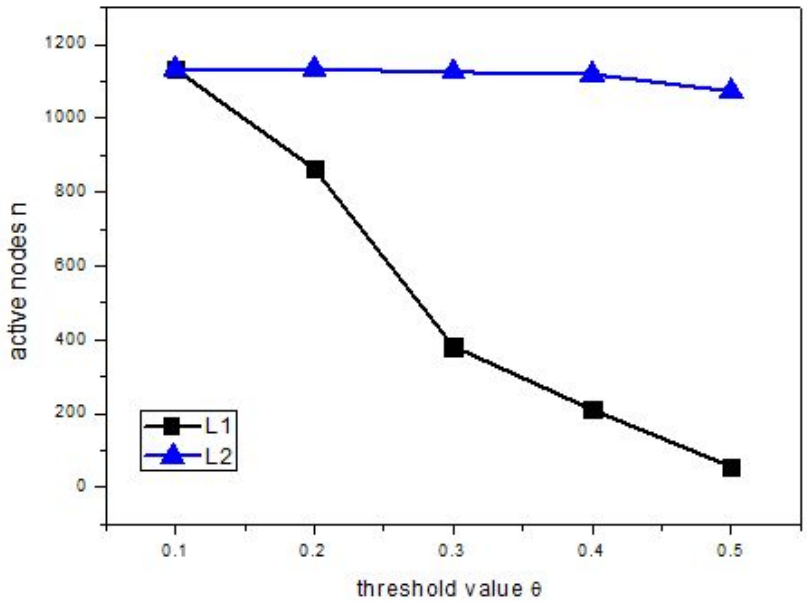


图 3.5 Email 网络不同阈值的影响范围

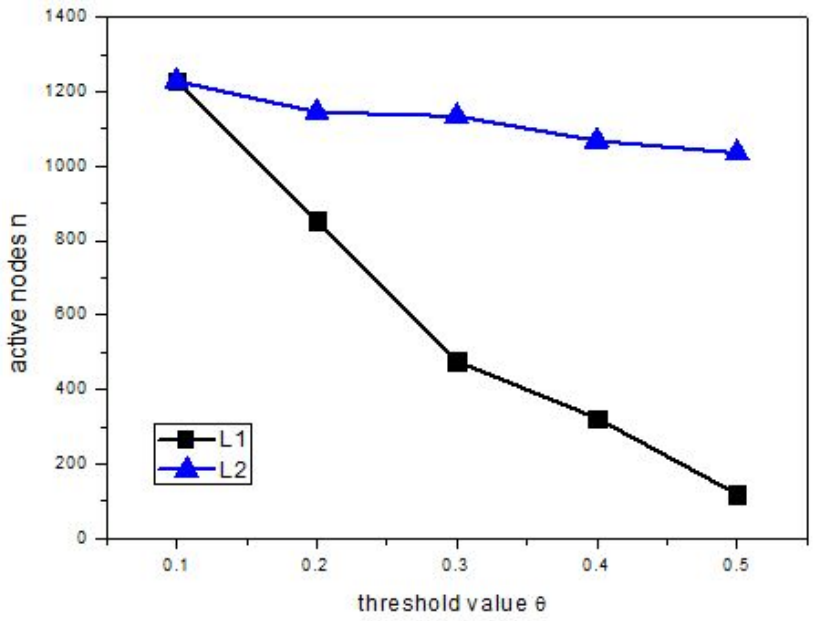


图 3.6 polblogs 网络不同阈值的影响范围

在图 3.5 和图 3.6 中可以看出随着θ值的增大，激活节点的数目具有下降的趋势，在 Email 网络和 Polblogs 中 L1 使得激活节点数目减少幅度比 Zachary 和 football 网络更大，是因为 Email 网络和 Polblogs 网络相比 football 网络和 Zachary's karate club 网络稀疏，且网络结构特征差异明显，因此同样具有随着θ值增加，L1 激活节点数目具有急剧下降的趋势。由上述论述可知，L1 不适合在相对稀疏的网络中计算节点之间的影响力，而现实生活中很多真实的网络都是稀疏网络，因此 L1 计算节点之间的影响力具有很大的局限性。而 L2 在θ取不同值的情况下，获得较多的激活节点的数目，可以克服 L1 的缺点。

在以前的论文的研究中θ取 0.5，应用 L1 计算节点之间的影响力是不合适的，因为在现实生活中，某一商品或服务被越多的人接受则会使范围逐渐扩大，按照 L1 计算节点之间的影响力则会出现营销信息一开始就没有传播出去的现象，因此本文提出的节点之间的影响力计算方法更加有效。

2.传播模型对比实验

本文采用的对比模型是独立级联模型(IC)，n 表示网络中激活节点的数目，其中移除状态和已被激活的节点都接受了营销信息，都看做为激活节点。与独立级联模型进行收敛速度以及传播范围的对比，证明在实现传播最大化时，本模型所需要的时间步最少，更加有利于扩大营销信息传播范围。

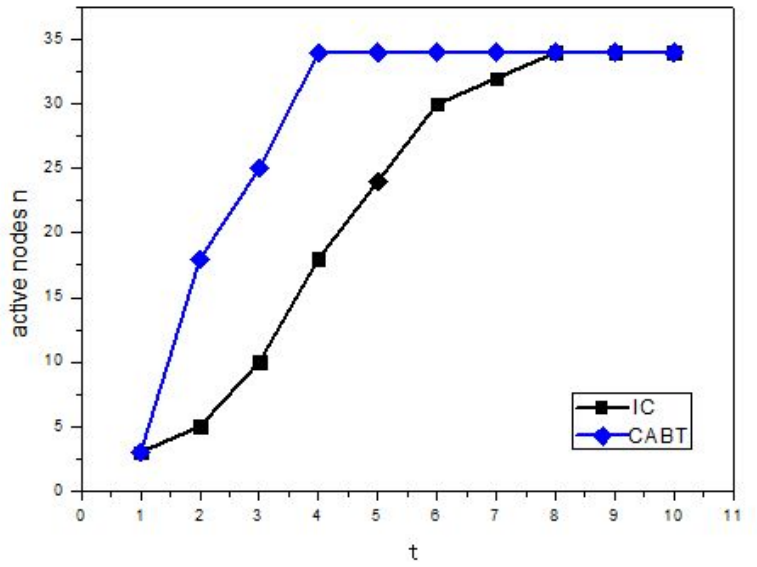


图 3.7 Zachary's karate club 对比实验

在 Zachary's karate club 网络中，该数据集只包含 34 个节点，因此只任意选取一个节点作为初始激活节点，运行结果如图 3.7 所示，CABT 模型在 t=4 时，该数据集中的所有节点都受到营销信息的影响，n 达到最大值。IC 模型在 t=8 时才达到最大值。从图中可以看出，CABT 模型的传播效果明显好于 IC 模型，这是因为在 Zachary's karate club 网络中，节点的中心性差异比较明显，分别以节点 1 和节点 33、34 为中心的网络，而 CABT 模型使中心性越大的节点越具有传播效力，但是 IC 模型对任何节点都是以一定概率进行传播，若节点的中心度很高，但是在激活概率很小的情况下，同样不具备较好的传播能力。

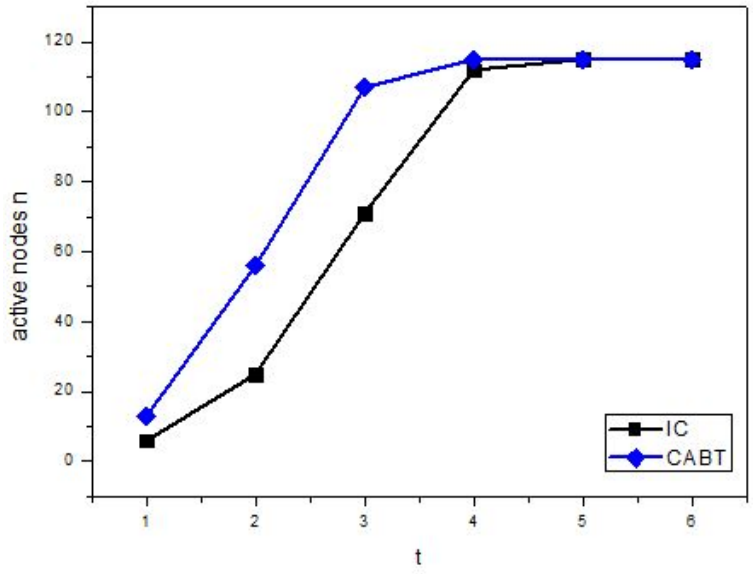


图 3.8 American College football 对比实验

在 American College football 数据集中，运行结果如图 3.8 所示，本文提出的 CABT 模型在 t=4 时，n 达到最大的感染节点数目 115，IC 模型在 t=5 时 n=115。在 football 数据集中，节点的度和聚集系数分布比较平均，且这两个度量的数值相对其他网络比较大，网络中的节点聚集程度相对比较平均，网络中节点的中心性并不明显，在这种情况下削弱了 CABT 模型的传播效果，两种模型的传播差异没有图 3.7 明显，但是 CABT 模型获得的传播效果仍然比 IC 模型好。

在 Email 数据集中，如图 3.9 所示，由于该数据集包含节点数目较多，因此任意选择 4 个节点作为初始激活节点，从图中可以看出 IC 模型在 t=11 时，获得最大感染节点的数目为 1133，CABT 模型在 t=5 时，获得最大感染节点的数目为 1133，CABT 模型比 IC 模型提前 6 个时间步达到 n 最大值。在图中可以看出，CABT 模型在初始传播阶段与 IC 模型能够激活节点的数目相似，并没有取得较好的效果，是由于 Email 网络比较稀疏，与 football 网络相反，度和聚集系数的平均值都很低，因此，CABT 模型与 IC 模型传播效果差异不是很明显，随着激活节点的数目增多，网络中邻居节点的累计影响力逐渐增大，因此节点被激活的数目逐渐增多，CABT 模型的传播效果逐渐优于 IC 模型。

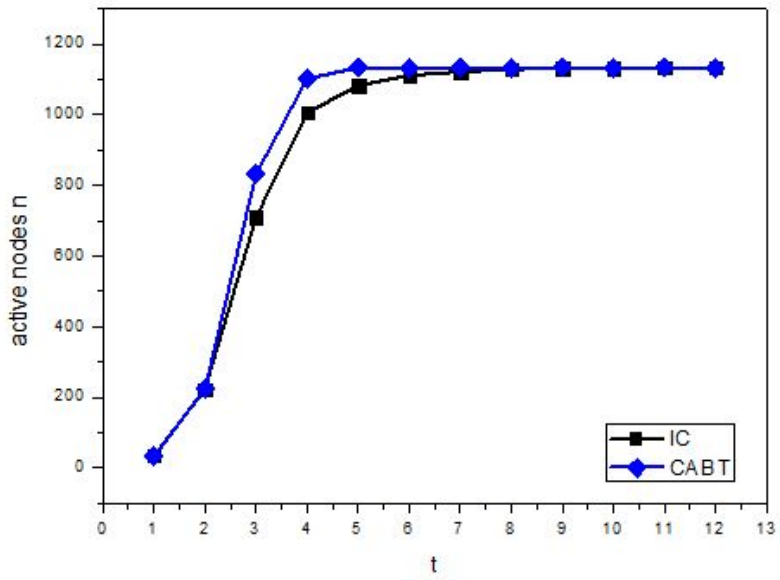


图 3.9 Email 网络对比实验

如图 3.10 所示，在 Polblogs 数据集的实验中，任意选择 5 个激活节点，IC 模型在 t=9 时，激活节点的数目 n 达到 1222 个节点，CABT 模型在 t=6 时 n 达到 1222 个节点，CABT 模型比 IC 模型提前三个时间步达到最大影响范围。IC 模型在 t=6,7,8 时，激活节点的数目分别是 1209,1220，1221，从实验中可以看出在前 4 步本文实验模型比 IC 模型传播范围要小很多，是因为 Polblogs 的度和聚集系数分布极其不均衡，存在很多边缘节

点和孤立节点，而 CABT 模型中，若随机选择度值较小的节点作为初始激活节点，则其对其邻居节点的影响很有限，而 IC 模型是基于概率的传播模型，激活概率不受网络拓扑的影响，因此，在传播过程的前半段 CABT 模型的传播效果不比 IC 好。随着激活节点的增加传播范围迅速加快，越来越多的中心度较大的节点被激活，使得传播范围得到大幅度提升，符合了实际网络中用户的从众心理。

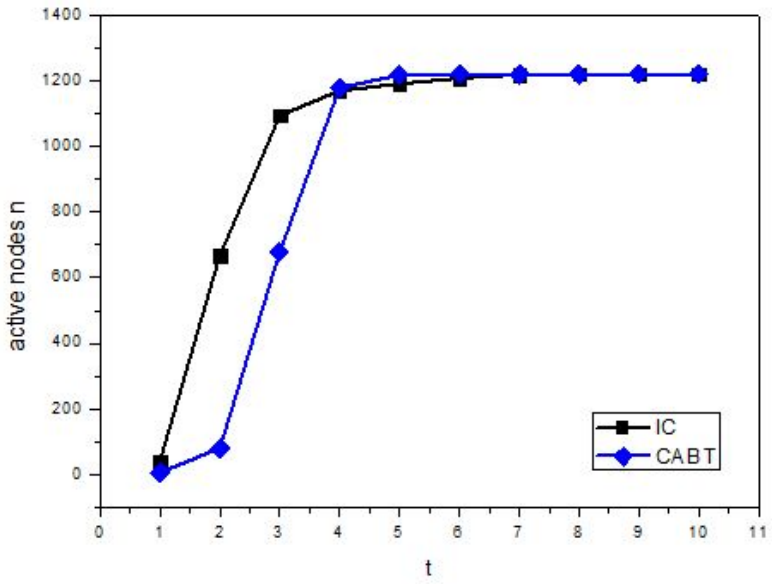


图 3.10 Polblogs 网络对比实验

3.5 本章小结

病毒营销是社会网络中研究的热点，其实质是对社交网络传播动力学的研究。实现传播最大化是病毒营销的最终目的。如何建立传播模型是研究社会网络病毒营销的一个挑战。在本文中，通过节点之间影响力计算及对个性阈值的分析，建立了元胞自动机模型，模拟病毒营销的传播过程，并通过实验证明其有效性，通过对比实验，结果表明，该模型可以有效的加速传播过程。

第4章 种子节点选择算法

4.1 引言

种子节点为初始传播节点，若在网络上进行网站推广，商家要付出一定的营销成本，使网络中的有限的几个节点作为激活节点的集合，使其成为网站推广的第一步。其中怎样选择具有较大传播能力的节点可以使网站得以最大范围的扩展。因此，不管是进行商品营销还是网站推广或创新传播，都热衷于对网络中种子节点的选择，其中种子节点的选择算法已经成为病毒营销研究中的热点问题。种子节点选择算法决定如何在网络中确定利于信息传播的节点。

本章主要是针对目前基于网络拓扑结构的种子节点选择算法存在的缺陷，提出了基于社区结构的种子节点选择算法，首先对社区结构划分算法进行简要介绍和评价；其次对现有的种子节点选择算法进行分析，分析现有算法的优劣性；本文的第三节主要提出结合两种网络结构特征的种子节点选择算法，通过实验论证其有效性。

4.2 种子节点选择问题定义及评价标准

病毒营销过程中种子节点选择算法是确定在网络中影响力最大的前 *top k* 个节点，这 *k* 节点能使病毒营销传播实现最大化即影响力最大化问题，且使这个含有 *k* 个节点的集合成为进行病毒营销传播的起点。

4.2.1 问题定义

种子节点选择问题可以归结为影响力最大化问题，在病毒营销过程中可以描述为：一家公司要对一个网站进行推广，这家公司首先要使网络中的最具有影响力的 *k* 个节点接纳该网站并且进行推广，因为网络中影响力越大的用户越有利于对其邻居用户进行信息传播，网络中的用户之间存在着某种结构和联系，使得网络用户可以快速地从其他用户获取信息，随着商品或网站推广的深入，网络中越来越多的用户接纳该营销信息。

用图论来定义影响力最大化问题，首先用图 $G(V,E)$ 来描述复杂网络，其中把复杂网络中的用户看作节点，用户之间的边看作用户之间的联系，因此 V 表示节点的集合，E 表示边的集合。令 S 表示种子节点的集合，$S=\{v_1, v_2, \cdots, v_k\}$，$k$ 为 S 集合中元素的个数，将这 k 节点作为激活节点去影响网络中的其他节点接受营销信息，在影响力传播过程中，需要指定传播模型，例如 IC 模型或者 LT 模型。目的是采取这 k 节点作为种子节点，可以使网络中受到影响的节点个数达到最大值，即接受营销信息的用户数量达到

最大值。根据上述描述，给出形式化定义，如下式（4-1）所示：

$$\max\{\delta(S) \| S \|= k, S \subseteq V\} \quad (4\text{-}1)$$

其中 $\delta(S)$ 表示 S 集合所激活的节点的数目。

4.2.2 算法评价标准

在以上的论述中可以看出，种子节点选择算法即影响力最大化算法主要关注两个方面：

（a）影响范围，是影响力最大化算法关注的主要方面，主要是指在选择 k 个最具有影响力的节点作为种子节点能够使网络中受到影响的节点数目达到最大值。影响范围是判断算法优劣的重要标准。

（b）算法效率，算法运行的耗时程度是评价算法的又一指标，算法需要在尽可能短的时间内确定最具影响力的 k 个节点。

由于影响力最大化算法的最终目标是实现网络中受到影响的节点最终数目达到最大值，因此较好的算法应该使受到影响的节点数目尽量多，同时，尽可能保证算法的时间效率。若受到影响的节点数目较多，但算法时间复杂度高导致算法运行起来十分困难，或者影响范围较小，则该算法不是好的算法。因此，既能保证影响范围又能兼顾传播效率则是较为提倡的最优算法。

目前对种子节点的选择算法的研究主要是针对贪心算法的的改进算法，以及基于社区结构的影响力最大化算法，还有对网络结构中的节点重要性的评价标准，如度、介数等。

4.3 基于社区结构的种子节点选择算法

在本小节中，提出一种基于网络结构的种子节点选择算法，在社区划分及最大度算法的基础上进行种子节点的选择，本文采用 Fast 算法进行社区划分，并采用在第三章提出的 CABT 模型在真实网络上进行模拟病毒营销传播过程，统计影响范围。

4.3.1 BCI 算法的提出

（1）社区划分与病毒营销的关系

如今复杂网络已经被证明存在社区结构的特征，目前对社区结构研究逐渐深入，多种社区发现算法被提出，且获得了不断改进。社区结构的存在对复杂网络和现实社会具有重要的意义。网络的社区结构分为社区内部关系和社区之间的关系，在社区内部边的

分布是相对比较紧密的，而社区之间的边分布就相对比较稀少，即呈现了网络拓扑结构的不均衡现象。社区在现实生活中代表具有共同爱好或职业的的人群，在社交网络中社区内部也是因为节点之间存在某种共同点才逐渐发展成为联系紧密的节点组。由此可知，社团对于病毒营销有着重要的意义，针对特定社团的特点进行营销策略的制定比盲目营销具有更高的成功率。例如一种新产品被社区中的一个成员获得很好的口碑，则在该成员会把该产品主动介绍给社区中联系比较紧密的成员，这样新产品在该社区中逐渐传播开来，这也是病毒营销在社区结构基础上利于传播的主要原因。

对社区发现算法的研究已经成为研究复杂网络特征的一个重要的研究课题，对充分认识复杂网络具有重要意义。社区结构是由 Newman 最先阐明，目前，社区发现算法主要有 GN 算法，谱平分法，边聚集系数的算法，Kernighan-Lin 算法[64-66]，以及 CNM 算法以 Fast 算法[67]等。

（2）存在的问题

在第二章中的种子节点的相关算法的介绍中，基于单一网络结构的种子节点选择算法具有其一定的局限性，如在最大度算法中，即按照网络中度的大小进行种子节点的选择，但是该算法并没有考虑到影响力的传播路径，最大度算法选择出最大度的节点可能是互为邻居的节点，对离其较远的节点则不利于信息传播。这一缺点在对网络社区划分后的拓扑图中表现的十分明显，例如在图 4.1 中，可以看出，该图是由虚线隔开的具有两个社区的网络拓扑结构，若选取 2 个种子节点，即 k=2 时，则种子节点分别是节点 A

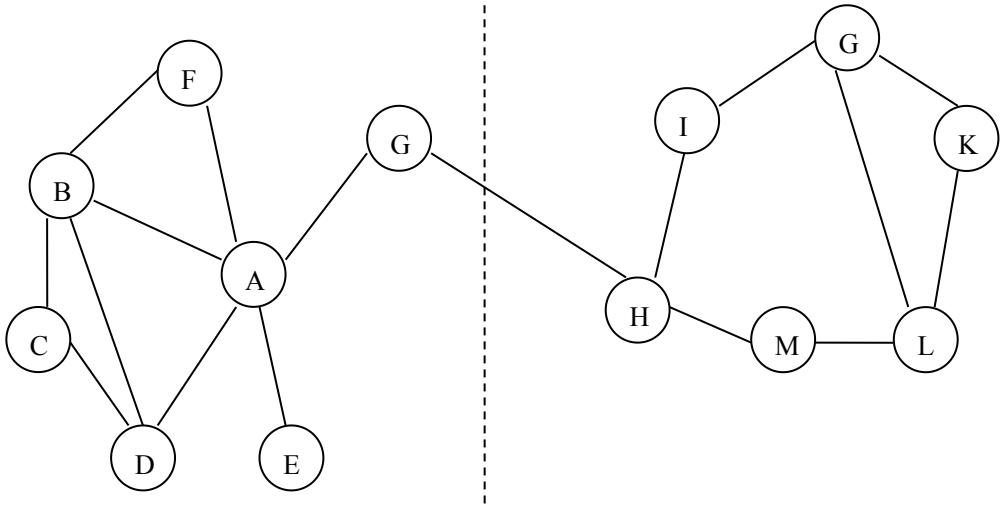


图 4.1 网络拓扑

和节点 B，A 的度值为 5，B 的度值为 6，可见节点 A 和节点 B 互为邻居，且两个种子节点都在同一社区，大大减弱了信息传播的速度，阻碍了信息的传播。因此要解决这一问题，需要把种子节点分散到两个社区中，由于同一社区中的节点联系相对紧密，会促

进影响力的传播。基于以上论述对最大度算法的分析，针对该算法的缺点，本文提出了基于社区结构的种子节点选择算法。

4.3.2 BCI 算法

研究表明，复杂网络是由许多社区构成的，由于社区是按照职业、兴趣以及爱好等形成的，则存在某些节点与多个社区有连接，因为个人的兴趣爱好是多种多样的，与多个社团都会有交集，例如，在现实生活中，一个学生是学生会的成员，同时对绘画感兴趣，参加了素描协会，因此就该学生就连接了两个社区，该学生可以同时收到两个社区的消息。N_c 则是描述网络中单个节点连接社区的数目。在最大度选择算法的讨论中，出现了邻居重叠的问题，导致度值较大的节点分布相对集中，造距离较远的节点受到信息影响的效果小。然而处于多个社区的共同节点，即 N_c 值较大的节点，更有利于把信息传播到多个社区，因此，N_c 值较大的节点可以看作是信息传播的枢纽。

本文提出的基于社区结构的种子节点选择算法（Based on the Community structure Influence algorithm of node，BCI），在该算法中，主要引入了 dN_c 这一度量，该度量考虑了网络节点的及社区划分，衡量了度和连接社区个数 N_c 的网络节点，因此比其他单一的衡量网络节点的重要性度量具有很大的优势。

1.dN_c **的定义**

$N_c(i)$定义为节点 i 连接不同社团的数目。如下图 4.2 所示，网络共有 20 个节点，分为四个社区，即 C_1，C_2，C_3，C_4 四个社区，C_1={A,B,C,D,E,U,V}，C_2={F,G,H,I}，以及

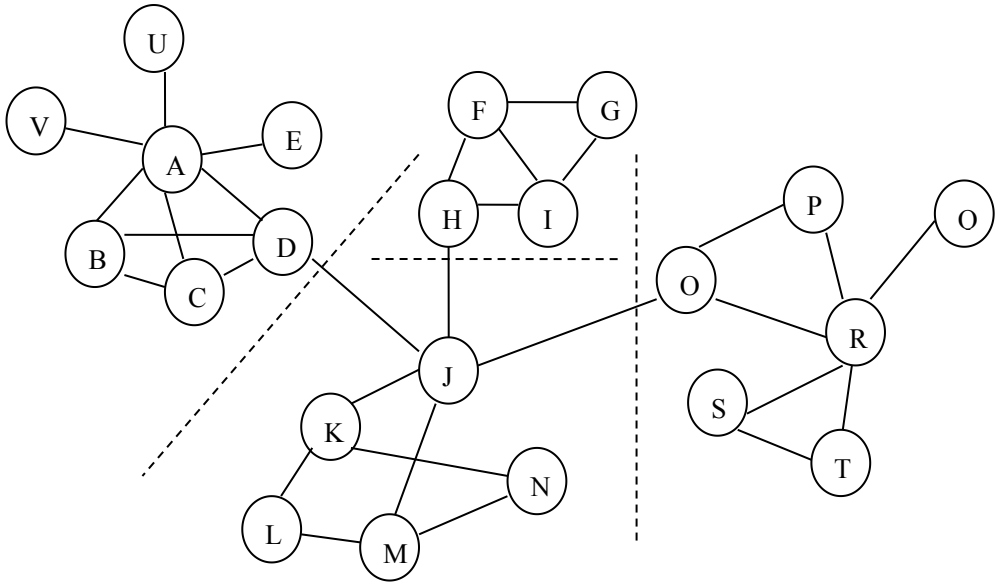


图 4.2 部分节点的网络拓扑

C_3={J,K,L,M,N}，C_4={O,P,Q,R,S,T}。其中 $V_c(J)$=4，图中节点 J 连接节点 K 和 M，由于

K 和 M 处于同一个社区，因此 $V_c(J)$=4。按不同的网络结构特征排序就会得出不同的结果，如按度的最大值进行排序，节点 A 的度值是最大的，即 $d(A)$=6。

描述网络结构的度量总是在某一方面对节点的影响力排序，而排序结果会存在一定的缺陷。BCI 算法是在最大度算法和 N_c 排序的基础上提出的混合算法，对按照单一网络结构特征排序靠前的节点但对传播过程贡献较少的节点进行了有效地控制，发现更多具有传播能力的节点。

dN_c 是衡量节点在网络中的重要程度的度量，即影响力大小的度量，在复杂网络 $G=(V,E)$ 中，$d(i)$表示节点的度，$N_c(i)$表示节点 i 连接不同社区的数目，节点 i 的 dN_c 计算如下所示：

$$dN_c(i)=\frac{d(i)}{\max d(j)}+\frac{N_c(i)}{C} \tag{4-2}$$

其中，j 表示网络中的任意节点，允许 $i=j$，C 代表网络包含的社区个数。

2.BCI 算法的实现

BCI 算法主要是在网络社区划分的基础上进行的，根据节点的 dN_c 值的大小进行排序，选择 *top-k* 个节点，即选择最大的 dN_c 的前 k 个节点作为种子节点。

算法伪代码见算法 4.1：

算法 4.1　BCI 算法

Algorithm 4.1 BCI(G,k)
Input: (*G,k)*
1: Initialize：S= ø，k
2: compute C//进行社区划分
3: compute each d_v ，$v\in A$
4: max$\{d_v\}$
5: for i=1 to k
6:　　compute $dN_c(v)$，$v\in A\backslash S$
7:　　v=max($dN_c(v)$)
8:　　$S=S\cup\{v\},v\in V\backslash S$
9: end for
Output S

3.BCI 算法的应用举例

为了对 BCI 算法进行更加深入的描述，采用较小的数据集进行举例说明，Zachary 网络只有 34 个节点且为无向图，被分为两个社区，见图 4.2。图中节点圆圈的大小代表节点的度大小，在网络中节点的度最大值是 17，节点 id 为 34，Zachary 网络分为两个

社区，根据式（4-2）计算 dN_c 的值，并按照 BCI 算法找出具有最大影响力的 k 个节点，在此令 k=7，即只对前 7 个节点进行排序，见表 4.1，d 表示节点的度，第二列表示节点连接社区的数目，第三列表示本文提出的 dN_c 度量的大小。

由表 4.1 可以看出，BCI 算法在社区连接个数相同的情况下，节点的影响力是按照节点的度进行排序的, id 为 34 是网络中最具影响力的节点。从图 4.3 中可以看出 id 为 33 的节点具有较大度，其度值为 12，但是其只能影响到所在社团内的节点，出现了具有较大影响力的节点邻居重叠的现象，但是应用 BCI 算法，可以有效的克服最大度选择算法的缺陷。

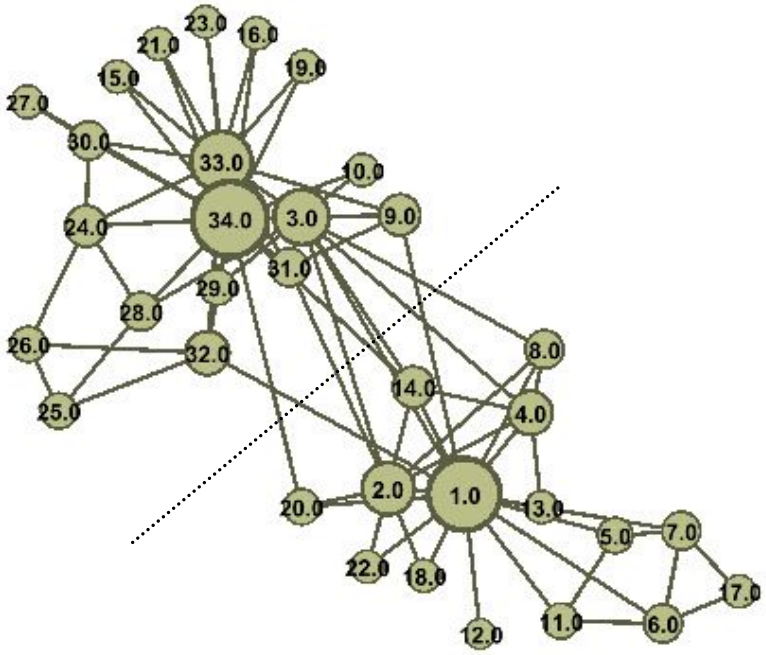


图 4.3 Zachary 网络的社区划分及度分布

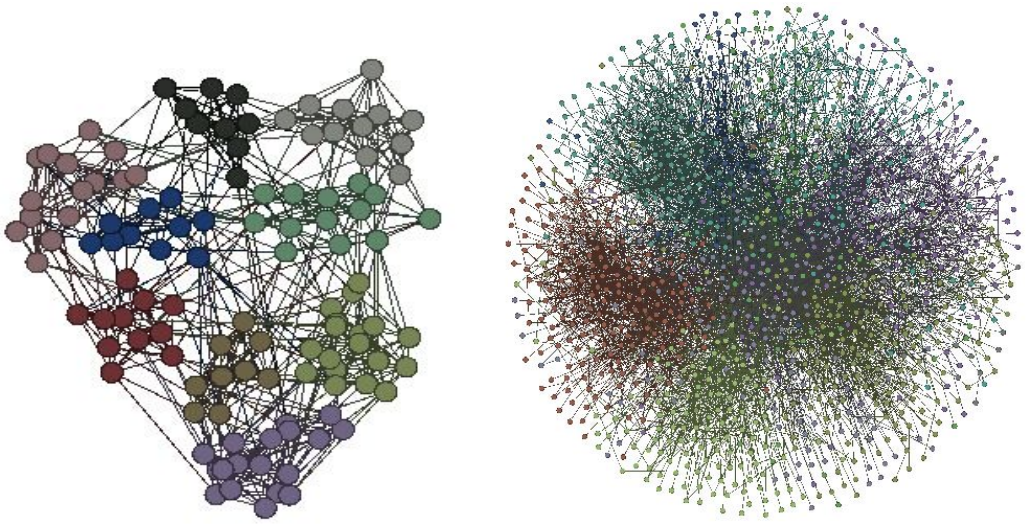
表 4.1 节点的影响力排序

id	d	N_c	dN_c
34.0	17	2	2.000
1.0	16	2	1.941
2.0	9	3	1.529
4.0	6	2	1.352
14.0	5	2	1.294
8.0	4	2	1.235
20.0	3	2	1.176

4.4 实验

4.4.1 实验环境及数据集

本文实验的硬件环境为：处理器：Intel(R) Core(TM) 2 Quad Q8400,主频为 2.66GHz，安装内存为 2.00GB，32 位操作系统。操作系统为 windows7 旗舰版 SP1，模拟软件为 matlab7.0 及 Gephi。



（a）football 网络结构及社区划分　　（b）Email 网络拓扑及社区划分

图 4.4 网络结构及社区划分

本章实验采用 American College football，Email 等两个网络，这两个网络的各项数据统计见第三章，如图 4.4(a-b)所示，分别为以上四个真实网络的拓扑结构。Email 网络是 football 网络的规模的几倍，用节点的颜色区分不同社区，同一颜色的节点处于同一社区，相反则处于不同社区。

4.4.2 种子节点选择算法实验

本节实验分为两部分，第一部分是取不同种子节点集合，即讨论 k 取不同大小的值对传播过程的影响，第二部分实验是讨论单源节点对营销信息传播的影响。

1. k 个种子节点及其影响力实验

在本组实验中采用独立级联模型来评价节点的影响力大小，为了更加突出种子节点对信息传播的影响，IC 模型的激活概率取 0.01，即 p=0.01。对于不同算法得出的节点度量排序，以不同组合作为种子节点，为了尽量减少其他影响，对每组种子节点独立运行 50 次，对运行后得到的激活节点数目取平均值作为评价种子节点的标准。本文提出

的算法与最大度选择算法、介数中心性以及特征向量中心性等基于网络拓扑结构的种子节点选择算法进行比较。

由图 4.5 可以看出，在 k=1 时，按照 BCI 种子节点选择算法，由于 dN_c 在很大程度上考虑到了节点连接社区的个数，且社区之间连接相对比较稀疏，有相对较少的邻居节点，在传播概率低的情况下，单个种子节点的传播能力有限，在 k=5，10，15，20 时，种子节点的个数增加，综合考虑了节点的度和连接社区的数目，因此邻接节点的数目大幅增加，不同社区之间的桥节点逐渐增多，这样比其他种子节点选择方法在社区之间更容易传播信息。最大度选择算法在 k=1 时获得了较好的影响范围，由于其度最大，邻居节点最多，因此在单源节点的传播过程中，其邻居节点中被激活的个数最多，因此在 k=1 时获得了较好的影响范围，在种子节点的数目 k 逐渐增大的过程中，其影响范围的增加幅度相对比较小，如在 k=10 时，激活节点的数目是 493，而当 k=15 时，激活节点的数目则为 507，影响范围的增幅是比较小的，这是因为此时种子节点中，度数大的节点距离相对较近，能够影响的节点范围发生了重叠现象。按照介数中心性选择的种子节点，在 k=1 时，与 BCI 算法选择的种子节点是同一个节点，即为节点编号为 333 的节点，作为单源节点获得了较小的传播范围，但随着 k 的数目的增大，影响范围迅速扩大，是因为节点的增加，其邻居节点逐渐增多，且介数越大能够影响到的节点越多。从图中可以看出，影响范围最差的是按照特征向量中心性算法选择的种子节点。

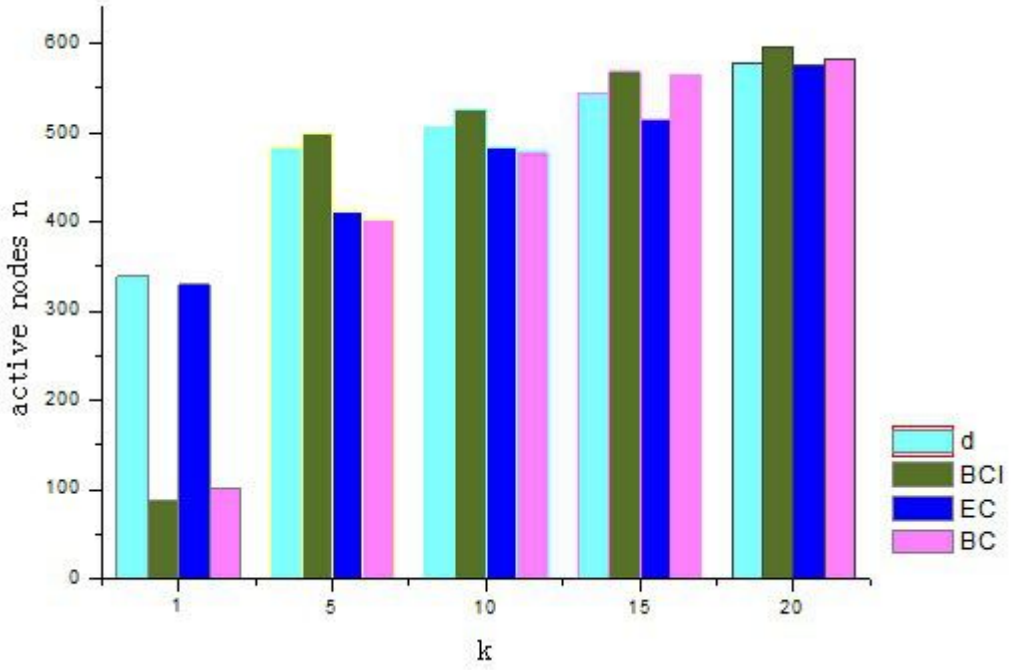


图 4.5 Email 网络 k 个节点及影响范围

由于 football 网络规模较小，因此选择 1 至 5 个种子节点进行讨论其影响范围。当

k=5 时，在计算最具有影响力节点的结果中，特征向量中心性算法与最大度算法得出的

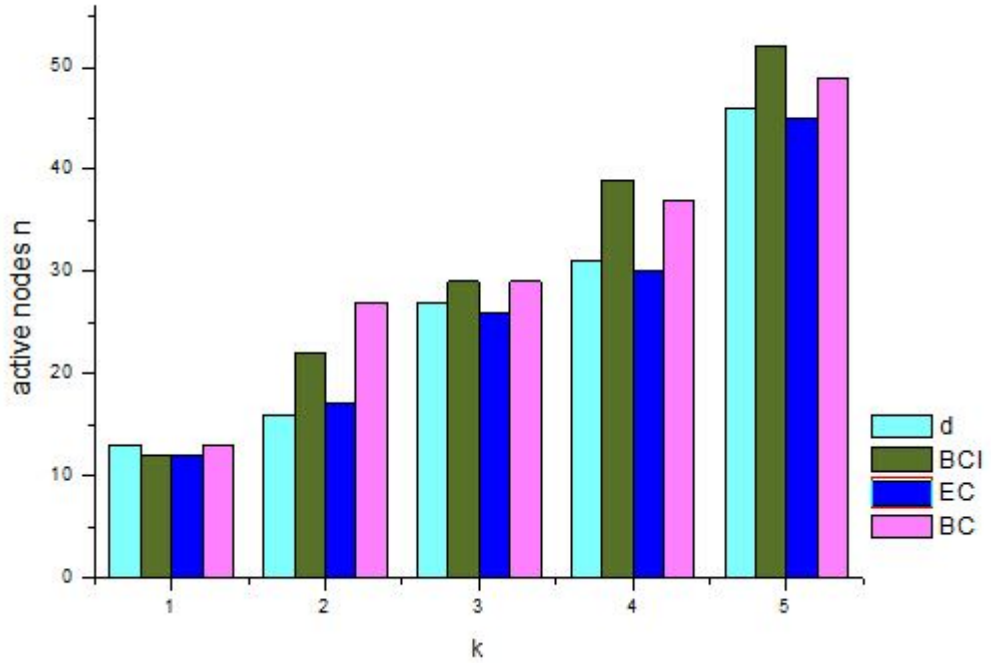


图 4.6 football 网络 k 个节点及影响范围

节点编号相同，因此在图 4.6 中特征向量中心性比最大度算法的影响范围要小。从图中可以看出，在 k=1 时，几种选择算法的影响范围相近，是因为节点的度值分布比较平均，在 football 网络中，节点的最大度数为 12，整个网络的平均度数为 10.66，因此，当种子节点为单源节点时，四种种子节点选择算法的影响范围区分不明显。在 k=2 时，按介数中心性计算得出的种子节点获得了最好的影响范围，此时种子节点的编号为 83 和 1，两个节点的度数和是 23，连接社区的总数是 9 个，而 football 网络总共被分为 9 个社区，可能是由于在度数较大且连接社区数目较多的情况下，营销信息会以较大概率向其他社区传播，因此获得了较好的影响效果。但是从整体上看种子节点的影响范围，最大度选择算法和特征向量中心性算法获得的影响范围比前两种相对较差，BCI 算法在充分考虑度和社区条件下选择种子节点获得了较好的效果。

2.种子节点影响力排序实验

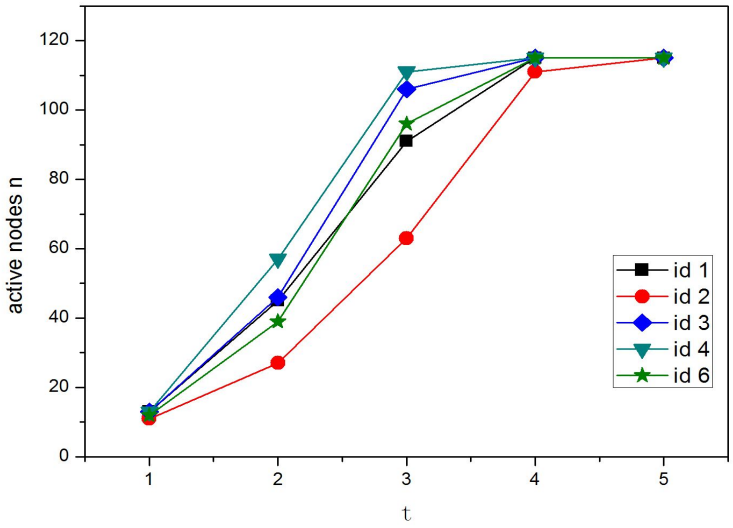
为了对单个种子节点的影响范围进行深入研究，对比不同算法对 *top-k* 个种子节点的影响力进行讨论，本组实验采用 football 数据集，主要针对网络中的单源节点传播范围及传播速度进行讨论，由于 football 网络规模较小，因此只选择 *k*=5 时的种子节点进行讨论。

在 football 网络中，四种种子节点选择算法排序如下表所示，id 表示节点编号。下表中不同算法选择的种子节点存在重复节点，而独立级联模型模拟相同种子节点的影响

范围可能存在差异，因此，在本组实验中采用第三章提出的基于阈值的元胞自动机模型进行营销信息传播过程的模拟。

表 4.2 football 四种度量最大的 5 个节点

Degree	dN_c	EC	BC
id 1	id 59	id 68	id 83
id 2	id 81	id 54	id 1
id 3	id 1	id 89	id 81
id 4	id 2	id 3	id 59
id 6	id 7	id 16	id 39



4.7 基于度选择算法的种子节点传播过程

用图 4.7 表示最大度算法选择的种子节点传播过程。在图 4.7 中可以看出，id 为 4 的节点具有较好的传播范围和速度，但是在最大度算法中节点的影响力却排在第四位，而 football 网络中 id 为 1 的具有最大度，但是其激活范围在 t=1 时，比 id 为 4,3,6 节点差，实际上的影响力排在了第四位，说明节点的度最大未必在传播过程中获得较好的效果，因此在最大度选择算法中不能对节点的影响力进行准确地排序。

图 4.8 描述的是按介数中心性进行排序的种子节点传播过程，从图中可以看出，id=59 的节点在传播过程中处于绝对的优势，但是见表 4.2 中，介数中心性选择算法中节点 59 的影响力排在第 4 位，其他节点在 t=2 和 t=3 时有节点影响范围的更替，id 为 81 的节点在 t=3 时，获得了较好的影响范围，仅次于节点 59。节点的影响范围随时间步有很大的变化是因为在初期节点的度起着很大的作用，但是随着时间步的增长，度的优势逐渐降低或消失，节点其他结构特征逐渐发挥作用。在 t=3 时刻时，而排在第一位的节点 83 实际的影响范围在 id 为 59,81 的节点之后，因此度和以及介数中心性选择算

法并没有挖掘出具有最优影响效果的节点，因此以上算法对节点影响力的排序具有一定的缺陷。

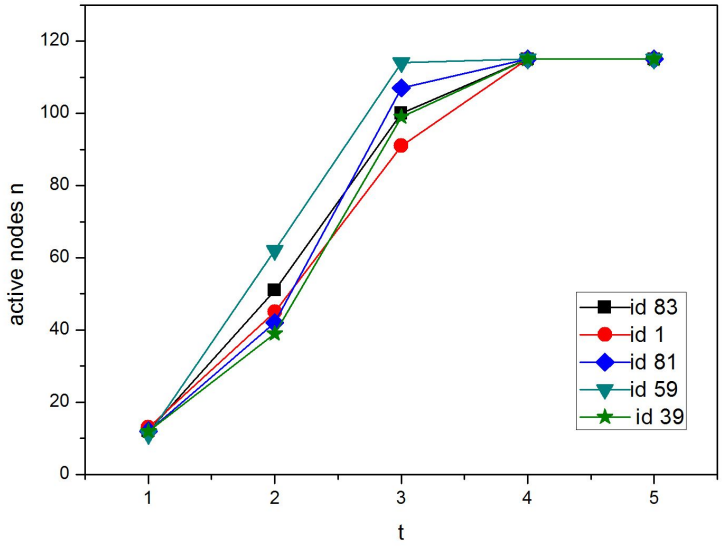


图 4.8 基于介数中心性选择算法的种子节点传播过程

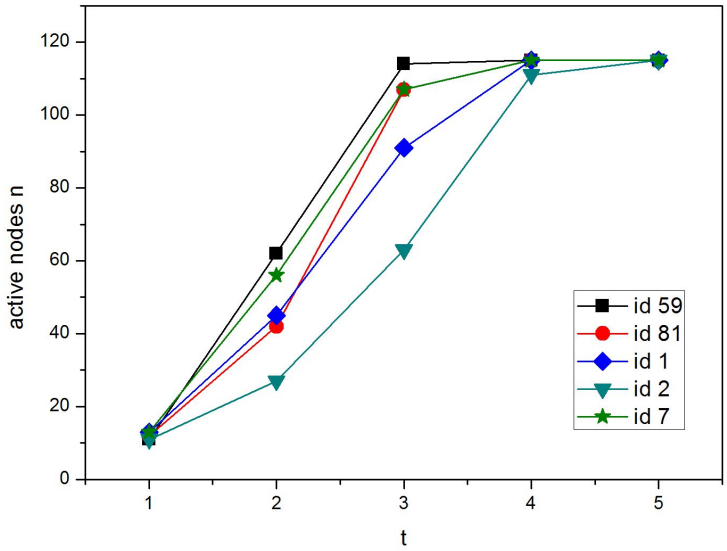


图 4.9 BCI 算法的种子节点传播过程

在图 4.9 中，描述的是 BCI 算法排序的最具影响力节点的传播过程，从图中可以看出，id 为 59 的节点具有较好的传播效果，BCI 算法能够挖掘出最具有影响力的节点。同时还可以发现，按度选择算法得出的最有影响力的节点 1 在图中获得了相对一般的影响效果，同时，在图中也可以看出节点的影响力并不是完全按照 *dNc* 度量进行排序的，

由图中可以看出节点 7 的节点重要性排序在第五位，但是其传播效果比节点 1 和 2 传播效果要好。因此，BCI 选择算法能够更加有效地找出最具影响力的节点。

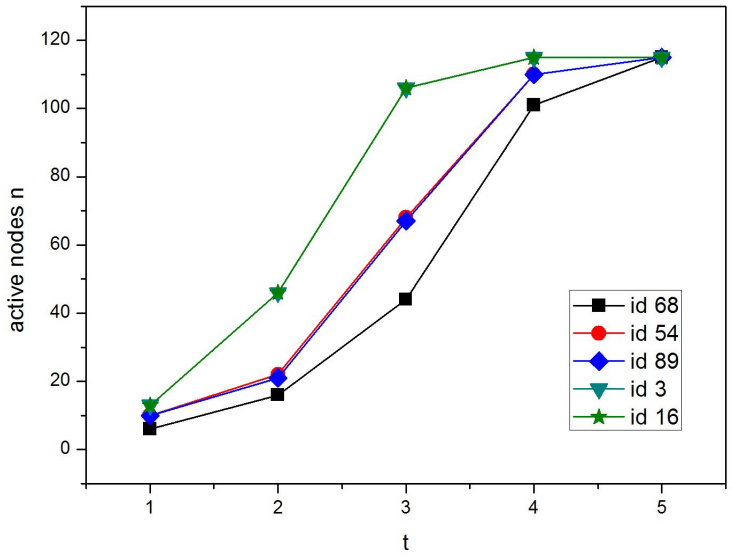


图 4.10 基于特征向量中心性选择算法的种子节点传播过程

图 4.10 描述按照特征向量选择算法获得的种子节点的传播过程，从图中可以看出，id 为 68 的节点应具有最大的影响力，但是其传播速度和影响范围在前五个节点中是最差的，id 为 3 和 16 的节点同时获得了较高的传播速度，但是和节点 59 的传播速度仍然存在一定差距。由上述可知，特征向量中心性不能对节点的影响力进行准确的排序， 且没有挖掘出最具影响力的节点。

BCI 算法是基于网络的拓扑结构的种子节点选择算法，具有较低的复杂度，由算法 4.4 可知，该算法的时间复杂度为 $O(N)$，小于介数中心性的时间复杂度 $O(N^3)$。

4.4.3 实验结论

由以上实验模拟可知，基于社区结构种子节点选择算法能够较准确的选择具有影响力的节点，且该算法是基于网络的拓扑结构时间复杂度相对较低。在与最大度算法，特征向量中心性以及介数中心性对比时的算法具有较好的影响范围。最大度选择算法也具有较低的复杂度，但是其得出的最具有影响力的节点在传播模拟过程中表现出较差的影响效果；基于特征向量中心性计算得出的种子节点也具有较差的影响效果；介数中心性比最大度算法和特征向量中心性算法的影响效果好一些，但是仍然没有 BCI 算法的影响范围好。综上所述，BCI 算法综合考虑了网络节点的邻接状况以及社区多样性，能够挖掘出具有最大影响力的节点，带来较大的影响范围。

4.5 本章小结

在本章中，主要研究在病毒营销过程中种子节点选择算法，即对网络中最具影响力的节点进行挖掘。本章的主要贡献是提出了 BCI 算法，运用关键度量 dN_c，由于该算法是建立在复杂网络社区划分的基础上的，因此，本章首先介绍了几种主要的社区划分算法以及社区划分的意义，而且还介绍了几种主要的种子节点选择算法，并对针对算法存在的缺陷进行了分析，提出了基于 dN_c 的 BCI 算法，并通过 MATLAB 和 Gephi 对 BCI 算法及复杂网络进行了分析，通过实验证明 BCI 算法的有效性。

结论

病毒营销作为一种新型的营销方式已经备受关注，对它的研究涉及到多个方面，许多产品和服务在主流的社交网络上都采用病毒营销进行推广，病毒营销已经成为最具有潜力和发展前景的营销模式。病毒营销传播机制研究主要是针对传播模型和种子节点的选择算法进行研究，传播模型用来刻画营销信息在网络中的传播，深入理解传播模型可以为商家更加熟练的应用病毒营销模式，种子节点的选择算法则是用来挖掘网络中具有影响力的节点，对于实施病毒营销的商家来说，致力于找出最大影响力的节点会带来营销信息的快速传播。随着复杂网络的深入研究，传播过程中众多复杂性和不确定性逐渐表现出来，因此在现实和理论上都需要对病毒营销传播机制做进一步深入的研究。

本文主要是对复杂网络的病毒营销传播机制进行研究，通过结合网络结构研究节点之间的影响力关系，并对传播模型和种子节点的选择算法进行研究。本文提出了一种有效的节点之间的影响力的计算方法，建立了基于阈值的元胞自动机模型来模拟营销信息的传播过程，针对如何选择种子节点这一问题，本文提出了基于度和社区结构的种子节点算法。

（1）节点之间的影响力估计

本文提出的节点之间的影响力计算方法主要是考虑到节点周围的网络结构，以及被影响节点的邻接节点的数目。现有的研究认为节点的聚集系数越大，即其邻居之间联系越密切，则节点的中心性就降低，影响力相应减少。现有的传播模型中节点之间的影响力值主要是通过设置随机数和只考虑度值的度量，这些度量在初始初始传播过程中需要设置很低的阈值才能使得信息得以传播，这是不符合实际的，因此具有很大的局限性。

（2）基于阈值计算的元胞自动机模型

现有的传播模型如独立级联模型以及线性阈值模型等在传播过程中并未对网络节点之间的影响力做有效的度量，在节点之间影响力的基础上，提出了基于阈值计算的元胞自动机模型，该模型的阈值考虑了被激活节点本身的度值及邻居节点激活节点的数目以及平均影响力，并与独立级联模型进行对比，基于阈值的元胞自动机模型可以使营销信息进行快速传播，具有较好的效果。

（3）基于社区结构的种子节点选择算法

本文根据复杂网络结构定义了 dN_c 度量，在此基础上提出了 BCI 算法，该算法结合了网络的社区划分以及节点的度，与其他基于网络拓扑结构的种子节点选择算法进行了对比，实验证明，对于不同规模的种子节点集合，BCI 算法获得了较好的影响效果；在对单源种子节点的传播速度方面，BCI 算法同样获得了较好的效果。

由于自身水平有限，本文对病毒营销的传播机制研究仍存在许多缺陷，总结如下：

（1）本文提出的基于阈值的元胞自动机模型在模型信息传播过程中并没有考虑到个体传播的意愿，没有考虑到用户的主观因素，因此对传播效果有一定的影响。

（2）基于阈值的元胞自动机模型虽然没有传播概率的影响，不会产生同一节点传播范围明显不同的后果，但是在非常稀疏的网络中传播效果表现不明显。

（3）本文提出的基于社区结构的种子节点选择算法对社区划分算法的依赖性很大，社区划分结果直接影响本文提出的种子节点选择算法，对于不同的社区划分算法，则会产生不同的种子节点集合，因此该算法具有一定的局限性，将进一步研究社区划分方法。

参考文献

[1] Amaral L A N, Scala A, Barthelemy M, et al. Classes of small-world networks[J]. Proceedings of the National Academy of Sciences, 2000, 97(21): 11149-11152.

[2] Latora V, Marchiori M. Is the Boston subway a small-world network?[J]. Physica A: Statistical Mechanics and its Applications, 2002, 314(1): 109-113.

[3] Williams R J, Martinez N D. Simple rules yield complex food webs[J]. Nature, 2000, 404(6774): 180-183.

[4] Montoya J M, Solé R V. Small world patterns in food webs[J]. Journal of theoretical biology, 2002, 214(3): 405-412.

[5] 邵培仁.传播学[M].北京：高等教育出版社，2007:24.

[6] Steve Jurvetson,Tim Draper.Viral Marketing [J].The Netscape M-Files,1997.

[7] 蔡占华. 社会网络社区对病毒营销意义[J]. 商场现代化, 2013, (27): 95-96.

[8] 袁宁,吴志军.“病毒”营销在网络中的应用[J].电子商务，2005,(28)：93-94.

[9] 徐孟晓, 王旭. 病毒营销传播渠道研究[J]. 云南财经大学学报, 2015,24(1):98-100.

[10] 孙鹏志. Google 的“病毒营销”策略[J]. 市场周刊，2005,(5):90-91.

[11] 卢智慧. 病毒性营销在网络营销中的应用研究[J].电子技术, 2012,(01):42-43.

[12] 刘向阳. 病毒营销的理论基础及传播机理模型研究[J].中国商贸，2010，(28):34-35.

[13] 张紫琳. 病毒营销实施策略研究[J].经济理论研究，2009，(07):39-40.

[14] 梁喜. 病毒营销传播意愿的影响因素及其传播策略[J]. 江苏商论，2010,(12):73-74.

[15] 陆昊菁. 病毒营销的社会化媒体传播动力挖掘[J].今传媒, 2012,(4):78-81.

[16] Domingos P, Richardson M. Mining the network value of customers[C]//Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2001: 57-66.

[17] 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.

[18] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, Natalie Glance, Cost-effective outbreak detection in networks[C], Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2007: 420-429.

[19] Forrest Stonedahl , William Rand , Uri Wilensky, Evolving viral marketing strategies, Proceedings of the 12th annual conference on Genetic and evolutionary computation, 2010: 07-11.

[20] Christine Kiss, Martin Bichler.Identification of influencers—measuring influence in customer networks[J]. Decision Support Systems, 2008, 46(1): 233-253.

[21] 田家堂, 王轶彤, 冯小军. 一种新型的社会网络影响最大化算法[J].计算机学报, 2011,(10):1956-1965.

[22] 冀进朝, 韩笑, 王喆. 基于完全级联传播模型的社区影响最大化[J]. 吉林大学学报（理学版）,2009,(05):1032-1034.

[23] 夏秀峰，赵秀涛. 客户信任网络下病毒式营销核心群体的挖掘[J].小型微型计算机系统,2011,(01):56-60.

[24] 赵秀涛. Web 病毒式营销中的挖掘技术研究[D]. 2011.

[25] 张伯雷，钱柱中，王钦辉，陆桑璐. 面向目标市场的信息最大覆盖算法[J]. 计算机学报，2014,24(4):894-904.

[26] 苏晓萍，申情，宋玉蓉，蒋云良. 利用社会网络上最有影响力节点实现高效病毒营销[J].小型微型计算机系统，2014，35(8)：1803-1807.

[27] F. Bass. A new product growth model for consumer durables[J]. Management Science, 1969, 15(5): 215-227

[28] Yongkun Li, Bridge Qiao Zhao, and John C. S. Lui. On Modeling Product Advertisement in Large-Scale Online Social Networks[J], IEEE/ACM Transactions on Networks, 2011, 20(5):1412-1425

[29] Tsung-Hao Meng-Fen Chiang, Wen-Chih Peng. Inferring social relationships across social networks for viral marketing[C]//Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on. IEEE, 2012: 143-150.

[30] Jianmei Yang, Canzhong Yao, Weicheng Ma, Guanrong Chen. A Study of the Spreading Scheme for Viral Marketing Based on a Complex Network Model[J]. Physica A. 389(2010):859-870.

[31] 张彦超. 社交网络服务中信息传播模式与舆论演进过程研究[D].北京交通大学，2012.

[32] 郭艳燕，童向荣，刘其成，龙宇，李晔. 基于博弈论的网络信息传播模型的研究[J]. 计算机科学，2014,41(3):238-244.

[33] 吴凯，基于微博的信息传播建模与节点影响力研究[D]. 信息工程大学，2013.

[34] 郭进时，汤红波，吴凯，杨森. 基于社区结构的影响力最大化算法[J].计算机应用，2013,33(9):2436-2439.

[35] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M], 北京:清华出版社, 2006.

[36] Newman M E J. A measure of betweenness centrality based on random walks[J]. Social networks, 2005, 27(1): 39-54.

[37] Goh K I, Oh E, Kahng B, et al. Betweenness centrality correlation in social networks[J]. Physical Review E, 2003, 67(1): 017101.

[38] Barthelemy M. Betweenness centrality in large complex networks[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 163-168.

[39] Newman M E J. The structure and function of com plex netwo rks [J]. SIAM Rev, 2003, 45(2): 167-256.

[40] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics[J]. Physics reports, 2006, 424(4): 175-308.

[41] Flake G W, Law rence S R, Giles C L, et al. Self-organization and identification of web communities [J]. Computer,2002, 35(3): 66-71.

[42] Cheng X Q, Ren F X, Zhou S, et al. Triangular clustering in document networks[J]. New Journal of Physics, 2009, 11(3):033019.

[43] Guimera R, Amaral L A N. Functional cartography of complex metabolic networks[J]. Nature, 2005, 433(7028): 895-900.

[44] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences of the United States of America, 2002, 99(12): 7821 −7826.

[45] Boguná M, Pastor-Satorras R. Epidemic spreading in correlated complex networks[J]. Physical Review E, 2002, 66(4): 047104.

[46] Ravasz E, Barabási A L. Hierarchical organization in complex networks[J]. Physical Review E, 2003, 67(2): 026112.

[47] Song C, Havlin S, Makse H A. Self-similarity of complex networks[J]. Nature, 2005, 433(7024): 392-395.

[48] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing letters, 2001, 12(3):211-223.

[49] Granovetter M. Threshold models of collective behavior[J]. American journal of sociology, 1978: 1420-1443.

[50] 罗英伟. 基于 Agent 的分布式地理信息系统研究:[D]. 北京:北京大学, 1995.

[51] Ma Xiaofei. Study on modeling and simulation of recrystallization annealing process for material with second phase particle by celluar automata method[D].Jinan:Shandong University,2008.

[52] Von Newman. Theory of self-reproducing automata[M].Urban:University of Illiois Press,1966.

[53] Bailey N T J. The mathematical theory of infectious diseases and its applications [M].New York: Hafner Press, 1975.

[54] Jin Y, Wang W D, Xiao S W. An SIRS model with a nonlinear incidence rate [J]. Chaos, Solitons and Fractals, 2007, 34(5): 1482-1497.

[55] Hayashi Y, Minoura M, Matsukubo J. Oscillatory epidemic prevalence in growing scale-free networks [J]. Physical Review E, 2004, 69(1): 016112.

[56] Galeotti A, Goyal S, Jackson M O, et al. Network games[J]. The review of economic studies, 2010, 77(1): 218-244.

[57] Young H P. The diffusion of innovations in social networks[J]. The Economy As an Evolving Complex System III: Current Perspectives and Future Directions, 2006, 267.

[58] Ma H, Yang H, Lyu M R, et al. Mining social networks using heat diffusion processes for marketing candidates selection[C]//Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008: 233-242.

[59] Bass F M. A new product growth for model consumer durables [J]. Management Science,1969, 15(1): 215-227.

[60] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks[C]. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009:199-208.

[61] T.Cao,X.wu,S,wang,X.Hu. OASNET:An Optimal Allocation Approach to influence Maximazation in Modular Social network.SAC'10 March 22-26,2010.

[62] 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.

[63] 于鑫, 段晓东, 刘向东, 周福才. 基于元胞自动机的流行病传播模型及模拟[J].计算机工程与应用, 2005,(02):205-209.

[64] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(2): 298-305.

[65] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Journal on Matrix Analysis and Applications, 1990, 11(3): 430-452.

[66] Kernighan, B. and S. Lin. An Efficient Heuristic Procedure for Partitioning Graphs. The Bell System technical journal, 1970. 49(1): 291-307.

[67] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

攻读硕士学位期间发表的论文和取得的科研成果

致　谢

在论文即将完成之际，我由衷的向在生活上和学习上给予我支持、教导、鼓励、帮助的各位老师和同学们表示最真挚的感谢。

首先我要感谢我敬爱的导师董宇欣副教授，感谢她在研究生期间对我学习和生活上的指导、支持和帮助。在工作非常繁忙的情况下，董老师仍然给予学生无微不至的关怀。论文从开题、实验、撰写论文、修改等每个阶段，董老师都对我进行了悉心的指导，提出了许多宝贵的意见和建议。同时，董老师为学生营造了开阔、自由、充满智慧的学术氛围，使我的思路更加开阔。老师严谨的学术态度和渊博的知识以及对学术的敏锐思维让我受益匪浅。在此，对我的导师表示深深的感谢。

感谢董红斌教授，老师对学术思路清晰，锐意进取，学识渊博，对我的学术思路给予了很大的指导和帮助。尤其每周的研讨会使我的视野更加宽广，充实了自己的知识储备，给学生营造了浓厚的学术氛围，锻炼和提高了我的思维能力。在此由衷的感谢董红斌老师。还要感谢实验室的印桂生老师和冯小宁老师。

同时，感谢实验室的同学们：崔薇、迟阔、由柏娇、杨永亮、尹宛厶、龙丽丽、宋敬环、沙与海、施铮、陈会龙等同学对我日常学习和生活上的帮助。在研究生生活的两年多时间里，共同学习，共同讨论与奋斗，这份深厚的友情使我的研究生生活充满快乐。此外，还要感谢实验室的王莹洁博士，滕旭阳博士，张亚楠博士，杨雪博士对我的帮助。

感谢我的室友李泳昕，在两年半的时间里对我生活上的照顾，在这里表示深深的感谢。

感谢我的父母对我学习和生活上的支持，你们的鼓励是我前进的源动力，感激之情，无以言表。

最后，还要感谢各位论文评阅老师，感谢百忙之中审阅我的论文，并提出宝贵的意见。