北京交通大学

硕士学位论文

社区发现技术的研究与实现

姓名：薄辉

申请学位级别：硕士

专业：计算机软件与理论

指导教师：黄厚宽

20090101

中文摘要

摘要：近年来随着复杂网络的发展，在现实世界中复杂网络无处不在，从因特网到万维网，从航空路线到大型电力网络，从超大规模集成电路图到人际关系网等等。随着近年来对复杂网络性质的物理意义和数学特性的深入研究，人们发现许多实际网络都具有一个共同的性质，即社区结构。所以在复杂网络中自动搜寻或发现社区具有重要的实用价值，发现这些网络中的社区有助于更加有效的理解和开发这些网络。

特别是 Internet 的迅速发展，互联网上的信息量越来越庞大，它已成为全球最大的信息发布库。目前互联网上的信息纷繁复杂，如何对其内容进行分析从而挖掘出人们所需要的内容这一问题亟待解决。社区发现技术可在一定程度上解决这个问题，不仅节省了用户的时间，而且提高了分析的效率。因此，将此技术用于 Web 挖掘具有重要的理论意义和实用价值。本文将从理论、算法和实现等三个方面研究社区发现技术。

但是，由于我们事先不知道到底应该将网络分为多少个社区，这使得这一问题极具挑战性。本文首先阐述了社区发现技术基本理论，对现有的典型社区发现算法作了分析，例如，Belief Propagation 算法、k-means(k-centers)算法、Kernighan-Liu 算法、谱平分法、W-H 算法、GN 算法、派系过滤算法等，并研究了各个算法的核心思想，算法复杂度，以及适用范围等。

我们将社区发现经典的 GN 算法加以实现，并用数据集进行实验。此外，我们还实现了 Frey 提出的 Affinity Propagation(AP)算法，通过编程和实验分析，对 AP 算法的算法思想，及算法复杂度有了深刻的认识和理解，并在原有 AP 算法的基础上，做了一些改进。并且还利用了社会网络中的联系关系设计了一个新的社区发现方法，称之为联系关系算法。并通过实验结果对这三种算法进行分析研究，为以后的进一步研究做准备。

关键词：GN 算法；AP 算法；社区发现；联系关系

分类号：TP182

ABSTRACT

ABSTRACT: In recent years, with the development of the complex networks, there have been numerous and various complex networks with the development of science, technology and human society, such as the Internet, the World Wide Web, the network of air lines, large-scale electric power networks, the structure of a piece of very large scale integration, the human social relationships, etc. Empirical studies and theoretical modeling of networks have been the subject of a large body of recent research in statistical physics and applied mathematics. And people find a property that seems to be common to many networks is community structure. Community detection in large networks is potentially very useful. Detecting communities in networks lead us to more efficiently understand and develop these networks.

Especially, with the rapid development of Internet, it has increasingly large amounts of information, and it has become the world's largest reservoir of information dissemination. Currently numerous information is on the Internet, but the method of analyzing the content of information yet to be resolved. Community detection technology can solve this problem to some extent; by using it, users will not only save time but can also greatly improve efficiency. Consequently, community detection technology which is used in information mining has importantly meaning of theory and value of practicality. This thesis aims to discuss the theory, algorithms and implementation of community detection technology.

However, we often have no idea how many communities we wish to discover in networks, the problem of community detection is quite challenging. First we describe the basic concept of the community detection technology and analyze the existing models, such as the Belief Propagation algorithm, k-means (k-centers) algorithm, Kernighan-Lin algorithm, spectral algorithm, W-H algorithm, GN algorithm, Clique Percolation Method and so on. And as for these algorithms, we study their core idea, the complexity, the scope of application and so on.

We implement the typical algorithm in community detection: GN algorithm and detect by using data set. Further, we also program Affinity Propagation (AP) algorithm which is proposed by Frey. By implementing and analyzing, we have deeply understanding of the idea, the complexity of AP algorithm. We also improve the AP algorithm by doing a little change on the original algorithm. And we design a new

method of community detection by make use of communication relation in social networks, which it is called communication relation algorithm. By comparing the experimental results of these algorithms, we analyze and study these algorithms, and get ready for further study in the future.

KEYWORDS: GN algorithm; AP algorithm; community detection; communication relation

CLASSNO: TP182

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他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。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：薄琴　　导师签名：李长安

签字日期：2009 年 6 月 20 日　　签字日期：2009 年 6 月 20 日

致谢

本论文的工作是在我的导师黄厚宽教授的悉心指导下完成的，黄厚宽教授严谨的治学态度和科学的工作方法给了我极大的帮助和影响。因此，在此衷心感谢近两年来黄厚宽老师对我的关心和指导。在课题研究期间，黄老师在学习和生活上给予了我很多的指导、帮助和关心，他高尚的师德、宽以待人的作风以及对学生认真负责的态度都使我终身难忘。黄厚宽老师对我的科研工作和论文都提出了许多宝贵的意见。在此，我再次向黄厚宽老师表示诚挚的敬意和由衷的感谢。

林友芳副教授悉心指导我完成了实验室的科研工作，在学习上和生活上都给予了我很大的关心和帮助。在近两年的学习期间，无论是科学理论研究还是日常学习安排，都给予了我很多的指导，让我受益匪浅，在此我向林友芳老师表示衷心的感谢和敬意。

在实验室工作及撰写论文期间，万怀宇师兄对我论文中的算法研究工作给予了热情帮助，在此向万师兄表达我的感激之情。

另外，我也要感谢我的家人，他们的理解和支持使我能够在学校专心完成我的学业。

1 绪论

1.1 研究背景

自从 D.J.Watts 和 S.H.Strogatz 提出 Small World 网络模型，并将统计物理学的方法应用于复杂网络研究以来，经过 Albert 和 Barabási，S. N. Dorogovtsev 与 J.F.Mendes，M.E.J.Newman 等人前瞻性工作的推动，复杂网络已经成为科学研究特别是复杂性研究的一个重要领域。

从 Internet 到万维网，从生态环境中的食物链网到生物体中的新陈代谢网络，从科研合作网络到各种政治、经济、社会网络，从大型电力网络到全球交通网络，人们生活在一个充满着各种各样的复杂网络的世界中，这也使得复杂网络的研究成为必要。所谓复杂网络是复杂系统的抽象，网络中的节点是复杂系统中的个体，节点之间的边则是系统中个体之间按照某种规则而自然形成或人为构造的一种关系。由于复杂网络节点众多，结构复杂，使得研究复杂网络非常困难。社区发现以及在复杂网络中寻找社区结构的方法使复杂网络化为若干个节点较少，结构较简单的子网络，从而使研究较为简单。

社区发现技术是当今飞速发展的数据挖掘和探查性分析中的一个极为重要的技术，在社会网络分析、数据挖掘、统计学、机器学习、空间数据库技术、生物学和市场学等领域有着广泛的应用。近年来，寻找复杂网络中的社区结构的方法已经成为复杂网络中研究的热点，将社区发现技术应用到复杂网络中社区结构的研究。并且这方面的研究也蕴藏着巨大商机，也是了解人类行为模式的基础，从而成为数据挖掘、信息检索、社会网分析等领域日益关注的焦点[1-7]。

同时丰富的数据资源也为这方面的研究提供了必不可少的支持，由于在网络上发布和获取信息越来越便利，所以网络上可供社区发现技术研究的各种数据资源非常丰富，如著名的 DBLP 数据资源(dblp.uni-trier.de，可以免费下载)；此外其它数据资源，例如 Blog、E-mail、电信数据等和我们日常生活有关的资源，很多研究已在这些数据资源上取得了研究成果，这不仅为社区发现研究提供了支持和帮助，还说明了社区发现技术研究已深入到人们社会生活的各个方面，具有重大的意义。

除此之外，其研究价值除体现在学术方面外（因为社区发现技术其本身就是一个学术课题），还具有巨大的实用价值，例如，它可以很便利的帮助我们找到某一研究领域并且关系紧密的专家，并提供专家推荐，促进科技的发展；此外，在商业领域，例如在电信业，可以发掘联系紧密的客户，并专门对这一部分用户提

供特别套餐服务，来进一步提高企业效益。另外在一些特殊领域，如犯罪侦查，从犯罪侦查的角度出发，我们可以利用已在公安机关记录在案嫌疑人的交往数据，结合社区发现和犯罪学的知识，经过一系列的人为干预和计算机分析之后，得到一些关于黑社会组织或者犯罪团伙的集团信息，用于辅助犯罪侦查。

1.2 研究现状和目的

1.2.1 研究现状简介

经过近几年的发展，社区发现的研究在很多领域取得了重要进展。大量实证研究表明，许多网络是异构的，即社会网络不是一大批性质完全相同的节点随机地连接在一起的，而是许多类型的节点的组合。相同类型的节点彼此之间存在较多的联系，因此具有较多的连接，而不同类型的节点之间的连接则与此相反。我们把满足同一类型中的节点以及这些节点之间的边所构成的子图称为网络中的社区(Community)。

与社区发现相关的理论包括图论[8]以及模式识别[9,10]等。社区发现的研究起源于社会学的研究工作[11]，Wu 和 Huberman[12]以及 Newman 和 Girvan[13]的研究成果，使得复杂网络中的社区发现成为近几年复杂网络领域的一个研究热点并形成了复杂网络中一个重要的研究方向。Newman 和 Girvan 把社区发现问题定义为将网络节点划分成若干组，使得组内节点之间的连接比较稠密，而不同组节点之间的连接则比较稀少。他们提出了基于边介数(Edge Betweenness)概念的分割方法，尽管该方法计算量很大，但由于其性能优越，从而成为社区发现研究的重要参考模型。其它一些方法涉及到很多概念，如电路理论[12]、超顺磁聚类[14]、网络的三角环(Triangular Loop)[15]、网络的谱性质分析[16]、网络的 Laplace 矩阵以及聚类技术[17]等等（具体的研究算法将在后面的章节中具体介绍）。

尽管社会网络中的社区发现问题得到了大量的研究，但还存在一些尚未解决的基本问题，如社区概念虽然大量使用，但却缺少严格的数学定义；大多数社区发现算法虽然性能优越，但所需计算量却很大。这说明社会网络中社区发现的研究还需要付出大量的努力。

1.2.2 研究目的

社区发现作为社会网络分析研究领域内的一个独立的研究课题，特别是基于 Blog、E-mail、电信通话记录等数据的研究，在国内外目前的研究过程中已经取得

了一些研究成果，并且得到了广泛的应用。

首先，在理论研究方面，社区发现研究有较高的学术价值。从广义上来说，社区发现的结果，即若干内部关系紧密的社区，可用于某个特定领域的推荐。但是对于各种社区发现的方法，从发现的效果、效率上来衡量方法的好坏的话，还有许多工作要做。特别是近些年来，随着一些社区发现的方法和技术相继出现，并应用于各行各业，其方法的优势和劣势也展现无遗，例如，当社会网络规模较大时，其方法的效果和效率都受到了影响，这些都成为目前社区发现领域受到国内外学者的研究焦点。

其次，近些年随着复杂网络研究的兴起，社区发现技术显得日趋重要。由于现实世界中复杂网络无处不在，并且复杂网络与社区发现技术联系紧密。尤其伴随着近年来对复杂网络性质的物理意义和数学特性的深入研究，人们发现了这些网络具有一个共同或相似的性质，整个网络是有若干个“社区”构成的，即每个社区内部节点之间的的连接相对非常紧密，但是各个社区之间的连接去比较稀疏。因此伴随着社区发现技术研究的同时，对正在兴起的复杂网络研究也起到了一定的推动作用。

另外，在实际应用方面，在复杂网络中自动搜寻或发现社区具有重要的实用价值，发现这些网络中的社区有助于我们更加有效的理解和开发这些网络，将其应用于我们的日常生活当中。例如，社会网络中的社区代表根据兴趣或背景而形成的真实的社会团体；引文网络中的社区代表针对同一主题的相关论文；万维网中的社区就是讨论相关主题的若干网站[18,19]；而生物化学网络或者电子电路网络中的社区可以是某一类功能单元[20,21]。

1.3 研究内容及主要工作

本文研究的对象主要是社区发现的算法及应用，主要工作是研究目前一些社区发现算法的效果及遇到的主要问题，同时对算法进行改进，通过实验进行效果比对，证明其改进效果，具体如下：

1、组织探索各种算法的核心知识点，适用范围、运算效率、准确度、以及可操作性。分析各种算法的优缺点，并分别对各种算法进行详细的介绍。

2、将著名的社区发现算法，GN 算法，作为实验对比算法进行研究，并完成设计与实现。将最新聚类算法，AP 算法应用于社区发现应用中，并将其作为重点研究对象，完成设计与实现。

3、通过实验对比，发现 AP 算法的一些不足，分析其不足的原因，并对其进行改进和实现。

4、实验结果分析，通过实验数据说明 AP 改进算法的效果，分析其优点和不足。

5、此外，还利用社会网络特点设计一个新的社区发现算法。

6、对将来的工作进行展望。

1.4 论文结构安排

本文共分为六章，文章结构及文章内容简介如下：

第一章 绪论。探讨了社区发现研究领域的背景及研究现状，说明了本文的研究目的，并分析了社区发现的研究意义，介绍了本文的研究内容及主要工作，最后给出了本文的结构安排。

第二章 相关理论知识。介绍了社会网络的涵义、研究内容、及其它相关概念，还详细介绍了社会网络的形式化表达。

第三章 社区发现的主要技术。本章详细介绍了几种社区发现算法——Belief Propagation(BP)算法，k-means(又称 k-centers)算法，基于密度的聚类分析(Density-based Methods)，基于网格的聚类分析(Grid-based Methods)和基于模型的聚类分析(Model-based Method)等方法，最后介绍了一些社区发现的典型算法，如 Kernighan-Liu 算法，谱平分法(Spectral Bisection Method)，W-H 算法，基于层次的聚类分析(Hierarchical Methods)，Girvan Newman 算法及派系过滤算法(Clique Percolation Method, CPM)。最后，本章还介绍了本篇文章的重点，AP 算法及其 AP 改进算法。

第四章 首先完成 GN 算法的总体设计与实现，详细介绍算法的设计与实现流程。然后对 AP 算法进行设计和实现，由 AP 算法的实验结果，并结合前面第三章的分析，对其进行改进和实现。最后，利用社会网络的特点，我们自己设计一个新的社区发现算法。

第五章 实验设计和结果分析。针对第四章中提出的算法，首先介绍了实验数据和实验评价标准，并提出实验方案，通过实验得到结果，并对结果进行分析。

第六章 结论与展望。对本文所做的工作进行总结，指出存在的不足之处，以及对未来工作的展望。

1.5 本章小结

本文介绍了社区发现技术的研究背景、研究现状、研究目的和意义，阐明了论文的主要研究内容和将要所做的主要工作，并对论文的结构安排进行了简要说明。

2 相关理论知识

由于社区发现技术在社会网络分析方面有广泛的应用，并且是社会网络分析的一个研究方面，所以在此首先要介绍一下有关社会网络分析方面的一些理论知识。

2.1 社会网络分析

二十世纪三十年代，Jacob Moreno 和哈佛大学的一组研究人员分别提出了社会网络模型来分析社会学中的现象和问题。社会学家发现社会实体之间存在着相互的依赖和联系，并且这种联系对于每个社会实体有着重要的影响。基于这样的观察，他们通过网络模型来刻画社会实体之间的关系，并进一步用来分析社会关系之间的模式和隐含规律。和以往社会学研究的方法不同，它提供了一种形式化、概念化的途径来看待“社会”这个研究对象的性质和发展进程，是一种应用性很强的社会学研究手段，有很多的应用。当社会学家建立一个准确一致的社会网络模型之后，就可以通过逻辑推理的方式来研究社会的性质。

由于数据收集方式的限制，早期的社会网络局限于一个小的团体之内，往往仅包含几十个结点。借助于图论和概率统计的知识，人工处理可以从中分析出一些简单的性质和模式。但是，随着现代的通信技术的发展，越来越多的数据被收集和整合在一起，建立一个大的社会网络成为可能。例如，可以通过电子邮件的日志来建立使用者之间的联系网络，或者通过网络日志及网络通讯录等方式将用户提交的联系人信息建立社会网络。所以，现在的社会网络规模比早期网络庞大，通常包含几千或者几万的结点，甚至有多达百万个节点的网络。面对这样庞大复杂的网络，简单的数学知识和原始的人工处理已经不可能进行有效的分析。社会网络分析是一种应用性很强的社会学研究方法，成功地解决了一些社会学问题上，得到了广泛的关注。随着信息技术的发展，越来越多的社会关系数据被收集。如果能够有效地对它们进行分析，必将加深人们对社会学的理解，促进社会学的发展。

从本体论的角度看，社会网络分析坚持一种实在论的本体论。认为社会结构是客观存在的各个行动者之间的关系，可以作为外在物对行动者产生作用。社会网络分析提供的就是对这种结构的分析，利用量化的语言对网络数据的结构进行描述。从认识论的角度看，社会网络分析认为世界是由网络而不是群体组成的，它把世界看成是网络的结构，把行动者之间的关系看成是资源流动物质的或者非物质的渠道，从而可以通过分析发现复杂的资源流动网络而不是简单的分层结构。

基于这种认识论，我们就可以根据行动者之间的关系模式来理解行动者的属性特征和网络的整体特征。它认为社会网络的结构特征决定了行动者之间关系，发生的环境只有在由各种关系构成的结构脉络中才能理解两个行动者的之间的互动关系。从方法论的角度看，社会网络分析用图论工具代数模型技术描述关系模式，认为从社会关系视角进行的社会学解释要优越于从个人属性的视角进行的解释。

2.1.1 社会网络的涵义

社会网络指的是社会行动者及其之间关系的集合。换句话说，一个社会网络是由多个点社会行动者和各点之间的连线行动者之间的关系组成的集合。用点和线来表达网络，这是社会网络的形式化界定。社会网络这个概念强调每个行动者都与其它行动者有或多或少的关系。社会网络分析者建立这些关系的模型，力图描述群体关系的结构，研究这种结构对群体功能或者群体内部个体的影响。下面对社会网络这个概念做进一步说明：

1、点　社会网络中的点是指社会行动者，边是行动者之间的各种社会关系。在社会网络研究领域，任何一个社会单位或者社会实体都可以看成点，或者行动者。例如行动者可以是个体、或集体性的社会单位，也可以是一个教研室、系、学院、学校，更可以是一个村落、组织、社区、城市、国家等，当然也包括网络上每一个虚拟社区的成员或社区本身。

2、关系　每个行动者是通过各种关系联系在一起。在社会网络分析中，一些得到广泛研究的关系有：

- 个人之间的评价关系如喜欢、尊重等。
- 物质资本的传递如商业往来、物资交流。
- 非物质资源的转换关系如行动者之间的交往，信息的交换等。
- 隶属关系如属于某一个组织。
- 正常的角色也是有关系性的，如教师学生、医生病人、老板职员关系等。
- 行为上的互动关系如行动者之间的自然交往，如谈话、拜访等。
- 生物意义上的关系如遗传关系、亲属关系以及继承关系等。

2.1.2 社会网络分析

社会网络分析作为人文社科领域内一门独立的学科，在国内外数十年的研究过程中已经形成了比较完整的学科体系，也得到了广泛的应用。社会网络分析主要是研究社会实体的关系连结以及这些连结关系的模式、结构和功能。社会网络

分析同时也可用来探讨社区中个体间的关系以及由个体关系所形成的结构及其内涵。换句话说，社会网络分析的主要目标是从社会网络的潜在结构中分析发掘其中次团体之间的关系动态。

社会网络分析主要研究行动者以及彼此之间的关系。通过对行动者之间关系与联系的连结情况进行研究与分析，将能显露出行动者的社会网络信息，甚至进一步观察并了解行动者的社会网络特征。而通过社会网络，除了能显示个人社会网络特征外，还能够了解许多社会现象。因为社会网络在组织中扮演着相当重要的无形角色，当人们在解决问题或是寻找合作伙伴时通常都是依循所拥有的社会网络来寻找最可能提供帮助和协作的对象。

社会网络分析是社会科学中的一个独特视角，它是建立在如下假设基础上的：在互动的单位之间存在的关系非常重要。社会网络理论、模型以及应用都是建立在数据基础上的，关系是网络分析理论的基础。网络模型把结构社会结构、经济结构等概念转化为各个行动者之间的关系模型。随着社会网络研究的深入，学者们渐渐在以下几个观点上达成共识：

- 社会行动者及其之间的行动是相互依赖的，而不是独立的、自主性的单位。
- 行动者之间的关系是资源（物质的或者非物质的）传递或者流动的“渠道”。
- 个体网络模型认为，网络结构环境可以为个体的行动提供机会，也可能限制其行动。
- 网络模型把结构（社会结构、经济结构等）概念化为各个行动者之间的关系模型。

2.1.3 社会网络分析的研究内容

社会网络分析用于描述和测量行动者之间的关系或通过这些关系流动的各种有形或无形的东西，如信息、资源等。自人类学家 Barnes 首次使用“社会网络”的概念来分析挪威某渔村的社会结构以来，社会网络分析被视为是研究社会结构的最简单明朗、最具有说服力的研究视角之一。20 世纪 70 年代以来，除了纯粹方法论及方法本身的讨论外，社会网络分析还探讨了小群体(clique)、同位群(block)、社会圈(social circle)以及组织内部的网络、市场网络等特殊的网络形式。

特别是近年来随着万维网的广泛应用和 Web2.0 理念的兴起，社区结构作为 Web2.0 网站的重要属性，因此这方面的研究蕴藏着巨大的商机，也是了解人类行为模式的基础，从而成为数据挖掘、信息检索、社会网络分析等领域的关注焦点。社区结构发现是将网络的整体结构分解为若干个社区，社区内部节点之间的连接相对紧密，而不同社区之间的连接相对稀疏。图 2.1 所示是一个小型网络中的社区

结构示意图，图中的网络包括3个社区，其中的虚线中的就是网络中的各个社区。

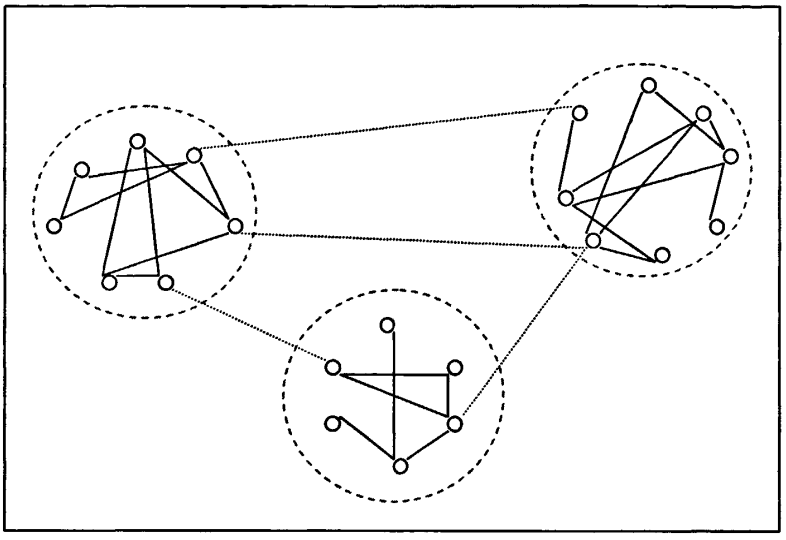

图2.1 一个具有社区结构的小型网络，图中有3个社区

2.1.4 社会网络分析的相关概念

如果没有数学工具图论、矩阵代数的支持，社会网络分析就不可能取得重要进展。在表达关系数据的时候，社会网络分析者主要利用数学领域中的两种工具社群图和矩阵代数。当然，社会网络方法论上的突破也离不开统计技术的发展。拥有了这两种工具，我们就能够计算一些网络测度例如密度、出入度等参数。在社会网络中，与“关联性”密切相关的研究是行动者之间的距离。有的行动者可能与网络中的任何一个人都建立了联系，与其他人的距离都很“近”。有的人可能交往比较少，相对“孤立”一些。如果行动者之间的距离不一样，我们就可能找到这些行动者在网络意义上的社会分层来，也可能有助于我们理解社会群体的“同质性”、“团结性”等特点。

下面将介绍几种距离相关的概念，并用这两种工具阐述它们在社会网络分析中所代表的含义。

1、图的概念

直观地说，给定多个点，把其中的一些点用曲线或者直线段连接起来，不考虑点的位置与连线的长短，这样所形成的点与线的关系结构就是一个图。即由点集合V和点与点之间的连线集合E所组成的集合对(V,E)称为图，用G(V,E)来表示。

V中的元素称为节点，E中的元素称为边。

若图中所有的边都具有方向，即区分它的起始节点与终止节点，则该图为有向图；若图中所有的边都没有方向，则该图为无向图；若部分边有方向，另一部分边没有方向，则该图为混合图。

若图中任意两个节点之间都存在边相连，则该图为完全图。

若图中任意两个节点之间的边只有存在于不存在之分，则该图为二值图。若图中任意两个节点之间的边都赋了权值，则该图为赋权图。

若图中任意两个节点都可以经过一系列的边互相到达，则该图为连通图。

非连通图中的最大连通子图称之为组件。

从图中取出一个点集，再加上这些点之间存在的关系，构成一个新的子图，这种生成子图的方式称为点生子图。

从图中取出一个边集，再加上与这些边相连的节点，构成一个新的子图，这种生成子图的方式称为线生子图。

2、点的度数

与某个节点相邻的那些节点称为该点的“邻点”(neighborhood)，一个点v_i的邻点的个数称为该点的“度数”(nodal degree)，记作$d(v_i)$，也叫“关联度”(degree of connection)。这样，一个点的度数就是对其邻点多少的测量。实际上，一个点的度数也是与该点相连的线的条数。如果两个点由一条线相连，称这两个点之间为“相邻的”(adjacent)。相邻是对两个行动者直接相关的图论表达。如果一个点的度数为 0，称之为“孤立点”(isolate)。度数这个概念在对有向图进行分析时必须考察线的方向，因此，一点的“度数”包括两类，分别称为“点入度”(in-degree)和“点出度”(out-degree)。一个点的点入度指的是直接指向该点的点的总数；点出度指的是该点所直接指向的点的总数。

3、线路、迹、途径

“线路”(walk)是由许多点和线首尾相接所构成的有序序列，这个序列起始于一点并且终止于一点，记作 W。在线路中，点和线都允许重复。“线路的长度”(the length of a walk)指线路中线的条数。

“迹”(trail)是特殊的线路，如果在一个线路中没有重复出现的线，则这样的线路叫做迹。

“途径”(path)也是一种特殊的线路，如果在一个线路中既没有重复出现的线，也没有重复出现的点，则这样的线路叫做途径。“途径的长度”(the length of a path)指构成该途径的线的条数。

4、测地线、距离、直径

在给定的两点之间可能存在长短不一的多条途径。两点之间的长度最短的途径叫做测地线。如果两点之间存在多条最短途径，则这两个点之间存在多条测地线。

两点之间的测地线的长度叫做测地线距离，简称为“距离”(distance)。也就是说，两点之间的距离指的是连接这两点的最短途径(shortest path)的长度。我们把点和之间的距离标记为：如果两点之间不存在途径即二者之间是不可达的，则称二者之间的距离是无限的或者无定义。如果一个图是不关联图，那么其中至少有一对点的距离是无限的。

一个图一般有多条测地线，其长度也不一样。我们把图中最长测地线的长度叫做图的直径(diameter)。如果一个图是关联图，那么其直径可以测定。如果图不是关联的，那么有的点对之间的距离就没有界定，或者就距离无穷大。在这种情况下，图的直径也是没有定义的。

5、图的密度

“密度”(density)这个概念是为了汇总图中线的总分布情况，以便测量该图与完全图的差距有多大。规模一定的点之间的连线越多，该图的密度就越大。具体地说，密度是指一个图中各个点之间相关联的紧密程度。

密度的形式化定义用图中实际拥有的线数与最多可能存在的线数之比来表示。假设一个图的实际线数为 l，节点数 n，则对于无向图而言，其密度为 2l/n(n-1)，对于有向图而言，其密度为 l/n(n-1)。

2.2 社会网络的形式化表达

从数学角度上讲，有两种方法可以描述社会网络：社群图法和矩阵代数方法。社群图法常常应用于结构对等性和块模型的研究。代数学法可用于分析角色和关系。当然，其他统计方法也可以用来描述社会网络。社群图是由莫雷诺最早使用的，现已在社会网络中得到广泛使用。用来表达一种关系的矩阵叫做社区矩阵。社群图主要由点（代表行动者）和线（代表行动者之间的关系）构成。这样，一个群体成员之间的关系就可以用一个出点和线连成的图表示。

如果根据关系（线）的方向，可以分为“有向图”和“无向图”。无向图是从对称图中引申出来的，它仅仅表明重要关系的存在与否。如果关系是有方向的（例如借款关系、权力关系等），也就是说，a 到 b 的关系与 b 到 a 的关系是不同的，那么，就应该用有向图来表示。我们用代表有向线的集合，用代表其中的单条线，用箭头代表关系的方向。行总和与列总和构成一个有向图及其邻接矩阵。

利用社群图表达关系网络的一个优点是比较清晰、明确，并且社会行动者之

间的关系一目了然。但是，如果社群图涉及的点很多，例如人，那么图形就相当复杂，很难分析出关系的结构，这是社群图的一个缺点。在这种情况下，我们最好利用矩阵代数法表达关系网络，用来研究多元关系，研究两种关系或者多种关系的“叠加”。这种方法最先由怀特和伯德提出来。

如果行和列都代表来自于一个行动者集合的“社会行动者”，那么矩阵中的要素代表的就是各个行动者之间的“关系”。这种网络是1-模网络。如果行和列代表来自两个行动者集合的“社会行动者”，那么矩阵中的元素分别代表的就是两个行动者集合中的各个行动者之间的“关系”，这种网络是2-模网络。如果“行”代表来自一个行动者集合的“社会行动者”，“列”代表行动者所属的“事件”，那么矩阵中的元素就表达行动者隶属于“事件”的情况，这种网络也是2-模网络，具体地说是“隶属关系网络”。

2.3 本章小结

本章从整体上介绍了社会网络的涵义，社会网络分析的概念、研究内容、及关概念，以及社会网路的形式化表达。

3 社区发现的主要算法及改进

3.1 社区发现的主要算法及成果

下面分别从概率图模型中的 Belief Propagation(BP)算法的提出和发展、聚类算法以及社区发现的已有算法等几个方面简要介绍前人的工作成果。

3.1.1Belief Propagation 算法

Belief Propagation(BP)算法是出机器学习领域泰斗 Pearl 提出的[22]，它的重要意义在于提出了一种有效的求解条件边缘概率的方法。

在现实世界中成千上万的因素相互联系，无法按照传统方法求解。人们曾提出许多近似推理方法，如变分方法(Variational Method)和基于随机抽样的蒙特卡罗方法(Monte Carlo Method)，但对于大规模复杂系统仍然难以实用。BP 算法改变了这一局面，Pearl 认为虽然影响世界的因素众多，但每个因素只与其他少数几个因素相关，这就构成推理网络。在概率图模型中，推理网络有两种：贝叶斯网络反映了因果推理关系；而马尔科夫网络反映了相互影响关系。在此基础上 Pearl 提出把推理局部化和分布化，把全局积分转变为局部的消息传递。节点通过与邻近节点交换信息来估算自身概率，如图 3.1 所示。通过这种方式，可以使算法复杂度由 $O(2^n)$ 降低为近似 $O(n)$，使得推理能在复杂系统中可用。

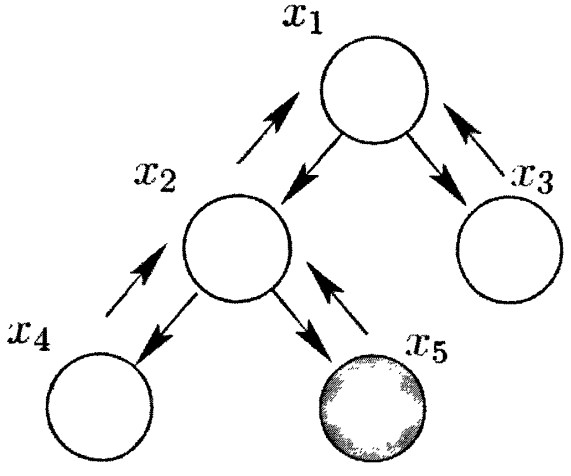

图 3.1 Belief Propagation(BP)算法的局部消息传递示意图

数学已证明，对于有向无环贝叶斯网络，BP 算法的解与严格积分的结果一致。这时的 BP 算法的功能只是利用联系的局部性降低运算复杂度。对于无向有环的马尔科夫网络，Pearl 曾指出这种信息传播可能导致不稳定：消息可能在带环的传播中无限加强，导致整个系统不能收敛。而对于大部分实际问题，BP 算法在带环系统中工作良好，从而发展成为 Loopy Belief Propagation 算法[23]，成为概率图模型中用于近似推理的经典算法之一。

3.1.2 聚类算法

基于相似度的数据聚类是数据挖掘领域的重要话题[24]，几十年来一直是研究者们关注的焦点。基于划分的聚类分析(Partitioning Methods)是最常用的聚类算法，一般计算得到若干聚类中心，使得它们与其邻近节点的方差之和最小，从而得到对数据的划分。例如经典的 k-means(又称 k-centers)算法[25]，随机选择聚类中心构成初始集合，然后不断改变该集合成员，直到找到方差之和(一种度量距离的方法)最小的集合。此外还有基于层次的聚类分析(Hierarchical Methods)、基于密度的聚类分析(Density-based Methods)、基于网格的聚类分析(Grid-based Methods)和基于模型的聚类分析(Model-based Method)等方法，都是根据不同任务提出的有针对性的聚类方法。而基于划分的 k-means 算法是应用最为广泛的聚类算法。

但是，k-means 算法存在的主要缺陷是，聚类效果严重依赖于初始聚类中心集合的选取，因此初始值敏感度比较高，因此许多次迭代，才能保证算法效果稳定性。而且算法复杂度较高，约为 $O(nkr)$，其中 n 为数据样本数目，k 为聚类个数，r 为迭代次数，因此 k-means 算法不适用于大规模数据分析和处理。

3.1.3 社区发现算法

传统的社区发现算法，主要来自计算机科学界和社会学界。并分别发展出不同的社区发现算法。计算机学界把社区发现任务描述为图分割问题(Graph Partitioning Problem)。多数图分割方法是基于迭代二分(Iterative Bisection)的：首先把整个图分解为最优的两个子图，然后对这两个子图分别进行最优分解，反复进行同样的处理，直到得到足够数目的子图。

1、图分割算法

两种著名算法分别是 Kernighan-Liu 算法[26]和谱平分法(Spectral Bisection Method)[27,28]。这些算法可以很好地解决那些事先知道社区数目的问题。

(1)Kernighan-Liu 算法

Kernighan-Liu 算法，简称 KL 算法，该算法基于贪婪算法原理可以将网络划分为两个大小已知的社区。其基本思想是为网络的划分引入增益函数 Q，表示两个社区内部的边数减去两个社区间的边数,然后寻找使 Q 值最大的划分方法。

需要说明的是，在交换节点对的过程中，增益函数 Q 的值并不一定是单调增加的。如果某次交换会导致 Q 有所下降，Q 必然会在其后的交换过程中得到更大的值。其算法复杂度约为 $O(n^2)$，其中 n 为网络节点数目。该算法的主要问题在于要求必须事先知道网络两个社区的大小。

算法的执行过程：

第一步：制定社区的规模（节点数）；

第二步：随机配置两个社区 A 和 B 内的节点；

第三步：定义增益函数 Q；

第四步：选定社区 A 中的一个节点 N；

第五步：计算将 N 与社区 B 中未被交换过所有节点的 Q 增益，即 $\Delta Q = Q_{交换后} - Q_{交换前}$，选中使 ΔQ 最大的节点 M；

第六步：交换 N 和 M；

第七步：重复执行第四步至第六步，直到社区 A 或社区 B 中的所有节点均被交换过。

(2)谱平分法

谱平分法是通过分析网络的拉普拉斯算子(Laplacian)的特征向量完成社区发现，在拉普拉斯矩阵的不为 0 的特征值所对应的特征向量中，同一个社区内的节点所对应的元素是近似相等的，这就是谱平分法的理论基础。

考虑网络社区结构的一种特殊情况：在一个网络中仅存在两个社区的情况下，网络的拉普拉斯矩阵就对应了两个近似的对角矩阵块。对一个实对称的矩阵而言，它的非退化的特征值对应的特征向量总是正交的。因此，除最小特征值 0 以外，其他特征值对应的特征向量总是包含正、负两种元素。这样，当网络由两个社区构成时，就可以根据非零特征值相应特征向量中的元素所对应的网络节点进行分类。其中，所有正元素对应的那些节点都属于同一个社区，而所有的负元素对应的节点则属于另一个社区。因此，我们可以根据网络的拉普拉斯矩阵的第二小的特征值 λ_2 将其分为两个社区。这就是谱平分法的基本思想。

当网络的确是分成两个社区时，用谱平分法可以得到非常好的效果。但是，其缺点在于每一次分割必须把网络分解成两个部分，通过反复调用算法来完成多社区划分，如果不能事先知道网络的社区数目，算法将很难达到满意效果。一般情况下谱平分法运算相当快，n*n 矩阵的特征向量的计算复杂度通常情况下为：$O(n^3)$。但在实际当中，由于大多数情况下拉普拉斯算子是个稀疏矩阵，特征向量

的计算可以用 Lanczos 方法在较短的时间内完成，该方法的时间复杂度大致为 $O(m)$，m 是网络中边的数量。

2、W-H 算法

传统的图分割法最大的缺陷就是它每次只能将网络平分，如果要将一个网络分成两个以上的社区，就必须对划分的子社区多次重复该算法。针对这个问题，Wu 和 Huberman[12]于 2003 年提出了一种基于电阻网络电压谱的快速谱分割法，简称 W-H 算法。

W-H 算法的基本思想是：如果将两个不在同一社区内的节点看成源节点（电压为 1）和终节点（电压为 0），将每条边视为一个阻值为 1 的电阻，那么，在同一个社区内的节点之间的电压值应该是比较接近的。因此，只要通过正确的方法找到源节点和终节点，选择一个合适的电压阈值，就可以得到正确的社区结构。

由于计算每个节点处的电压值需要求解拉普拉斯矩阵的逆矩阵，所需计算量通常为 $O(n^3)$，显然速度太慢。为此，Wu 和 Huberman 采用了下列近似方法：依次更新每个节点处的电压值为其相邻节点的电压值的平均值，如此进行多次，则将得到每个节点处的电压的近似值，并且运行次数与网络的大小无关。Wu 和 Huberman 已经证明了该近似算法是一种线性时间复杂度的算法。该算法与传统的谱分析算法一样，需要事先知道社区的数目。

W-H 算法的一个重要特点是可以在不考虑整个社会网络社区结构的情况下，寻找一个已知节点所在的整个社区，而无须计算出所有的社区。该特点在 WWW、Web 搜索引擎等大规模的网络中可以很好的应用。W-H 算法的不足之处是，如果预先不知道关于网络社区结构的部分信息，则很难应用该算法确定社区结构。

3、层次聚类法

由于事先并不知道一个网络中有多少个社区存在，各个社区所包含的节点个数也是未知的，使得社区发现问题是一个比图分割更加困难的问题。另外，网络的社区结构通常呈现出层次特征，一个社区可以进一步划分成几个子社区，也增加了社区发现的难度。社会学界中的社会网络分析则以社区结构作为研究对象，为此发明了层次聚类(Hierarchical Clustering)的分析方法。

社会网络分析中的层次聚类方法的思想更接近社区结构的思想，目的是根据各种衡量节点之间相似程度和节点之间连接的紧密程度的标准找出社会网络中的社区结构。

该方法基于各节点间连接的相似性或强度，把网络划分为社区。该方法根据向网络中添加边还是从网络中删除边可以分为两类：凝聚方法(Agglomerative Method)和分裂方法(Divisive Method)[29]。

(1)凝聚方法

凝聚方法首先计算各节点之间的相似性，然后从相似性最高的节点对开始（从高到低），向一个只有 n 个节点的无边网络（本来为空的网络）中添加边，这个过程可以在任意时刻终止，得到相应的社区结构。

算法的执行过程：

第一步：初始化起始状态为 n 个孤立节点；

第二步：计算网络中每一对节点的相似度；

第三步：根据相似度从强到弱连接相应节点对，形成树状图；

第四步：根据需求对树状图进行横切，获得社区结构。

凝聚法中边添加过程可用一个树状图(Dendrogram)来表示，如图 3.2 所示，为层次聚类常用的记录算法结果的层次树状图，按照图中虚线进行横切，可得到四个社区网络。

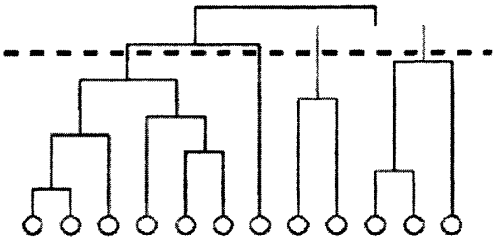

图 3.2 层次聚类通常采用层次图来记录算法结果

此外，在凝聚法过程中，根据社区的不同划分方法，可进一步细分为单连接法和全连接法，在此不作详细介绍。

(2)分裂方法

分裂方法与凝聚方法相反，是从图中逐渐删除相似度最低的节点对的连边进行的。分裂法的一般做法是找出相互关联最弱的节点，并删除它们之间的边，通过这样的反复操作将网络划分为越来越小的组件，连通的网络构成社区。在划分过程中，可以在任何时刻终止，并将当时的结果作为社区结构。其终止条件可以自己定义，比如划分的社区数量、社区内的节点数等。

算法的执行过程：

第一步：初始化网络；

第二步：计算网络中每一对节点的关联度；

第三步：根据关联度从弱到强，逐步删除节点对之间的边；

第四步：根据需求停止删除；

第五步：输出连通子网络，构成社区。

由于分裂法在寻找关联最弱的节点时，需要大量计算。对于稀疏网络，可能存在大量孤立节点不属于任何社区的问题，因此在实际应用时也难以取得令人满意的结果。

分层聚类方法的缺点是算法的适应性问题，算法对于那些社区结构已知的网络有较好的结果，比如在电影演员合作网络（演员被看成结点，如果两个演员在同一部电影充当角色，那么认为他们之间存在连接）中能够清楚地得到演员合演电影的个数。对于社区结构未知的网络，算法准确度较差。另外，算法虽然能对核心结点进行很好的分类，但是对于外围结点的分类却经常出错，如图 3.3 所示，核心结点（深色）在网络中有较强的连接，它们能够在算法执行的早期被连接在一起。浅色的为外围结点，这些点与社区往往只有一条边相连，图中结点 1 属于社区 A 是比较合适的。

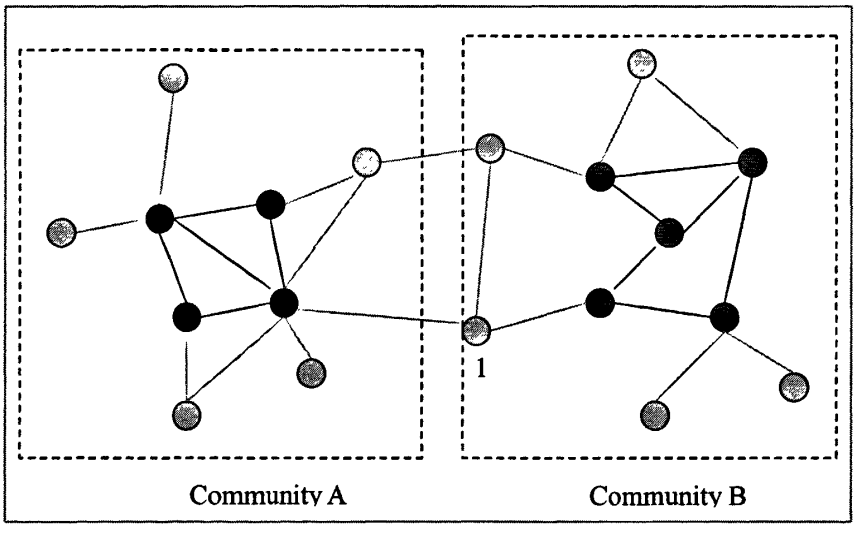

图 3.3 分层聚类得出的社区图

4、GN 算法

层次聚类方法实际上是一种基于加边的方法，正因为如此，应用层次聚类算法将导致若干节点无法确定归属。下面介绍的一个基于去边的方法。

Girvan 和 Newman 在 2002 年提出一种新的基于层次聚类的分裂算法[30]，称为 Girvan Newman 算法（简称 GN 算法），该方法是基于删边的方法。对于一般网络发现算法来说，其基本要求是可发现网络的最自然的分割。

基本思想是不断从网络中移除介数(Betweenness Centrality)最大的边。边的介数是指通过该边的最短路径的数目。由于同一社区内的节点对介数较小，而处在不同社区的节点对介数较大，因此可以比较好的划分社区。Girvan Newman 算法复

杂度为$O(m^2n)$，其中 n 是网络节点数，m 是网络边数，而对于稀疏网络，复杂度约为$O(n^3)$。

算法的执行过程：

第一步：计算网络中每条边的 Betweenness；

第二步：删除 Betweenness 值最高的边；

第三步：重新计算所有边的 Betweenness；

第四步：重复第二步和第三步，直到所有的边都被删除。

Girvan Newman 并不能提供判断社区划分准确性的标准，因此在不知道社区数目的情况下，该算法同样不能判断算法终止位置。因此 Newman 等人引入了一个衡量社区划分质量的标准——模块度(Modularity)[31-33]。Newman 认为，社区并不意味着绝对数量上社区之间的边少，而应该是比期望边数要少。模块度就是指落在社区内的边数减去随机生成图中的期望边数。形式化定义如下：首先设某种社区划分情形下，网络划分为 k 个社区，则定义 k*k 的矩阵$E=(e_{ij})$，元素e_{ij}表示网络中社区 i 与社区 j 之间的边占所有边的比例；矩阵的迹$Tr(E)=\sum_i e_{ii}$表示网络中社区内部的边占所有边的比例，矩阵中的第 i 行的和$a_i=\sum_j e_{ij}$表示与社区 i 中的点连边占所有边的比例。如果不考虑社区，假设节点之间随机连接，那么$e'_{ij}=a_i a_j$。因此模块度可以定义为：

$$Q=\sum(e_{ii}-a_i^2)=Tr(E)-\|E^2\| \quad \ldots(1)$$

其中$\|x\|$表示 x 所有元素之和。Newman 等学者发现，Q 的最大值与期望的划分位置密切相关，因此 Q 值可以作为社区划分方法正确性的判断标准。如图 3.4 所示是计算机生成的 64 节点、明显分为 4 个社区的人工网络，通过 Girvan Newman 算法生成的层次树状图在不同层次进行社区划分和对应的模块度值。可以看到，模块度在划分为 4 个社区的时候取最大值。

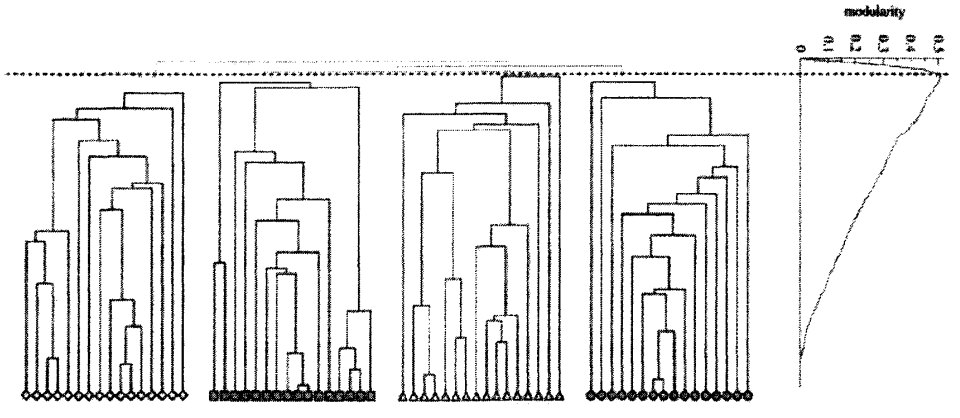

图 3.4 社区划分与模块度值比对

Girvan Newman 算法准确率较高，但算法复杂度比较大，因此 Newman 在该算法基础上提出一种快速算法[32]，简称为 Newman 快速算法。该算法实际上是基于贪婪算法思想的凝聚算法，通过模块度增量来指导合并的方向。该算法复杂度为$O(mn+n^2)$，对于稀疏网络约为$O(n^2)$。

从本质上讲，社区发现算法和聚类算法的任务是一致的，都是把根据距离等因素把数据划分为不同的部分，而社区发现算法实际上可以看作在网络上特殊的聚类问题。

3.1.4 其它社区发现算法

以上算法都将网络划分为若干互不重叠的社区，但现实中很多网络的社区结构并不互相独立，每个节点按照不同的角色同时属于多个社区，我们称之为社区交错现象，如图 3.5 所示。因此 Palla 等人于 2005 年提出了派系过滤算法(Clique Percolation Method,CPM)[34]分析这种互相重叠的社区结构，并给出了 k-clique 的社区定义。

k-clique 是一个全连通的子图，可以看成构成网络的基本元素，它们是重复出现的重要网络连接模式，这些小的有清晰定义的子图对理解网络的结构有重要的意义。k-clique 的社区定义：社区可以是为多个 k-clique 的连接，它们之间可以通过邻接的 k-clique 互相到达。这个定义的好处是能够表示一个社区内部连接的紧密性，说明了社区内部的成员是可达的。这类似于搭积木，每个 k-clique 为一个积木，整个社区是由多个积木搭建而成的。

基于 k-clique 定义的社区发现算法能够分析出社区之间的交错情况，首先求出网络中所有的 k-clique，得到网络的 clique 交错矩阵，然后通过分析该矩阵，得到划分的社区。算法的时间代价为指数形式。交错社区发现算法对真实数据的分类较为科学和合理。GN 算法及其改进算法均使用层次结构的树状图表示划分的结果，无法表示社区的重叠现象，Palla 给出的算法通过分析 clique 交错矩阵直接得到社区的划分，从而避免了层次树图的产生。

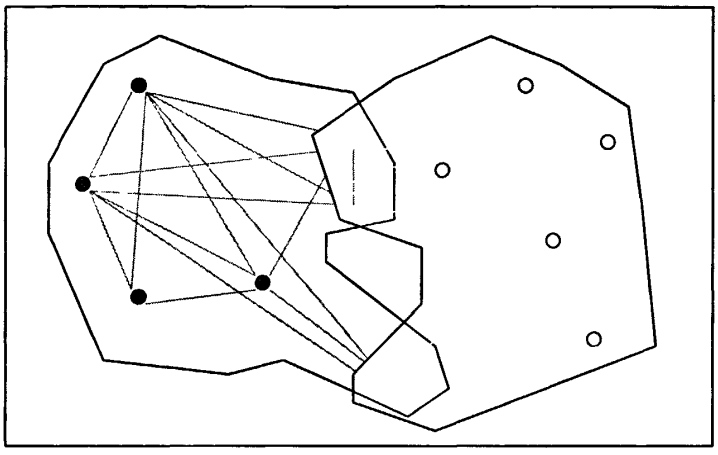

图 3.5 社区的交错现象，图中的空心结点均为社区的交错结点，同时属于多个社区

3.2AP 算法及改进

3.2.1AP 算法

2007 年 Frey 等在 Science 上提出 Affinity Propagation 算法[35]，即把每个数据看作网络中的一个节点，通过节点与邻居节点间的消息传递，最终涌现出聚类中心(exemplar)。每个时刻，传递的消息代表了节点选取其他节点作为聚类中心的亲密程度(affinity)。算法过程如图 3.6(A)所示，其中“exemplar”即为聚类中心，反映了聚类和聚类中心涌现的过程。

该算法以节点间的相似度$s(i,k)$为基础，$s(i,k)$表示x_k节点是x_i节点聚类中心的可能性。聚类实际上是以最小化方差为目标，因此相似度需要定义为方差的相反数（欧氏距离），如对于节点x_i和x_k，$s(i,k)=-\|x_i-x_k\|^2$。当然，$s(i,k)$也可以有其他定义方式。特殊的，对于$s(k,k)$表示节点 k 被选为聚类中心的可能性，称为偏好度(preference)。如果没有先验知识，所有节点的偏好度取相同的值，这个值决定了聚类数目，值越小聚类数目越少，可参考图 3.6(D)。

在节点间传递两种消息，分别为责任度(responsibility)$r(i,k)$，从节点x_i传递给聚类中心候选节点x_k，表示与其他x_i相比，x_k作为x_i的聚类中心程度；合适度(availability)$a(i,k)$，从聚类中心候选节点x_k传递给节点x_i，表示根据x_k邻居节点对x_k是否合适作为聚类中心的意见，x_k作为x_i聚类中心的合适程度。

在算法开始，所有合适度被赋值为 0。然后根据以下规则计算责任度：

$$r(i,k) \leftarrow s(i,k) - \max_{s.t.k' \neq k} \left\{ a(i,k') + s(i,k') \right\}$$

...(2)

而合适度可以通过以下规则计算：

$$a(i,k) \leftarrow \min\left\{ 0, r(k,k) + \sum_{i's.t.i' \notin \{i,k\}} \max\left\{ 0, r(i',k) \right\} \right\}$$

...(3)

而$a(k,k)$的更新方式为：

$$a(k,k) \leftarrow \sum_{i's.t.i' \neq k} \max\left\{ 0, r(i',k) \right\}$$

...(4)

以上的规则很简单，可参考图 3.6(B、C)。通过责任度和合适度之和可以确定聚类中心：对于节点x_i能够最大化$r(i,k)+a(i,k)$的 k 值对应的节点x_k是x_i的聚类中心。

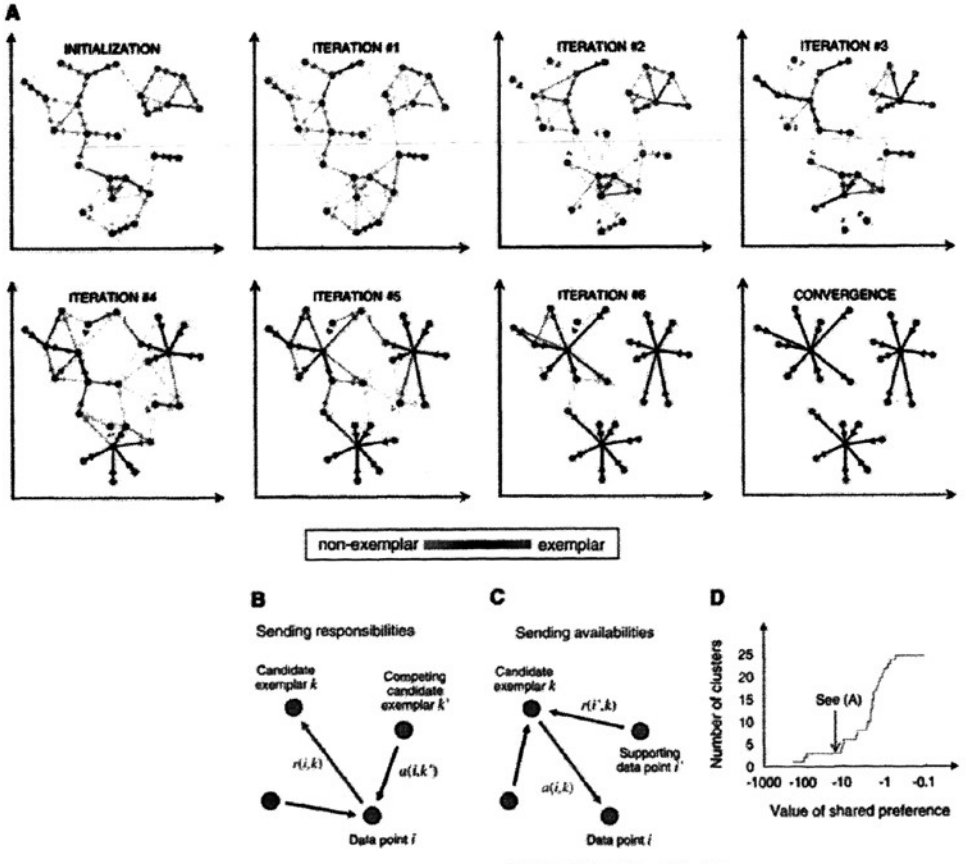

图 3.6 Affinity Propagation 算法迭代计算示意图[35]

Affinity Propagation 算法实际上是因子图(Factor Graph)[36]中的 BP 特例。如图 3.7 所示，网络中的信息传递可以表示成为这样的因子图，其中 s_i 表示 x_i 节点的聚类中心的下标，而 L_i 表示以 s_i 为聚类中心的 x_i 的似然概率。f_i 则代表了对聚类中心 x_i 的一系列约束条件。从 s 节点向 f 节点传递的是责任度信息，而从 f 节点向 s 节点传递的合适度信息。该算法的本质是学习得到一系列的聚类中心，这类似于混合模型(Mixture Model)中学习若干类别概率分布的方法。同时，它采用了 Pair-wise Clustering Algorithm 的方法，通过数据之间的相似度构造网络进行信息传递，因此该算法结合了聚类问题中两种重要方法的优势，并在算法稳定性上进行了改进[37]。

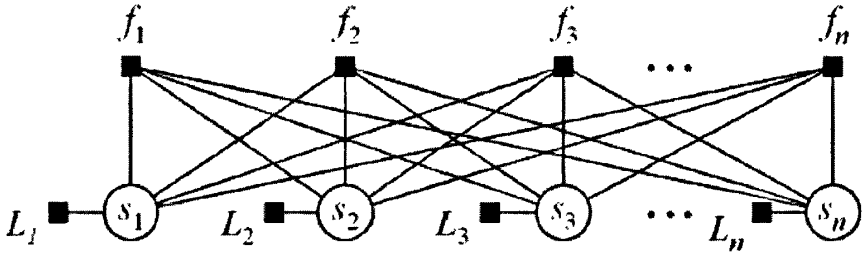

图 3.7 Affinity Propagation 可看作是因子图中的 BP 的特例[38]

Affinity Propagation 算法的优点除了在准确率、稳定性和速度上胜人一筹外，它的最大优点是对数据要求很低：(1)该算法不要求相似度的对称性，即可以处理 $s(i,k)\neq s(k,i)$ 的情形；(2)该算法不要求相似度满足三角不等式，即可以处理 $s(i,k)<s(i,j)+s(j,k)$ 的情形。实际上，在实际网络，尤其是社会网中，节点间往往不满足对称性和三角不等式。例如 x_i 把 x_k 看作最重要的合作伙伴，而 x_k 可能只把 x_i 普通合作伙伴；同样，x_i 与 x_j 熟识，x_j 与 x_k 熟识，但很有可能 x_i 与 x_k 并不熟识。因此，可以推测 Affinity Propagation 算法对于社区结构分析具有极大的优势。

采用 Affinity Propagation 算法进行社区发现的重要步骤是定义网络节点的相似度 $s(i,k)$。本文将主要探讨在无向无权图上如何应用 Affinity Propagation 算法进行社区发现。对于无权网络 $G=(V,E)$，其中 V 表示节点集，E 为边集，设该网络节点 x_i 与 x_k 的最短路径为 $p(i,k)$，节点 x_i 与 x_k 的相似度 $s(i,k)$ 有各种定义方式：

节点 x_i 与 x_k 之间有边，相似度定义为负数 $-v$，没有边定义为负数 $-w$，其中 $w>v>0$。即：

$$s(i,k)=\begin{cases}-w\\-v\end{cases}\quad -w<-v<0 \qquad \ldots(5)$$

相似度定义为节点 x_i 与 x_k 间最短路径的相反数。即：

$$s(i,k)=-p(i,k) \qquad \ldots(6)$$

如果两节点间没有连接，则定义$s(i,k)=-w$，其中 w 大于网络直径$d=\max_{i,k\in V} p(i,k)$，即网络任意节点之间平均最短路径的最大值。

相似度可以定义为连边介数的相反数。即：

$$s(i,k)=-b(i,k) \quad \ldots(7)$$

其中$b(i,k)$表示x_i与x_k间边的介数。

3.2.2AP 算法的改进

AP 算法的作者在文章中指出，AP 输出的聚类数目依赖于输入的偏好度 preference(后面都用 p 来简化表示)。而能否有效选择好偏好度，能否找到更好的相似度度量方法，能否在无向无权图上找到足够的相似度信息，这些都会影响偏好度的选取。

在目前相似度度量方法和相似度信息无法得到更多改进和帮助的情况下，偏好度 p 取何值能产生最准确的聚类结果是不得而知的，特别是在后面的真实数据的实验中，AP 算法虽然表现不错，但离真实情况还是存在一定差距的，这就需要对 AP 算法进行分析并改进，特别是在偏好度的选择上进行改进。

经过分析得知，由上一小节公式(2)、(3)所示，当 i=k 时，$r(k,k)\leftarrow s(k,k)-\max\{a(k,k')+s(k,k')\}$，偏好度$s(k,k)=p(k)$。当$p(k)$较大时，$r(k,k)$也变得较大，促使$a(i,k)$也增大，那么点 k 作为聚类结果中的一个类可能性就会增大；同样，当其它的节点$p(i)$相对而言也较大时，其它结点作为聚类中心的可能性也会随之增大，那样最终聚类结果将随着偏好度的变化而变化，很有可能造成结果的多样性，这样是不行的。因此，偏好度 p 在一定程度上影响着 AP 算法输出的聚类数目。而 AP 算法的作者并没有解决这个问题，虽然作者建议将偏好度设为为p_m（s 中元素的中值），但不能排除偏好度设置为其它值（非中值）时会降低最终结果的准确程度，很有可能还会提高准确度，这些都是未知的。

由于到目前为止，偏好度与输出的聚类数目之间，还没有找到它们之间某种一一对应的关系，也没有任何现实规律可循，所以可以考虑将偏好度的选择从某一个数值的选取，改为考虑在一定范围大小的数值域内选取。同时，在算法的改进过程中，为了继承和保留 AP 算法的一些不容忽视的优秀特性，其算法的总体内容和思想不会被改变，只是给它另加一些“东西”，以获得更好的效果。算法从初始给定的偏好度出发（可以考虑 AP 算法作者提出的中值p_m设想），运行过程中让其不断更新其责任度 r 和合适度 a；若运行过程中收敛到类数 K，以一定幅度p_{range}逐步减小偏好度，照此重复多次，以获得多个聚类结果。

3.3 本章小结

本章从整体上介绍了社区发现的主要算法，及本文将要重点介绍的 AP 算法及 AP 算法的改进。

4 社区发现算法设计与实现

4.1GN 算法的设计与实现

GN 算法是社区发现技术发展过程中的一个重要里程碑，它从网络的全局结构出发，避免了传统算法的若干缺点，成为目前进行网络社会分析的标准算法，得到了广泛的应用。因此，我们选择 GN 算法作为对比算法。

4.1.1GN 算法的设计

由于 GN 算法看作是一种分裂法，与一般分裂法不同的是，GN 算法不是寻找关联最弱的节点对，然后删除它们之间的边，而是寻找 Betweenness 值最高的边并删除它。就是寻找并删除那些连接节点对的边。GN 算法包括计算网络中每条边的介数、去除边介数最大的那条边、重复进行直至网络中没有任何边存在。其具体流程图如图 4.1 所示。

4.1.2GN 算法的实现

```
第一步：分析文件，建立群
init2();//初始化
checkout_pass_num();//计算边介值
make_group();//建立群
第二步：忽略文件中不合格群
for(i=0;i<group_end;i++)
{
if(group[i].count<the_shortest_num)
{
group[i].enable=false;//使群无效
int j=0;
while(group[i].arc[j]!=-1)
{
net[group[i].arc[j]].nelgect=true;//使该群中的边忽略
j++;
}
```

```
}
}
第三步：判断挖掘条件，数据挖掘
while(!no_edge())//如果有边可以删
{
data_mining();//数据挖掘
}
```

4.2AP 算法的设计与实现

Girvan Newman(GN)算法作为一个非常经典的社区发现算法，是社区发现技术发展过程中的一个重要里程碑，它从整个网络的全局结构出发进行社区识别，避免了其它传统算法的很多缺点，成为目前进行社会网络社区分析的标准算法，得到了广泛的应用。但是，GN 算法的缺点也暴露无遗：首先，因为要重复计算边的 Betweenness 值，而每次重复过程都要计算每对节点间的最短路径，算法的时间复杂度高；其次，GN 算法无法预知网络最终应该分裂成多少社区，而且通过树状图把网络分解到节点，每一个节点都必须属于某一个社区，这就对数据的要求比较高。所以，针对 GN 算法的缺点，在此我们选择了 AP 算法作为主要研究对象。

4.2.1AP 算法的设计

AP 算法的设计，关键是定义网络节点的相似度 $s(i,k)$。本文主要探讨的是在无向无权图上的应用，因此对于无权网络图 $G=(V,E)$，其中 V 表示节点集，E 为边集，设该网络节点 x_i 与 x_k 的最短路径为 $p(i,k)$，节点 x_i 与 x_k 的相似度 $s(i,k)$ 我们选择最短路径的定义方式。

相似度定义为节点 x_i 与 x_k 间最短路径的相反数，即 $s(i,k)=-p(i,k)$。

算法设计如下：

输入：

s(i, k)：节点 i 到节点 k 的相似度

preference(k)：聚类中心候选节点 k 的偏好度

输出：

K：聚类中心的个数

在此我们将 AP 算法设计成如下步骤：

第一步：初始化合适度 $a(i,k)=0$；

第二步：通过公式(2)更新责任度；

第三步：通过公式(3)更新合适度，并通过公式(4)更新自己的合适度；

第四步：经过$r=(1-lam)*r+lam*r_{old}$和$a=(1-lam)*a+lam*a_{old}$一个固定次数的迭代后，该消息传递程序被终止。

4.2.2AP 算法的实现

对于我们使用到的第二个算法 AP 算法，具体代码描述如下：

```
N=size(S,1); A=zeros(N,N); R=zeros(N,N); % 初始化信息
S=S+1e-12*randn(N,N)*(max(S(:))-min(S(:)));
lam=0.5;
for iter=1:100,
    %计算责任度
    Rold=R;
    AS=A+S; [Y,I]=max(AS,[],2);
    for i=1:N, AS(i,I(i))=-realmax; end;
    [Y2,I2]=max(AS,[],2);
    R=S-repmat(Y,[1,N]);
    for i=1:N, R(i,I(i))=S(i,I(i))-Y2(i); end;
    R=(1-lam)*R+lam*Rold; %责任度
    %计算合适度
    Aold=A;
    Rp=max(R,0); for k=1:N, Rp(k,k)=R(k,k); end;
    A=repmat(sum(Rp,1),[N,1])-Rp;
    dA=diag(A); A=min(A,0); for k=1:N, A(k,k)=dA(k); end;
    A=(1-lam)*A+lam*Aold; %合适度
end;
E=R+A;
I=find(diag(E)>0); K=length(I); %聚类中心
[tmp c]=max(S(:,I),[],2); c(I)=1:K; idx=I(c);
```

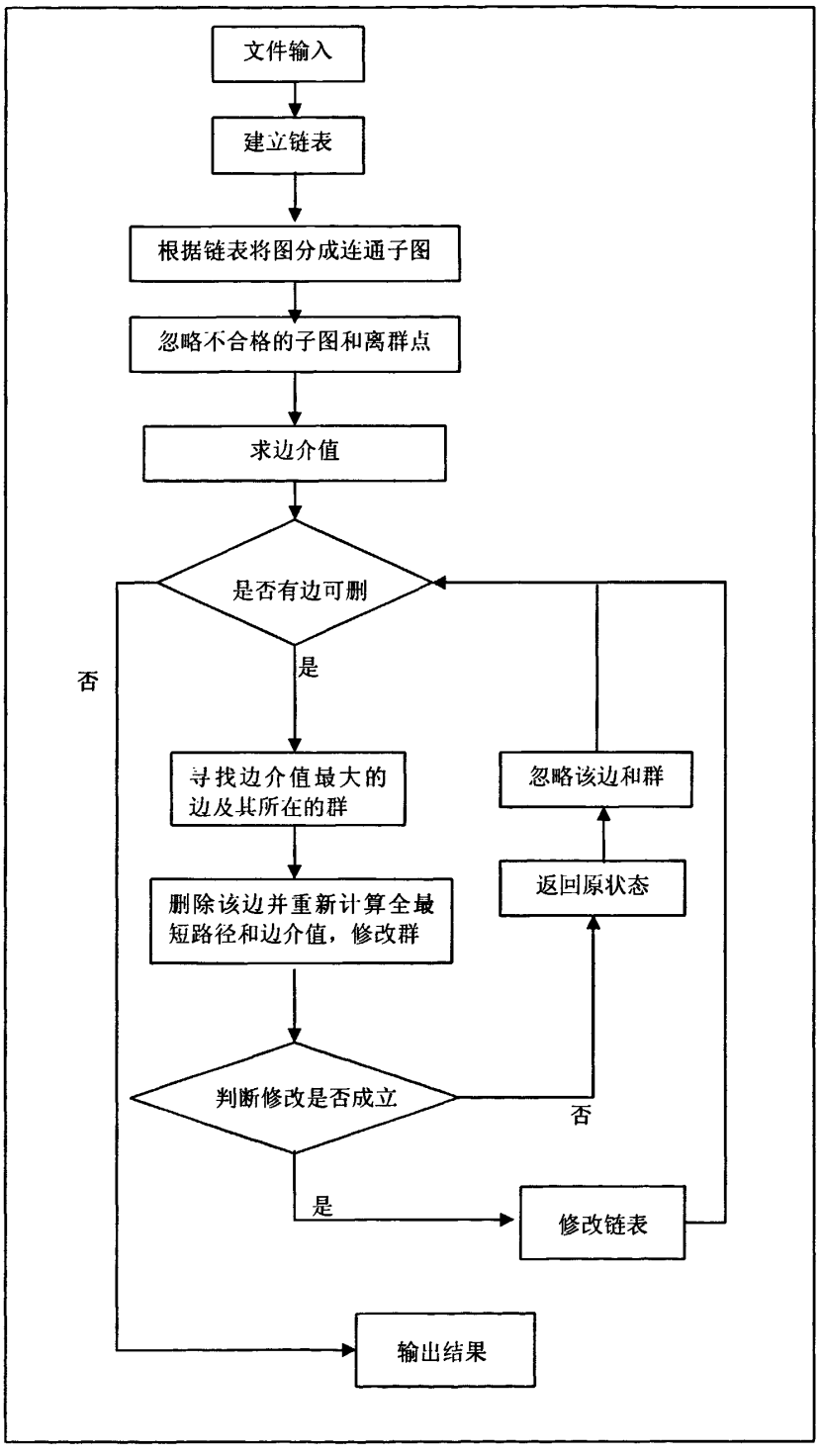
文件输入
建立链表
根据链表将图分成连通子图
忽略不合格的子图和离群点
求边介值
是否有边可删
否
是
寻找边介值最大的边及其所在的群
删除该边并重新计算全最短路径和边介值，修改群
判断修改是否成立
否
是
忽略该边和群
返回原状态
修改链表
输出结果

图4.1 GN算法流程图

4.3AP 改进算法的设计与实现

正如前面 3.2.2 节介绍的，由于 AP 算法中偏好度定为 p_m（s 中元素的中值），但该方法不能代表所有的情况，只能算是权宜之计，很有可能出于 p_m 设置不当，造成某个社区的丢失，所以我们提出了对 AP 算法进行改进。

本节将详细介绍 AP 改进算法的设计与实现。

4.3.1AP 改进算法的设计

在偏好度 p 选取过程中设计每次减小 p（即减小 $s(i,i)=p(i)$）后，以当前的责任度 $r(i,j)$ 和合适度 $a(i,j)$ 值作为新迭代的起点，继续更新 $r(i,k)$ 与 $a(i,k)$。

设计如下：

第一步：首先选择一个起始偏好度（可以使用 AP 算法作者提出的中值 p_m）；

第二步：通过原始 AP 算法得到类的数目；

第三步：检测这个数目是否收敛；

第四步：若收敛，进行下一步，否则，让算法继续重复执行直至收敛；

第五步：若连续数次循环，结果均收敛到某一个值，则以某个小幅度减小偏好度，否则转第二步；

第六步：继续执行第二步。

其中，下降幅度设定要适中。由于下降幅度太小或太大时都会对算法运行产生不利影响，太小时运算时间长，丢掉了 AP 算法的特点；太大时有的社区或类有可能被错过。因此固定的步幅难以确定，对不同类数很难找到一一对应的幅度。于是我们设计幅度调整功能，调整系数 $\alpha=0.015(N+60)$，幅度 $p_{range}=0.02p_m/\alpha$。这样，算法在运行的过程中可以适当灵活的调整 α，使当类的数目较大时下降步幅小一些，反之则大一些。此外，我们将得到的聚类结果通过聚类有效性指标 Silhouette 指标来确定那个聚类数目为最终结果[39]。

4.3.2AP 改进算法的实现

改进算法的具体实现如下：

初始化：$p \leftarrow p_m/2, p_{range} \leftarrow 0.02p_m, d \leftarrow 0, rec \leftarrow 0$

for i←1 to 400 do

N(i)←N　% 应用 AP 算法，N 是聚类的数目

if 如果点 w 是聚类中心

then W(w,j)←1　% 记录点 w 收敛次数

else W(w,j)←0

%调整偏好度幅度及偏好度

if 这 N 个类收敛，且连续 20 次收敛

then Converge←1

else Converge←0, d←0, rec←0

rec←rec+1

if Converge = 1 and rec >= 5

then d←d+1

$\alpha = 0.015(N+60)$

$p \leftarrow p + d * p_{range} / \alpha$

rec←0

if N <= 2

then 停止

4.4 基于联系关系的社区发现

4.4.1 算法的思路

很多社区发现的算法都是将关系复杂的社会网络分解成一个个内部关系更加紧密的小社区，虽然目前的很多算法能够产生不错的效果，但是不一定都能满足我们的要求，而且有些方法比较复杂。通过社区发现技术的挖掘，人们往往想要的社区是一个不仅在某一方面有共同的“爱好”，而且最好能显示出梯度的特性，和一定的对比效果。

因此我们可以从社会网络的特点出发，即社会是由人构成的，用人日常的思维习惯来分析研究，通过简单算法来实现类似社区发现的效果。由于社区发现是在社会网络中进行，即以社会为背景，可以利用社会的特征，以及人与人之间的关系亲疏紧密的特点对社会网络进行分析和挖掘，并设定一定的要求。例如，利用人们通常的想法，即“与权威专家合作的人必是优秀的人”，并结合“六度分割”假设，即地球上的任何一个人能通过不到五个中介人的一连串联系与地球上的另一个人联系上。该算法还有一个不同之处在于，很多社区发现算法基于对网络的划分，我们设计的算法则是利用网络中个体之间的关系，对个体进行抽取组成一个社区。

而且该算法也有很好的通用性，在别的应用领域，如犯罪侦查，可以通过与犯罪嫌疑人关系的亲疏程度来寻找线索，我们的设计思想可以适用。

首先将社会网络描述成一个图$G=(V,E)$，其中$v\in V$代表在社会网络中的一个人（点）；$e_{ij}^{r}\in E$代表了一个类型为r(relation)的关系，例如合著关系，表示人v_i和v_j之间的关系情况；并且每一个关系e_{ij}^{r}都被分配了一个权值$w(e_{ij}^{r})$(weight)，以此来表示两个人之间关系的亲近程度。之后，利用著名的最短路径算法 Dijkstra 算法[40,41]要找到从一个人v_i到另一个人v_j的关系队列。将这些队列中的人查找出来，并由这群人组成一个社区。

对于起始人v_i和目标人v_j的选取，我们可以使用两种方法：

1、人为选取——权威专家

对于某一个研究领域的权威专家，现实社会中信息流通发达，想找到一个权威专家不是一件困难的事情。例如，如果你是从事数据挖掘这个领域研究的，你肯定知道几个知名的专家学者。我们就可以如此选取我们所知道的两个权威专家，分别设置为v_i和v_j。

因为我们使用的是 DBLP 数据，里面存储的全部是学术研究领域的信息，任何专家都是可以找到的，所以此方法是行得通的。当然，如果换个别的数据集，如公安机关记录的犯罪嫌疑人的犯罪记录及同伙联系网，此方法也是行得通的。

2、技术选取——点度中心度

“点度中心度”(degree centrality)是一个最简单、最直观的指标，它描述的是一个节点位于图中“核心”位置的程度，刻画了该点与图中其他点发展交往关系的能力。

点度中心度又分为绝对点度中心度和相对点度中心度，点v_i绝对点度中心度(用C_{ADi}表示)就是该点的度数，即：

$$C_{ADi}=d(v_i) \qquad \ldots(8)$$

点v_i相对点度中心度(用C_{RDi}表示)是其绝对点度中心度的标准化形式，为该点的度数与其最大可能的度数之比，即：

$$C_{RDi}=\frac{d(v_i)}{v-1} \qquad \ldots(9)$$

其中 v 为图中的节点数。

4.4.2 相关符号及概念的定义

1、联系关系$a(v_i,v_j)$(association)：就是满足$e^r_{m(m+1)}\in E$，代表了一个类型为 r 的关系，表示人v_i和v_j之间的关系情况，而 E 为一系列关系队列$\{e^r_{i1},e^r_{12},\cdots,e^r_{lj}\}$，其中$m=1,2,\cdots,l-1$，其中$v_i$和$v_j$分别代表起源人和目标人。并且每一个联系关系被分配了一个分数：$p(a_k(v_i,v_j))$(preference)。

2、查找：找到从v_i到v_j所有可能的联系关系$\{a_k(v_i,v_j)\}$，给每一个联系关系分配了一个分数$p(a_k(v_i,v_j))$，最后返回这些联系关系，并按照分数越小越好的原则罗列：$A(v_i,v_j)=\{(a_k(v_i,v_j),p(a_k(v_i,v_j)))|p(a_k(v_i,v_j))<p(a_{k+1}(v_i,v_j))\}$。一个关联的分数被定义为公式(10)：

$$p(a_k(v_i,v_j))=\sum_{m=1}^{l}w(e^r_{m(m+1)}) \qquad \ldots(10)$$

4.4.3 算法的设计

在这个算法中，将联系关系查找问题看作是最近最短路径问题。

输入一个查询(v_i,v_j)。目的是找到所有联系关系$A(v_i,v_j)=\{(a_k(v_i,v_j),p(a_k(v_i,v_j)))\}$，并将其罗列出来。为了描述的方便，我们将省去$(v_i,v_j)$，并写成$A=\{(a_k,p(a_k))\}$。

下面我们将详细介绍算法中的每一步：

我们将社会网络看作是一个网络交际图，给社会网络中每一个关系 e 设定一个权值$w(e)$，权值的设定我们有不同的计算方法，在这里，由于使用的数据集记录的是每一个研究员的学术交流信息，在此，我们用研究员之间的合著次数作为两个人之间关系的权值；其次，我们的目的是在计算过程中，找到所有人$v\in V/v_j$到达目标人v_j的最短关联（包括从v_i到v_j最短的一条路径$Path_{\min}$），在图中，找到两点之间最短路径，我们使用的是基于堆栈的 Dijkstra 算法来找到最短路径；而后，找到最近最短联系关系，基于在第二步中找到的最短路径$Path_{\min}>0$，和之前定义的参数α，我们找到小于$(1+\psi)Path_{\min}$的所有联系关系，这样限制了查找结果中路径的长度，减少了计算量；最后，分数$p(a)$是基于公式(10)计算的，在第三步中找到的联系关系，全部基于分数进行罗列，并且依据‘较低的最好’的原则按分数从低到高排列。

4.4.4 算法的实现

具体的算法实现如下所示：

其中$Path_{\min}$代表最短路径值，而ψ是人为设定的系数。

{

%第一步，初始化权值

对于每一个$(e_{ij} \in E)$

$\{w(e_{ij}) \leftarrow w_{ij}$，对每一对人$v_i$和$v_j$设定的权值。$\}$

%第二步，找到最短路径，即联系关系

找到从每一个结点$v \in V$到结点 z 的最短路径$p'(v)$

%利用基于堆栈 Dijkstra 算法

{

对于每一个$(v \in V \setminus z)$

$\{p'(v) \leftarrow 8; path(v) \leftarrow -1; f(v) \leftarrow 0;\}$;

$p'(v) \leftarrow 0$;

heap←创建一个最小堆栈;

将 z 插入堆栈中;

While(堆栈不为空)

{

v_{min}←从堆栈中去除最小结点;

$f(v_{min}) \leftarrow 1$;

$E(v_{min})$←所有的边指向结点v_{min};

对于每一个$(e_{min} \in E(v_{min}))$

{

(t, v_{min})←边经过e_{min}表示(t, v_{min})权值;

if$(f(t)=0 \&\& w(e_{min})+p'(v_{min}) < p'(t))$

{

$p'(u) \leftarrow p'(v_{min}) + w(e_{min})$;

$path(u) \leftarrow v_{min}$;

if(t 在堆栈中)

{

将 t 从堆栈中移出;

} else

{

将 t 插入堆栈中;

}

}

}

}

}

%找到最短联系关系

对于每一个$(v \in V/v_j)$

{$p'(v) \leftarrow$找到从v到v_i的最短联系关系；}

$Path_{\min} \leftarrow p'(v_i)$；

%第三步，找到最近最短联系关系

设置堆栈$\leftarrow (v_i, NULL)$；

%$p(v)$用来表示当前联系关系的分数

对于每一个$(v \in V)$

{$p(v) \leftarrow 0$； $f(v) \leftarrow 0$；}

$f(v_i) \leftarrow 1$；

While(堆栈不为空){

$(h,e) \leftarrow$结点在栈的顶端；

$E(h) \leftarrow$所有边都从结点 h 指出；

对于每一个$(e_h \in E(h))$ {

$(h,t) \leftarrow$边经e_h指向；

if$(f(t)=0 \&\& p(h)+w(e_h)+p'(t)<(1+\alpha)L_{\min})$ {

if$(t=v_j)$ {

$a(v_i,v_j) \leftarrow$所有的边在堆栈中，包括e_h；

$p(a(v_i,v_j)) \leftarrow p(h)+w(e_h)+p'(t)$；

将$(a(v_i,v_j), p(a(v_i,v_j)))$放入结果列表 A 中；

}else

{if((堆栈的大小)<最大长度)

{push(t,e_h)于堆栈中；

$f(t) \leftarrow f(t)+1$； $p(t) \leftarrow p(h)+w(e_h)$；

}

}

}else

{pop(h,e)从堆栈中推出； $p(t) \leftarrow p(h)+w(e_h)$；

}

}

}

%得到结果

联系关系列表 A;

```
}
```

4.4 本章小结

本章介绍了 GN 算法、AP 算法、AP 改进算法的设计与实现，并且还介绍了另一种基于联系关系的社区发现算法，利用我们熟识的 Dijkstra 算法来实现，以此来说明通过简单的方法也可以达到社区发现的效果。

5 实验设计和结果分析

5.1 实验数据

我们使用的实验数据是共两部分。一部分则从网络上下载的以现实生活为背景并被很多研究人员使用的数据[42,43]，另一部分主要从著名的开源数据集 DBLP 中提取。

5.1.1 现实生活数据

虽然网络已经为我们的社区发现算法提供了数量庞大，甚至是可控制的测试数据平台，但是有些方面它显然还是无法与来自真实世界并专门用来测试算法的数据相媲美。为此了更好的进行实验对比，我们还选择了几组代表真实世界的网络的数据集，数据集中的社区结构已经知道。其中就有众所周知的 Zachary 空手道俱乐部社会关系网。20 世纪 70 年代初期，Zachary 用了两年的时间观察美国一所大学的空手道俱乐部成员间的社会关系，基于这些成员在俱乐部内外的社会关系，构造了他们的社会关系网。而碰巧在他调查过程中，该俱乐部的主管和校长之间产生争执，俱乐部分裂成为两个分别以主管和校长为核心的小团体，如图 5.1 中的节点 1 和 33 分别代表主管和校长，而方形和圆形也分别代表了分裂后的小团体中的各成员。

此外我们还选取了其它几组数据进行实验分析，这些都是在很多论文中经常被用作实验的数据集，其下载地址为：http://www-personal.umich.edu/~mejn/netdata。

5.1.2 DBLP 数据

DBLP 不是数据库，而是服务器，主要提供计算机科学期刊和程序方面的书目信息。最初的服务器主要集中在数据库系统和逻辑编程，其英文为名称为 Data Base systems and Logic Programming (即 DBLP)，现在它正在逐步扩大到计算机科学的其他领域。现在，我们可以将“DBLP”看作是“数字的参考文献和图书馆项目”。DBLP 中收录有大量的学者信息，如发表的期刊、杂志、文章的题目，作者及合著者的姓名等相关信息，信息非常全面。并且由于 DBLP 数据集是开源的，并且是免费的，所以我们提取 DBLP 中的一部分作为实验用数据。

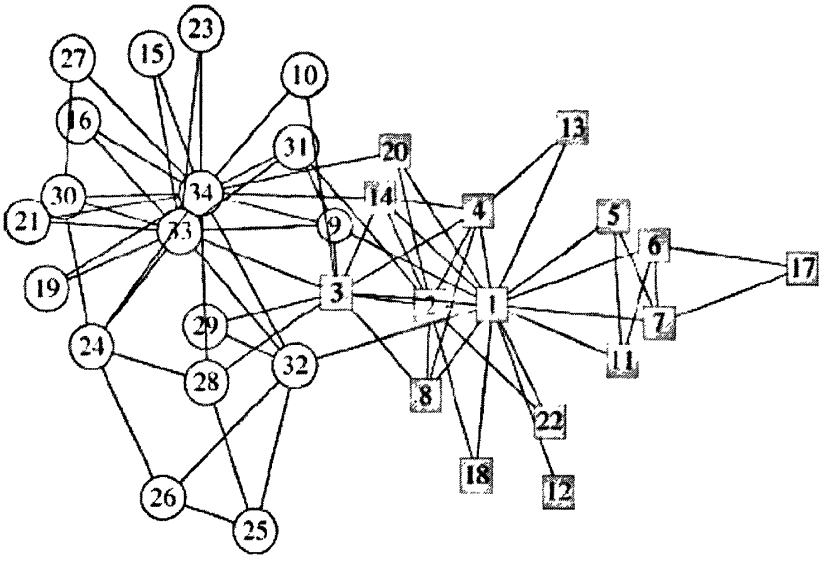

图 5.1 Zachary 空手道俱乐部社会关系网示意图

5.2 实验评价标准

在本文的社区发现研究实验中，我们设定了几个评价算法好坏的标准，下面分别介绍。

5.2.1 时间复杂度标准

随着数据量的增加，进行社区发现时需要耗费的时间也随之增加，现在很多研究都把实验过程中时间效率作为评价算法好坏的标准。

因此我们把时间复杂度也作为社区发现技术评价的一个标准，例如 Girvan Newman 算法，虽然准确率较高，但当社会网络规模庞大时，算法时间复杂度也随之增大，时间效率不得不纳入考虑的范畴。而 Affinity Propagation 算法的优点就是在速度上更优秀，这已经是被证明的了，而我们的 AP 改进算法，它的运行效率又如何呢。我们通过在不同数据集上的实验来进行评价。

5.2.2 准确度标准

社区发现研究的目的就是要准确发现所有内部联系紧密的社区。但由于社会网络内部纷繁复杂，对其内部的社区数目我们无法准确预先得知，如科学家关系

网络，里面的数据十分丰富，但我们无法预知里面准确的社区数量。而 Girvan Newman 算法是目前被公认的在社区发现算法里准确率较高的经典算法，所以将其算法作为对比算法来评价 AP 算法的准确程度。

描述社会网络主要是通过图来描述的，而图又分为有向图、无向图、有权有向图、无权无向图等多种，一种社区发现算法在一种图上表现优秀不代表在所有图上都表现优秀。

除了考查我们改进的 AP 算法之外，AP 算法能否在不同的图上保持较高准确率也是我们要考查的问题之一，即考查算法表现是否能始终如一。尤其是 AP 算法，效果的好坏关键是相似度度量方法的选取以及偏好度的设置，特别是在无权无向图上进行社区发现时，网络中特殊信息较少，这些都会影响相似度和偏好度的设置。这对 AP 算法和 AP 改进算法能否表现良好都是个重大的挑战。为此，本文使用的数据集均为无权无向图，目的就是考查 AP 算法及 AP 改进算法在无权无向图上的表现。

5.2.3 可视化标准

目前的社区发现算法通过可视化软件得到的结果只是几个关系紧密的小社区的连接，如图 5.1，从图中我们只能知道社区内的这一群人的关系较为紧密，但具体到一两个人我们就描述的不够清晰了，所以我们又设计了基于联系关系的社区发现算法。

由于该算法设计简单，展示效果较好，能看到被发现的社群中人与人 的联系紧密程度，而且还能呈现出一定的层次性，并且不需要额外编写可视化软件，所以我们在此也将可视化标准用来评价算法的标准之一。

5.3 实验方案

本节主要对 AP 算法及 AP 改进算法进行实验考查，还有对基于联系关系的社区发现算法的展示效果进行考查。

其中对 AP 算法及 AP 改进算法进行的考查是主要的，我们主要使用基于现实生活数据的 Zachary 空手道俱乐部社会关系网等多组经常被用作社区发现实验的数据进行实验。

5.3.1 基于现实生活数据进行社区发现

首先我们使用基于科学家合作网络数据集，对 GN 算法和 AP 算法分别进行实验，记录实验结果，并通过实验数据的对比来分析 AP 算法的表现，对其性能有个初步的认识，然后再在现实生活数据上进行实验。

我们使用的现实数据集是包括著名的 Zachary 空手道俱乐部社会关系网在内的多个标准现实生活数据集。Zachary 空手道俱乐部社会关系网是经过多年的观察所得，因此事先已经得知具体分为两个社区。图 5.1 就是通过 GN 算法得到的，结果准确。

在随后的实验中，我们在其它的现实生活数据集上进行实验，考查在多组数据相同的情况下 AP 算法及 AP 改进算法的发现效果。注意，实验是在无权无向图上进行的，上面含有的特殊信息比较少，考查在此情况下，算法的效果如何，并在后面进行分析。

5.3.2 基于联系关系的社区发现

通过另一种方式展现社区发现的效果，考查利用社会网络的特点，通过比较简单的算法，实现出社区发现的效果，并通过较好的可视化标准来展示社区内部的关系。其实验数据是我们从 DBLP 数据中选取的一小部分数据。

5.4 实验结果与分析

我们将上述算法分别在上面提到的两个实验方案中进行实验，得到如下实验数据。

5.4.1 时间运行比较

如表 5.1 所示，是 GN 算法、AP 算法及 AP 改进算法三种算法对相同数据集进行实验时，所消耗的时间对比。

表 5.1 GN、AP 和 AP 改进算法的时间消耗

数据集	GN	AP	改进 AP
Zachary	0.02	0.01	0.05
Football teams	0.30	0.21	0.32
Dolphin social network	0.06	0.03	0.15
Computer-generated	0.20	0.15	0.50

从实验数据中可以看到，AP 算法的时间效率是比较高的。但效率高不一定准确率高，在随后的实验中我们将证明这一点。

5.4.2 基于现实生活数据集

我们选取了几个已经知道社区数量的数据集，来分析其算法的准确性。

如表 5.2 所示，是 GN 和 AP 算法对现实生活数据集进行社区发现实验时划分得到的社区数和标准数目的对比。

表 5.2 GN 和 AP 算法得到的社区数

数据集	标准	GN	AP
Zachary	2	2	3
Football teams	6	6	6
Dolphin social network	5	4	5
Computer-generated	4	3	3

从实验结果中可以看到 AP 算法在实验结果中与标准结果尚有一定出入。AP 算法在这些实验数据集上的社区发现表现并不是令人十分满意，其主要问题如前面分析的，无法有效选择偏好度；另外，在无权图上还没有找到更好的相似度度量方法；除此之外，由于我们是在无权无向图上进行实验的，而无向无权图上缺乏足够的相似度信息。对于 AP 算法在无权无向图上表现不是很好的原因中，我们认为 $s(k,k)$ 表示节点 k 被选为聚类中心的可能性，即偏好度(preference)，是影响其效果的重要原因，在 3.2.2 节中提到的 AP 算法改进，就主要是针对这一块的。所以，从上面的实验结果看，AP 算法的表现除了效率之外，并不是在任何时候都表现的那样好，因此，AP 算法的使用还需要慎重，特别是在无权无向图上。

如表 5.1、表 5.3 展示的 AP 改进算法的运行效果和标准结果的对比，虽然改进后的 AP 算法的运行时间有所延长，但准确度得到了很大提高，其实验结果和标准结果完全一致。说明我们之前对 AP 算法不足之处的分析是对的，对 AP 算法的改进是有效的。我们正是考虑到了偏好度空间的特殊情况，最终得到了满意的效果。

表 5.3 改进 AP 算法得到的社区数

数据集	标准	改进 AP
Zachary	2	2
Football teams	6	6
Dolphin social network	5	5

Computer-generated	4	4

5.4.3 基于联系关系

对于本组实验，我们选取的是从 DBLP 数据集中抽取的一部分数据进行的模拟实验。我们采用点度中心度选取的方式，选取“Miron Livny”和“David De Roure”为起始人和目标人，其实验结果为：

Miron Livny——Ian T.Foster——David De Roure
Miron Livny——Luc Moreau——David De Roure
Miron Livny——Suresh Singh——Ian T.Foster——David De Roure
Miron Livny——Paul Avery——Ian T.Foster——David De Roure
Miron Livny——Thomas Fahringer——Luc Moreau——David De Roure

从实验数据可以看到，该算法将与“Miron Livny”和“David De Roure”有过合作经历的研究员按合作紧密程度进行排列，将合著过的并且次数较多的人列在前面，呈现了一定的梯度性，而且可以通过旁边的专家衬托别人的优秀程度，使我们也可以以此来参考这些人的优秀程度，这符合人们正常的思维习惯。特别是当起始人和目标人为同一科研领域的研究专家时，那么能与他们合著的人也一定是同一领域的优秀研究员，即“和权威专家合作的人都是优秀的”。我们将找到的这些人组成了一个社区。

但是我们也看到了此方法的不足，由于此方法是基于“六度分割”假设原理设计的，当起始人和目标人为同一个研究领域的研究员时，发现的人群为同一领域的可能性较大，但当研究员从事不同多个领域研究时，找到的人当然也是‘五花八门’；此外，如果数据规模庞大，将所有研究员两两组队来发现所有的社区，那样工作量是巨大的，我们实验主要是观察其效果，所以只选取了很小一部分进行，所以工作量不是很大，在这方面是不能与 AP 和 GN 算法相比的，而且不同的组队很有可能会出现社区重叠现象，对于这些问题此方法没有提出解决办法，有待进一步完善。

5.5 本章小结

本章对本文提到的 GN 算法、AP 算法、AP 改进算法和联系关系算法分别准备了相应的实验数据，设计了相应的实验方案及评价标准，并得到了相应的实验结果，还对实验结果作了简要的分析。实验结果表明 AP 算法确实有效，但在无权

无向图上表现不是非常理想，使用仍然需要慎重。我们通过对 AP 改进算法实验得知，新算法提高了准确性，实验数据证明改进是有效的。而联系关系算法实验结果说明这种方法简单易实现，有很大的应用前途，但仍有不足，也别是如何很好的解决工作量庞大的问题。

6 总结与展望

本文研究的内容是社区发现技术，特别是对 Affinity Propagation(AP)算法和AP 改进算法进行了研究，并通过实验证明 AP 改进算法的可行性和有效性。此外，本文还提出了一种新的社区发现的方法，通过分析其设计思想来说明其特点。

6.1 论文总结

本文首先分析了社区发现技术的研究背景和研究现状，并介绍了研究目的。提出通过经典的 GN 算法和最新的 AP 算法来研究社区发现技术，通过两种算法的实验对比来研究 AP 算法的发现效果，并对 AP 算法的不足进行了很好的分析，提出了改进的方法。并且在本文中，通过社会网络分析，还设计了一个基于联系关系的新的社区发现算法。

社区发现技术属于社会网络分析研究的一个领域，因此紧接着我们对研究过程所遇到的理论知识进行了比较详细的阐述，包括图论的基本概念、社会网络分析的基本概念。随后，介绍了目前主要的社区发现技术与计算方法。

最后分别对 GN 算法、AP 算法，AP 改进算法、以及基于联系关系的社区发现算法进行了设计与实现，尤其是 AP 改进算法是本文的核心研究内容。本文分别对 AP 改进算法进行了设计，详细介绍了算法的设计思想。同时选择了描述现实社会的真实数据作为实验数据。其中的算法已经完成了编码实现，并通过实验得到了实验结果，说明了算法的有效性与不足。

6.2 进一步的工作

社区发现技术在目前的研究与应用领域有着非常重要的地位，尤其是现在对复杂网络研究的不断深入，在复杂网络中体现的“社区结构特性”，也从侧面推动着社区发现研究不断向前。本文着重对 AP 算法和 AP 改进算法进行了深入的研究和探讨，今后将在以下几个方面进行研究和探索：

1) 将 GN 算法和 AP 算法进行实验对比，进一步研究 AP 算法的特点，以及在不同的社会网络中考查算法有效性的表现；
2) 研究更多无权无向图上的相似度度量方法，将来对 AP 算法是否适合解决无权无向图的社区发现问题作指导；
3) 研究 AP 算法对有权、有向图等含有特殊信息的网络上的应用；

4) 对 AP 算法进行更加进一步的改进，使其效果更好；
5) 基于联系关系的社区发现算法，方法和设计思想简单，体现了社会网络中人际关系的特点，但暴露出来的问题也比较多，研究将此算法思想和目前的社区发现算法结合起来，互相取长补短。

6.3 研究展望

将社会网络分析中的联系关系应用于社区发现之中，是一种全新的研究尝试。由于社会网络数据集的覆盖面广（几乎覆盖了社会的各个方面和领域）、信息资源丰富，所以此技术的应用范围十分广泛。

社会网络分析作为一门独立的学科，社区发现技术又作为社会网络分析中一个重要的研究范畴，在国内外的研究中已经取得了丰硕的成果，并且已经成功运用于社会网络的分析研究之中，所以将其应用于更多的环境中也应该是可行的。本文的研究表明，目前的很多算法在不同的社会网络中表现的效果不一，这方面的研究还有很广阔的余地。

在未来的研究过程中，我们应该致力于对第 6.2 节中所描述的几个方面进行更深入细致的研究。

参考文献

[1] Tantipathananandh, C., T.B. Wolf, and D. Kempe. A framework for community identification in dynamic social networks. in Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007. San Jose, California, USA: ACM.

[2] Flake, G.W., S. Lawrence, and C.L. Giles. Efficient identification of Web communities. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2000. Boston, Massachusetts, United States: ACM Press.

[3] Chakrabarti, D. and C. Faloutsos, Graph mining: Laws, generators, and algorithms. Acm Computing Surveys, 2006. 38(1): p. A1-A69.

[4] Chi, Y., et al. Structural and temporal analysis of the blogosphere through community factorization. in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007. San Jose, California, USA: ACM.

[5] Ino, H., M. Kudo, and A. Nakamura. Partitioning of Web graphs by community topology. in Proceedings of the 14th international conference on World Wide Web. 2005. Chiba, Japan: ACM.

[6] Bennouas, T. and F. Montgolfier. Random web crawls. In Proceedings of the 16th international conference on World Wide Web. 2007. Banff, Alberta, Canada: ACM.

[7] Andersen, R. and K.J. Lang. Communities from seed sets. In Proceedings of the 15th international conference on World Wide Web. 2006. Edinburgh, Scotland: ACM.

[8] West D B. Introduction to Graph Theory [M]. Prentice Hall, Upper Saddle River, 2001.

[9] Duda R O, Hart P E, Stork D G. Pattern Classification [M]. Wiley-Interscience, New York, 2001.

[10] L.da F.Costa, R M.C.Jr. Shape Analysis and Classification: Theory and Practice [M]. CRC Press, Boca Raton, 2001.

[11] Scott J. Social Network Analysis: A Handbook [M]. Sage Publication, London, 2000.

[12] Wu F, Huberman B A. Finding communities in linear time: A physics approach[J]. Euro. Phys. J B, 2003, 38: 331-338.

[13] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physcial Review E, 2004, 69: 026113.

[14] Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts model. Cond-mat/0402349, 2004.

[15] Radicchi F, Castellnao C, Cecconi F, et al. Defining and identifying communities in networks. Cond-mat/0309488, 2004.

[16] Capocci A, Servedio V D P, Caldarelli G, et al. Detecting communities in large networks. Cond-mat/0402499, 2004.

[17] Donetti L, Munoz M A. Detecting Network Communities: a new systematic and efficient algorithm. Cond-mat/0404652, 2004.

[18] Jeong H, Tombor B, Albert R, et al. The large-scale organization of metabolic networks [J]. Nature, 2000, 407: 651-654.

[19] Fell D A, Wagner A. The small world of metabolism [J]. Nature (Biotechnology), 2000, (18): 1121~1122.

[20] Pool l, Kochen M. Contacts and inuence [J]. Social Networks, 1978, (1): 1-48.

[21] Milgram S. The small world problem [J]. Psychology Today, 1967, (2): 60-67.

[22] Pearl, J., Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. 1988, San Francisco, CA: Morgan Kaufmann.

[23] Frey, B.J. and D.J. C. A Revolution: Belief Propagation in Graphs with Cycles. 1998: The MIT Press.

[24] Han, J. and M. Kamber, Data Mining: Concepts and Techniques. 2006: Morgan Kaufmann.

[25] MacQueen, J., Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967. 1(281-297): p. 14.

[26] Kernighan, B.W. and S. Lin, An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal, 1970. 49(2): p. 291-307.

[27] Fiedler, M., A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. Czechoslovak Mathematical Journal, 1973. 23(298): p. 619-633.

[28] Pothen, A., H.D. Simon, and K.P. Liou, Partitioning sparse matrices with eigenvectors of graphs. Siam Journal On Matrix Analysis And Applications, 1990. 11(3): p. 430-452.

[29] Scott, J., Social Network Analysis: A Handbook. 2000: Sage Publications.

[30] Girvan, M. and M.E. Newman, Community structure in social and biological networks. Proceedings Of The National Academy Of Sciences Of The United States Of America, 2002. 99(12): p. 7821-7826.

[31] Newman, M.E. and M. Girvan, Finding and evaluating community structure in networks. Physical Review E, 2004.

[32] Newman, M.E., Fast algorithm for detecting community structure in networks. Physical Review E, 2004.

[33] Newman, M.E., Modularity and community structure in networks. Proceedings Of The National Academy Of Sciences Of The United States Of America, 2006. 103(23): p. 8577 - 8582.

[34] Palla, G., et al., Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005. 435(7043): p. 814-818.

[35] Frey, B.J. and D. Dueck, Clustering by passing messages between data points. SCIENCE, 2007. 315(5814): p. 972-976.

[36] Kschischang, F.R., et al., Factor graphs and the sum-product algorithm Factor graphs and the sum-product algorithm. Information Theory, IEEE Transactions on.

[37] Dueck, D. and B.J. Frey. Non-metric affinity propagation for unsupervised image categorization. In ICCV 2007. 2007.

[38] Frey, B.J. and D. Dueck. Mixture Modeling by Affinity Propagation. In NIPS 2005. 2005.

[39] 邓庆山. 聚类分析在基因表达数据上的应用研究. 计算机工程与应用.2005,41 (35).

[40] 乐阳,龚健雅. Dijkstra 最短路径算法的一种高效率实现. 武汉测绘科技大学学报,1999.

[41] 陈萧枫,蔡秀云,唐德强.最短路径算法分析及其在公交查询的应用[J].工程图学学报,2001, (3) .

[42] Girvan, M. and M.E. Newman, Community structure in social and biological networks. Proceedings Of The National Academy Of Sciences Of The United States Of America, 2002. 99(12): p.7821 - 7826.

[43] Newman, M.E. and M. Girvan, Finding and evaluating community structure in networks. Physical Review E, 2004.

作者简历

薄辉，男，生于1982年11月19日。2001年9月至2005年6月就读于河北省燕山大学计算机科学与技术专业，学历本科，并获得学士学位。2006年9月至今就读于北京交通大学计算机与信息技术学院，专业为计算机软件与理论，学历为硕士研究生。