大规模社交网络的社区发现算法研究

朱杰

2018 年 5 月

中图分类号：TP311

UDC分类号：004

大规模社交网络的社区发现算法研究

作 者 姓 名　朱杰

学 院 名 称　计算机学院

指 导 教 师　张欣讲师

答辩委员会主席　** 教授

申 请 学 位　工程硕士

学 科 专 业　软件工程

学位授予单位　北京理工大学

论文答辩日期　2018 年 5 月

An Algorithm for Large-scale Social Network Community Detection

Candidate Name:	Zhu Jie
School or Department:	School of Computer Science & Technology
Faculty Mentor:	Lect. Zhang Xin
Chair, Thesis Committee:	Prof. **
Degree Applied:	Master of Engineering
Major:	Software Engineering
Degree by:	Beijing Institute of Technology
The Date of Defence:	May，2018

大规模社交网络的社区发现算法研究

北京理工大学

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位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理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

作者签名：_____　　签字日期：_____

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：① 学校有权保管、并向有关部门送交学位论文的原件与复印件；② 学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③ 学校可允许学位论文被查阅或借阅；④ 学校可以学术交流为目的，复制赠送和交换学位论文；⑤ 学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

作者签名：_____　　导师签名：_____

签字日期：_____　　签字日期：_____

摘要

本文……。（摘要是一篇具有独立性和完整性的短文，应概括而扼要地反映出本论文的主要内容。包括研究目的、研究方法、研究结果和结论等，特别要突出研究结果和结论。中文摘要力求语言精炼准确，硕士学位论文摘要建议 500~800 字，博士学位论文建议 1000~1200 字。摘要中不可出现参考文献、图、表、化学结构式、非公知公用的符号和术语。英文摘要与中文摘要的内容应一致。）

关键词： 形状记忆；聚氨酯；织物；合成；应用（一般选 3 ～ 8 个单词或专业术语，且中英文关键词必须对应。）

Abstract

In order to exploit …….

Key Words: shape memory properties; polyurethane; textile; synthesis; application

目录

插图

表格

第 1 章　绪论

1.1　研究背景和意义

近年来，随着科技的发展和网络的不断普及，在线社交网络如今已经成为互联网时代最为基础的一部分。诸如微信、微博、Facebook、Twitter 和 GitHub 等等国内外的社交类平台软件的出现，使得人们可以更加有高效的沟通交流。在如今的移动互联网时代下，人们的社交重心由线下更多的转到了线上。线上的社交也确实带来了很多的便捷。人们不再有地域的限制，可以轻松的与亲朋好友时刻保持联系；人们通过社交平台可以迅速的认识了解一个人并与之成为朋友；人们可以扮演起在日常生活中无法扮演的角色，任何人都可以成为信息的分享者和传播者。

如今这些成熟的社交软件每天都会产生海量的用户数据。在这个大数据的时代，这些看似杂乱无章、毫无交集的数据中，其实蕴含了丰富的信息等待着人们去挖掘与分析。在这样的背景下，如果将社交平台中所有用户抽象成点，而用户与用户之间的关联抽象成边，这就抽象出了一张网络关系图，那么对于在互联网这个虚拟世界中形成的这样一张巨大而又复杂的社交网络的研究与分析就显得意义非凡。

正所谓物以类聚，人以群分。社交网络中亦是如此。网络图内部连接比较紧密的节点子集合对应的子图称之为社区。网络图中包含一个个社区的现象称之为社区结构。社区结构是网络的一个普遍特征。而给定一个网络图，找出其社区结构的过程就叫做社区发现（community detection）。挖掘社交网络中的社区在人物分析、商业个性化推荐和舆情控制等领域有着很关键的作用。单单以商业个性化推荐而言，为了获得更大的用户群体以获取更多的流量关注，或者刺激促进用户更多的消费，在线求职招聘类平台要为用户推荐适合用户需求的职位，在线购物消费类平台要为用户推荐符合用户需求的商品，在线社交网络平台要为用户推荐和用户兴趣相投或者相关的好友，在线新闻媒体平台要为用户推荐符合用户口味的相关讯息。在这样的背景下，一个优秀的个性化推荐系统就显得尤为重要。在推荐系统中，免不了要对用户群体进行分类，而社区发现所形成的社区在此就有着先天的优势。直白的说，对社交网络的社区发现，无非也就是对社交网络中的用户分类（或者说聚类）。在同一个社区内的用户群体，往往有着相似的兴趣爱好。因此如果对整个用户群体进行社区发现，那么比如说在为用户进行商品个性化推荐时，可以重点关注与该用户在同一个社区的其他用户

购买过而他未购买过的商品，这样用户的接受率也会高很多。当然，推荐系统是一个复杂的 AI 系统，利用了远不止社区发现这一种手段。

对社区发现算法的研究其实远不止社交网络这一领域，社区发现其实是对复杂网络的一种分析手段。除了社交网络，还有科学文献引用网络、生物学分子结构分析等等领域也都使用着社区发现算法。截至当前，多年来的众多学者的贡献使得对社区发现算法的研究已经是比较成熟了，但是依然存在着问题。现如今的社交网络数据往往是海量的，对如此大规模数据的社交网络进行社区发现时，过往的一些经典算法显然已经无法胜任。因此，在这样的背景下，提出一种适应于当前大规模社交网络的社区发现算法就显得尤为迫切了。

1.2 国内外研究现状

本文这里将现有的社区发现算法分为三大类：基于链接的方法，基于内容的方法和融合链接和内容的方法。基于链接的方法也就是基于网络拓扑结构的方法。它将社交网络看做一张网络图，用户为节点，关系为边。基于内容的方法主要是基于社交网络中用户的个人信息和发表过的内容来进行社区发现。基于内容的方法也可被称为基于主题的方法，或者基于节点相似度的方法。而融合的方法则同时关注网络拓扑结构和用户属性，以此来获取更高质量的社区。

人们总是习惯于和自己相似的人结交关系。不论是友谊还是工作关系，任何关系网络中人们的缔结方式均有此趋势。因此，在社交网络中两个用户之间所谓的链接关系被认为是一个可以证明彼此之间是有着某些共同点的证据。这也有利于发现社区。这个发现最早是被 McPherson 等人[1] 提出的，文中提到了同质性原则“相似性产生联系”。这也被认为是大部分社区定义的最早的参考准则。

在基于链接的社区发现算法中，社交网络是由一张图为模型，节点代表社区成员，边代表成员之间的关系或者交互。这里社区所需的凝聚力属性就是成员之间的链接。在社区之中链接是较为密集的，而在社区之间链接相对而言较为稀疏。分别将原始图结构中的组件和派系当做是已知的社区[2]。然而，更多的更有意义的社区可以通过基于图划分（聚类）的方法来检测得到，其目标就是尽可能减少社区之间边的数量。这样一来，同一个社区中的节点之间就可以有更多的内连接，而与别的社区中节点的外连接就可以减少。大部分的方法都是基于二分迭代：不断地将一个社团划分为两个社团。然而在复杂网络中社区的数量显然是无法预先得知的。在这个层面上，

Girvan-Newman 的算法是最为广泛使用的基于链接的社区发现算法[3]。GN 算法的基本思想是删除那些社区之间的连接，这样剩下的每个连通部分就是一个社区。作者巧妙地借助了最短路径这一思想。GN 算法中定义一条边的边介数（betweeness）为网络中所有节点之间的最短路径中通过这条边的数量，而边介数高的边要比介数低的边更可能是社区之间的边。其实，这也比较好理解，因为两个社区中的节点之间的最短路径都要经过那些社区之间的边，所以它们的边介数会很高。在社区发现中几乎不可能预先知道社区的数目。于是必须有一种度量的方法，可以在计算过程中来衡量每个结果是不是相对最佳的结果。这同样也是算法好坏的评价指标。在 GN 算法中使用了模块度（modularity）这一概念。模块度的大小定义为社区内部的总边数和网络中总边数的比例减去一个期望值，该期望值是将网络设定为随机网络时同样的社区分配形成的社区内部的总边数和网络中总边数的比例的大小，模块度一般记为 Q。在每次划分的时候计算 Q 值，当 Q 取最大值时则是此网络较为理想的划分。Q 取值，越大越好，实际中一般 Q 最高点在 0.3 至 0.7。有时候，当不能或者不容易获取全部网络的数据时，可以用局部社区中的局部模块度来检测社区的合理性。局部模块度比全局模块度快很多，中小网络效果会比全局的差些，但是中等或大规模的网络中，局部模块度效果可能好要比全局的更好。其他的一些分图算法还包括：最大流最小割理论[4]、谱二分的方法[5]、Kernighan-Lin 划分算法[6] 和最小电导率分割算法[7] 等等。

基于链接的社区发现算法其实也可以被看作是一种数据挖掘或者说机器学习聚类算法，相当于无监督的用户分类。因此这其中可以用到的无监督学习的相关技术包括：k-means 算法、混合模型和层次聚类等等。

尽管基于链接的技术更加直观且基于社会学的同质原则，但也有两个原因导致它们在识别基于相似兴趣的用户社区方面存在缺陷。首先，许多社交关系不是基于用户的兴趣相似性，而是基于其他因素，如朋友和亲属关系，并不一定反映用户间的兴趣相似性。其次，许多有着相似兴趣的用户彼此之间可能并没有互相关注，以至于在网络之中似乎是没有关联的[8]。随着在线社交网络功能的不断增加，网络上除了用户之间的链接之外，还有许多用户自己提交的内容（称为社交内容）可用。用户可以维护个人资料页面，撰写评论，分享文章，标记照片和视频以及发布他们的状态更新。因此，研究人员已经探索了利用社交内容的主题相似性来检测社区的可能性。他们提出了基于内容或主题的社区检测方法，这样一来，不论社交网络结构如何，都可以检测到志同道合的社区用户[9]。

大多数基于内容的社区发现工作都侧重于检测社区文本内容的相似模型。比如，Abdelbary 等人提出的算法[10] 利用了高斯受限玻尔兹曼机（GRBM）来识别主题社区。尹志军等人[11] 将社区发现与主题建模结合在一个统一的生成模型中，以检测在结构关系和潜在主题方面相互一致的用户群体。在他们的框架中，一个社区可以围绕多个主题形成，一个主题也可以在多个社区之间共享。Sachan 等人[12] 提出了概率方案，将用户的帖子、社交关系和交互类型结合起来发现 Twitter 中的潜在用户社区。在他们的工作中，他们考虑了三种类型的互动：传统推文、回复推文和转载推文。其他学者还提出了隐含狄利克雷分布模型（LDA）的变体来识别主题社区，例如作者-主题模型（Author-Topic model)[13] 和社区-用户-主题模型（Community-User-Topic model)[14]。

另一个流派的工作将基于内容的社区发现问题建模为图聚类问题。这些方法都基于相似性度量标准，该度量标准能够根据用户都感兴趣的主题计算用户的相似度，并基于聚类算法来提取具有相似兴趣的用户群体（潜在社区）。例如，刘洪涛等人[15] 提出了一种基于用户间主题距离（topic-distance）的聚类算法来检测社交标签网络中基于内容的社区。在这项工作中，LDA 被用来提取标签中隐藏的主题。彭敦陆等人[16] 提出了一个层次聚类算法来检测推文中的潜在社区。他们在新浪微博中使用了预定义的类别，并根据用户在每个类别中的兴趣程度计算了用户的配对相似度。

像基于链接的方法一样，基于内容的社区发现方法也可以转化为数据的聚类，这里的一个社区只是一组节点的集合。代表用户的节点与同一社区内的节点相似度较高，而与社区外的节点相似度较低。从这个意义上说，亲密关系确实是社区所需的凝聚力属性。

基于内容的方法其实是为常规文本设计的，但是诸如 Twitter 或微博这类社交网络多是简短、混杂和非正式的社交内容。在这种情况下，社交内容本身并不是提取真实社区的可靠信息[17]。通过社交结构（即链接）丰富社交内容有助于我们找到更有意义的社区。研究人员们已经提出了几种方法将链接和内容信息结合起来用于社区发现。正如参考文献[18,19] 中所述，它们可以拥有更好的性能。大多数这类方法是通过共享隐含变量这一手段来为社区成员制定链接和内容的综合生成模型。

社区发现算法的常见方法是将网络划分为不相交的社区成员。这种方法忽略了个体可能属于两个或更多社区的可能性。但是，许多真实的社交网络都存在着社区的重叠[20]。例如，一个人可以属于多个社交群体，例如家庭群体和朋友群体。越来越多的研究人员开始探索允许社区重叠的新方法，即重叠社区（overlapping communities）。

重叠社区引入了另一个变量，即不同社区中用户的成员身份，称为 cover。由于与标准社区相比，重叠社区有大量可能的 cover，因此检测此类社区代价就很高。

一些重叠的社区发现算法利用网络中用户的结构信息将网络的用户分成不同的社区。这类方法的主导算法是基于集团渗透理论（clique percolation theory）[21]。然而，LFM 和 OCG 方法是基于对用户出入度适应函数的局部优化[22,23]。此外，一些模糊社区发现算法会计算每个节点属于每个社区的可能性，如 SSDE 和 IBFO[24,25]。几乎所有的算法都需要先验信息来检测重叠的社区。例如，LFM 需要一个参数来控制社区的大小。不过也有一些基于相似性的方法将社区看作分布在整个用户空间的隐含变量，如参考文献[26]。

Erosheva 等人[27] 介绍了 Link-LDA，一种重叠的社区发现方法。它可以同时根据摘要（内容）和参考文献（链接）对科学类论文进行分类。在它们的生成模型中，论文被假定为摘要和参考文献的一对模型，每个部分都用 LDA 抽取特征。在摘要和参考文献中相似性都很高的文章倾向于有着相同的主题。与 Link-LDA 相反，Nallapti 等人[28] 没有将参考文献视作待处理的单词，并提出需要明确引用文本和参考文献之间的主题关系。他们提出了 Pairwise-Link-LDA 来模拟文档对之间的链接存在，并通过使用这些附加信息获得了更好的主题质量。其他利用 LDA 融合链接和内容的方法可以参考文献[29,30]。除了相似度生成模型外，还有其他一些方法将链接和内容信息结合起来用于社区发现，如谱聚类中利用矩阵分解和核聚变的方法[31,32]。

1.3 论文主要工作

通过查阅大量关于社交网络、社区发现、聚类等方面的文献资料，深入理解社交网络及重叠社区特性的基础上，认真研究社区发现相关算法，本文完成了如下工作：

（1）研究问题具体化。本文将社交网络转化为无向带权网络模型，针对无向网络找寻社区划分方案。在允许节点属于多个社区的情况下，使得划分后的社区内节点关系紧密，社区间节点关系疏远。

（2）CDMMLPA 算法的设计。标签传播算法主要分为标签分配和传播 2 个阶段。CDMMLPA 算法中并未为每个节点分配标签，而是以小社区为单位，为每个小社区分配唯一的标签。采用模块度最大化的方法进行标签传播，由于模块度最大化是一个单向操作，因此能获得高质量的社区。CDMMLPA 算法中充分考虑了社区结构的属性，以确保得到稳定的社区。

1.4 论文组织结构

本论文主要对重叠社区的挖掘算法进行研究，并设计了相应的优化算法。本文主要包括四大章节，其主要的结构组织如下：

第一章为绪论。主要介绍了课题的背景、意义、国内外现状以及本课题的主要研究内容。其中，重点介绍了各类社区发现算法的国内外研究现状。

第二章为相关工作。首先介绍了复杂网络的定义；然后介绍了复杂网络中社交网络的特性；接下来给出了社区发现的定义，从社区结构定义和社交网络网络模型着手；最后介绍了社区结构的评价指标。

第三章为基于模块度最大化的标签传播社区发现算法。主要介绍本文设计的一种基于模块度最大化的标签传播社区发现算法（CDMMLPA）。本章首先是针对算法中的一些特殊名词给出了相对应的定义；然后详细解释了算法实现过程；最后给出了算法的伪代码。

第四章为社区发现算法相关实验与评估。首先介绍了实验用的数据集；然后是对比实验部分，主要和经典标签传播算法 LPA 和改进的标签传播算法 LPAm+ 的比较；比较维度主要是模块度、强社区数目和运行时间；最后是实验总结。

第 2 章　相关工作

2.1　复杂网络

随着近几年关于复杂网络（Complex network）理论及其应用研究的不断深入，已有大量关于复杂网络的文章发表在 Science，RL，NAS 等国际一流的刊物上，侧面反映了复杂网络已经成为物理界的一个新兴的研究热点。人们开始尝试应用这种新的理论工具来研究现实世界中的各种大型复杂系统，其中人们关注的热点问题是复杂系统的结构以及系统的结构与功能之间的关系。

在自然界中存在的大量复杂系统都可以通过形形色色的网络加以描述。一个典型的网络是由许多节点与节点之间的连边组成，其中节点用来代表真实系统中不同的个体，而边则用来表示个体间的关系，往往是两个节点之间具有某种特定的关系则连一条边，反之则不连边，有边相连的两个节点在网络中被看作是相邻的。例如，神经系统可以看作大量神经细胞通过神经纤维相互连接形成的网络；计算机网络可以看作是自主工作的计算机通过通信介质如光缆、双绞线、同轴电缆等相互连接形成的网络。类似的还有电力网络、社会关系网络、交通网络、调度网络等等。

复杂网络的研究由于其学科交叉性和复杂性的特点，涉及了众多学科的知识和理论基础，尤其是系统科学、统计物理、数学、计算机与信息科学等，常用的分析方法和工具包括图论、组合数学、矩阵理论、概率论、随机过程、优化理论和遗传算法等。复杂网络的主要研究方法都是基于图论的理论和方法开展的，并已经取得了可喜的成果。但近几年，统计物理的许多概念和方法也已成功地用于复杂网络的建模和计算，如统计力学、自组织理论、临界和相变理论、渗流理论等，如网络结构熵的概念，并用它来定量地度量复杂网络的“序”。复杂网络模型在很多科学领域都得到广泛的应用。

2.2　社交网络

社交网络（Social network）就是复杂网络中的一种，是由许多节点以及节点间关系构成的一个网络结构。节点通常是指个人或组织。社交网络代表各种社会关系，经由这些社会关系，把从偶然相识的泛泛之交到紧密结合的家人关系的各种人们或组织串连起来。因此，社交网络也被称为社会关系网络。社交网络依赖于一种或多种关系而形成，如价值观、理想、观念、兴趣爱好、友谊、血缘关系、共同厌恶的事物、冲

突或贸易。由此产生的网络结构往往是非常复杂的。

随着网络时代的来临，社交软件成为人们日常生活中必不可少的模块，社交网络也随之被作为当下流行的研究课题。社交网络是指人与人之间以某种关系建立的的社会网络结构，这种关系可以是兴趣爱好，朋友熟人，经贸交易等。将社会行动者视为节点，社会关系视为连线的边，从而构造出社交网络的拓扑模型。社交网络本着“节约社交成本，高效获取信息”的目的，依靠特定的功能诉求，将相应的用户人群集结在同一平台，从而实现社会关系网络化的延伸，跨越了时间地域等的限制。相对于传统网络，社交网络包含以下特性：

（1）以人为基点构建的关系型社会网络。社交网络以人为中心，着重强调人与人之间的关系。如兴趣类社交平台百度贴吧，贴吧通过社交网络的优势，使得拥有相同兴趣爱好的人聚集在一起，以兴趣建立用户之间的联系，构造庞大的主题社交平台。由于社交网络中人为基点的特性，现存的社交网络软件往往是以用户为中心来进行产品设计和组织管理。

（2）虚拟化社会关系，现实社交关系的延伸。社交网络是人们真实生活的虚拟映射，网络上的社交行为也是人们日常社交的一部分。社交网络以人为主题，每个用户有其对应的 id 账号，并围绕该账号展开一系列工作和生活相关的社交行为，形成相应的朋友圈，以此拓展现实社交行为。如生活类的社交软件微信，微信将现实生活中的人际交往以互联网形式呈现，每个用户通过自己的 id 账号管理着自己的社交圈，丰富了用户的社会交际。社交网络跨越了时间地域等的限制，为用户提供了便捷。

社交网络分析是用来查看节点、链接之间的社会关系的分析方式。节点是网络中的个人参与者，链接则是参与者之间的关系。节点之间可以有很多种链接。一些学术研究已经显示，社交网络在很多层面运作，从家庭到国家层面都有，并扮演着关键作用，决定问题如何得到解决，组织如何运行，并在某种程度上决定个人能否成功实现目标。

对于社交网络分析，社区发现是其中一个十分重要的问题研究方向，社区发现也可以被称为社区检测。对于所有社交网络中的节点，判断它隶属于哪一个社区，从而把有相似属性的节点划分到同一个集合，一个大的网格节点集合可以被划为多个不同大小包含不同节点的子集，这就是社区发现。独立的分析每个社区结构的特点，可以得到社区之间的不同之处，这样我们可以迅速简便的获得自己想要的信息。

2.3 社区发现

社区发现（community detection，也可以译作社区检测）是一个复杂而有意义的过程，它对研究复杂网络的特性具有重要作用。给定一个网络图，找出其社区结构的过程就叫做社区发现。近几年，发现及分析复杂网络中的社区结构得到了许多学者的关注，同时也出现了很多的社区发现算法。

2.3.1 社区结构定义

（1）一般社区结构定义

目前对网络社区的定义还没有一个统一的标准，人们从不同的角度给出了不同的社区结构的定义。典型的包括：基于子图的局部定义和基于节点相似度的定义。

基于子图的局部定义：社区结构可以被看成网络拓扑结构中具有高内聚特点的若干节点集合，这些节点集合往往是某种具有独立功能或者性质的相对独立组件的抽象。因此，可以根据网络局部拓扑结构特点来定义社区结构。当前，被各领域学者广泛接受的一个描述定义是基于子图的局部定义。即社区结构是复杂网络节点集合的若干子集，每个子集之间的节点之间连接非常紧密，不同子集的节点之间连接比较稀疏，如图2.1所示。图中的网络由 20 个节点组成，这 20 个节点被划分成了三个社区结构，分别对应着图中的三个虚线内部的结构。

基于节点相似度的定义：从物理意义上讲，社区往往代表了复杂系统或者复杂网络中的具有相似或者相同功能的元素集合，这些元素相互协作或者相互作用，共同完成整个系统中某些相对独立的功能或者组织结构。据此，可以基于节点的相似度来定义社区，该定义假定社区内部节点都是相似的，社区间的节点相似性低，采用某种指标来衡量网络节点间的相似性，根据节点之间的相似性来定义社区结构。总体而言，从本质内涵来看，已有的社区结构的定义都是一致的，是由网络中所有个体组成的集合的子集，该集合中的个体基于某种属性连接紧密并和子集外的个体连接稀疏。但是紧密和稀疏并没有一个可以定量分析的标准，这些定义就没有多少实用的价值。

（2）重叠社区结构定义

在真实世界的社交网络中，社区结构呈现复杂多样的特点，大部分社区结构是重叠的，这就是说网络的节点集合中存在一些同时属于多个社区的节点，即重叠节点。比如，在社交网络中一个个体可以同时属于多个社会团体，各个组织之间有一些共有

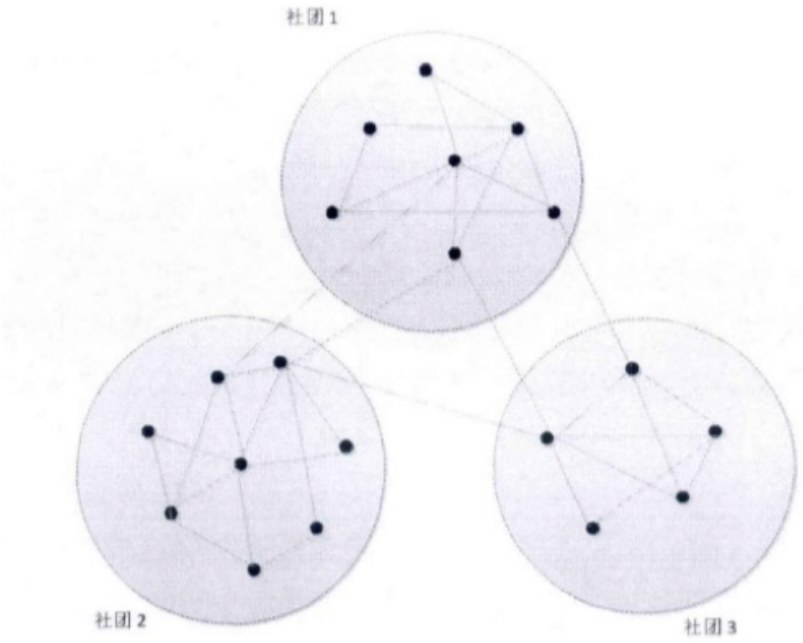

图 2.1　一个具有社区结构的网络示意图

的个体。在各种类型的网络中，重叠节点一般十分重要。所以，网络中的重叠社区发现获得了越来越多人的关注。

通常，重叠社区结构大致被分为两种类型：离散重叠社区和模糊重叠社区。对于前者，我们只要判断一个节点属不属于一个社区，也就是说节点要么属于一个社区，要么不属于这个社区。相反，模糊重叠需要计算节点对于不同的社区的隶属度，对于某个社区的隶属度有高有低。

2.3.2　社区网络模型描述

网络中社区结构表示的是网络中节点集合的子集。一般情况下，一个复杂的网络可以这样表示：由顶点集 V 和边集 E 组成的图 $G = (V, E)$。节点个数表示为 $n = |V|$，边数表示为 $m = |E|$。如果任意两个节点对 (i, j) 与 (j, i) 表示的是同一条边，该图被称为无向图，否则，该被称为有向图。如果我们给图中的每一条边都设置一个代表关系强弱程度的数值，我们把这种图定义为有权图；否则，该图被称为无权图。显然，我们也可以把无权图看成图中每条边权重值都相同的有权图，比如权值都为 1。在无向图中的定义中，节点 i 的度指的是以 i 为顶点的边的数目，记为，是所有含有该节点的边的数量的总和。在有向图的定义中，节点的度分为两种类型，入度和出度。以

该节点为终点的边的数量为该节点的出入度，以该节点为起点的边的数量为该节点的出度。在无向图中无出入度之分。此外，我们还可以用邻接矩阵或者邻接表来表示网络的真实拓扑结构，邻接矩阵如果是对称矩阵那么表示的是无向图，如果是非对称的矩阵表示就是有向图。

2.4 社区结构评价指标

迄今为止，出现了各种各样的社区发现算法，如何评价不同的的发现算法的好坏是一个非常重要的问题。为此，学者们提出了多种社区结构评价指标用来评价网络社区划分质量，其中比较有代表性的有模块度、NMI 等。下面详细介绍这些指标。

2.4.1 模块度

模块度是目前学者们最常用和经典的网络社区结构评价指标，它最初是被 Newman 等人于 2004 年提出来的[3]。其通过比较现有网络和基准网络在相同社区划分下的连接密度差来衡量网络社区的优劣，其中基准网络是由原网络具有相同度序列的随机网络。模块度计算公式如下：

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{2.1}$$

其中，A 表示网络中的邻接矩阵，m 表示网络中边的总数，k_i 和 k_j 表示节点 i 和 j 的度数，c_i 和 c_j 表示节点 i 和 j 所属的社区。如果 $i = j, \delta(c_i, c_j) = 1$，反之 $\delta(c_i, c_j) = 0$

.....

2.4.2 NMI

随着在线社交网络的发展，人们发现在线社交网络的很多数据中存在着暗示各个节点的社区属性信息。例如，在人人网的学校信息便揭示了网络节点中属于同一学校的社区结构，Facebook 中的兴趣信息同样表征了具有相同兴趣的虚拟用户群体。这些数据在为社区发现问题提供了丰富的信息的同时，也在一定程度上为虚拟社区结构优劣的评判提供了标准答案。针对这种预先拥有一定虚拟社区结构信息的情况下，Leon Danon 等人【34】提出了 Normalized Mutual Information（NMI）利用信息化熵来

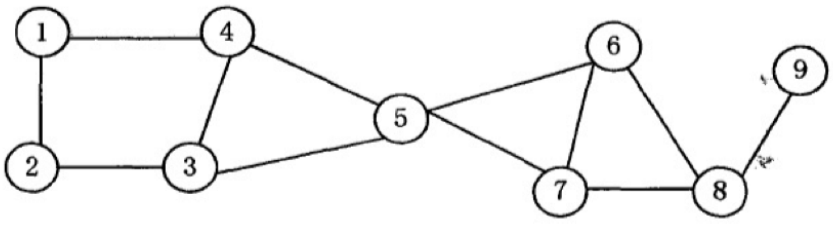

图 2.2　网络示例

衡量算法划分的社区结构和预先已知的社区结构之间的差异。NMI 是基于混合矩阵（Confusion Matrix）N 来计算的数字指标。NMI 公式如下：

$$NMI = \frac{-2\sum_{i,j} N_{ij} ln\frac{N_{ij}}{N_i N_j}}{\sum_i N_i ln\frac{N_i}{n} + \sum_j N_j ln\frac{N_j}{n}} \tag{2.2}$$

使用该数字指标，可以衡量划分出来的社区结构与已知的网络社区结构的差异程度值，该值越大，则表明获得的社区结构划分越好，当该值达到最大化值 1 时，说明算法发现的社区结构与已知社区结构完全已知，效果最好。

下面以图2.2为例来说明计算 NMI 的过程。假设已知的最佳社区结构划分为集合 1，2，3，4 和 5，6，7，8，相应的社区划分向量表示为 a = (1,1,1,1,2,3,3,3,3)，再假设某算法获得的社区划分结构可以用向量表示为 b = (3,3,3,3,2,1,1,1,1) 来表示。根据已知的社区划分向量，可以构造混合矩阵 N:

$$N = \begin{bmatrix} 0 & 0 & 4 \\ 0 & 1 & 0 \\ 4 & 0 & 0 \end{bmatrix} \tag{2.3}$$

根据上式计算可知，该划分的 NMI 值为 1。

结论

本文采用……。（结论作为学位论文正文的最后部分单独排写，但不加章号。结论是对整个论文主要结果的总结。在结论中应明确指出本研究的创新点，对其应用前景和社会、经济价值等加以预测和评价，并指出今后进一步在本研究方向进行研究工作的展望与设想。结论部分的撰写应简明扼要，突出创新性。）

参考文献

[1] Mcpherson M, Smithlovin L, Cook J M. Birds of a feather: Homophily in social networks[J]. Annual Review of Sociology, 2001, 27(1): 415–444.

[2] Fortunato S. Community detection in graphs[J]. Physics Reports: A Review Section of Physics Letters (Section C), 2010(3/5): 75–174.

[3] M G, J N M. Community structure in social and biological networks.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2002(12): 7821–7826.

[4] Jr L R F, Fulkerson D R. Maximal flow through a network[M]. [S.l.]: Birkhäuser Boston, 2009: 243–248.

[5] Pothen A, Simon H D, Liou K P. Partitioning sparse matrices with eigenvectors of graphs[J]. Siam J.matrix Anal.appl, 1990, 11(3): 430–452.

[6] Kernigan R. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal, 1970, 49.

[7] Leskovec J. Graphs over time: Densification laws, shrinking diameters, explanations and realistic generators[J]. Kdd, 2005: 177–187.

[8] Deng Q, Li Z, Zhang X, et al. Interaction-based social relationship type identification in microblog [M]. [S.l.]: Springer International Publishing, 2013: 151–164.

[9] Natarajan N, Sen P, Chaoji V. Community detection in content-sharing social networks[C]. Ieee/acm International Conference on Advances in Social Networks Analysis and Mining. [S.l.: s.n.], 2013: 82–89.

[10] Abdelbary H A, Elkorany A M, Bahgat R. Utilizing deep learning for content-based community detection[C]. Science and Information Conference. [S.l.: s.n.], 2014: 777–784.

[11] Yin Z, Cao L, Gu Q, et al. Latent community topic analysis:integration of community discovery with topic modeling[J]. Acm Transactions on Intelligent Systems Technology, 2012, 3(4): 1–21.

[12] Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks[C]. International Conference on World Wide Web. [S.l.: s.n.], 2012: 331–340.

[13] RosenZvi, Michal, Griffiths, et al. The author-topic model for authors and documents[M]. [S.l.: s.n.], 2012: 487–494.

[14] Zhou D, Manavoglu E, Li J, et al. Probabilistic models for discovering e-communities[J]. Proc. 15th Int. Conf. on World Wide Web (WWW'06), 2006: 173–182.

[15] Liu H, Chen H, Lin M, et al. Community detection based on topic distance in social tagging networks [J]. Telkomnika Indonesian Journal of Electrical Engineering, 2014, 12(5).

[16] Peng D, Lei X, Huang T. Dich: A framework for discovering implicit communities hidden in tweets [J]. World Wide Web-internet Web Information Systems, 2015, 18(4): 795–818.

[17] Yang T, Jin R, Chi Y, et al. Combining link and content for community detection: a discriminative approach[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.: s.n.], 2009: 927–936.

[18] Cohn D, Hofmann T. The missing link: a probabilistic model of document content and hypertext connectivity[C]. International Conference on Neural Information Processing Systems. [S.l.: s.n.], 2001: 409–415.

[19] Getoor L, Friedman N, Koller D, et al. Learning probabilistic models of link structure[J]. Journal of Machine Learning Research, 2003, 3(4): 679–707.

[20] Xie J, Kelley S, Szymanski B K. Overlapping community detection in networks[M]. [S.l.: s.n.], 2013.

[21] Palla G, Dere Nyi I, Farkas I S, et al. Uncovering the overlapping community structure[M]. [S.l.: s.n.], 2005.

[22] Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2012, 11(3): 19–44.

[23] Becker E, Robisson B, Chapple C E, et al. Multifunctional proteins revealed by overlapping clustering in protein interaction network[J]. Bioinformatics, 2012, 28(1): 84–90.

[24] Magdon-Ismail M, Purnell J. Ssde-cluster: Fast overlapping clustering of networks using sampled spectral distance embedding and gmms[C]. IEEE Third International Conference on Privacy, Security, Risk and Trust. [S.l.: s.n.], 2010: 756–759.

[25] Lei X, Wu S, Ge L, et al. Clustering and overlapping modules detection in ppi network based on ibfo [J]. Proteomics, 2013, 13(2): 278–290.

[26] Ren W, Yan G, Liao X. A simple probabilistic algorithm for detecting community structure in social networks[J]. Physics, 2007: 36–40.

[27] Erosheva E, Fienberg S, Lafferty J. Mixed-membership models of scientific publications[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 Suppl 1 (1): 5220.

[28] Nallapati R M, Ahmed A, Xing E P, et al. Joint latent topic models for text and citations[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August. [S.l.: s.n.], 2008: 542–550.

[29] Dietz L, Bickel S, Scheffer T. Unsupervised prediction of citation influences[C]. International Conference on Machine Learning. [S.l.: s.n.], 2007: 233–240.

[30] Gruber A, Rosenzvi M, Weiss Y. Latent topic models for hypertext[J]. Uai, 2008.

[31] Zhu S, Yu K, Chi Y, et al. Combining content and link for classification[M]. [S.l.: s.n.], 2007: 487–494.

[32] Yu S, Moor B D, Moreau Y. Clustering by heterogenous data fusion: framework and applications[J]. Dec-2008, 2008.

致谢

光阴荏苒，岁月如梭，两年的研究生生活转瞬即逝。值此毕业论文即将完稿之际，我对帮助过我的老师、同学以及亲友表达由衷的感谢，并对本硕共培养了我 6 年的母校北京理工大学致以诚挚的敬意。

首先感谢我的两位导师：讲师张欣和副教授金福生。张老师是我名义上的导师，但是实际上两年研究生生涯我接触更多的是金老师。在参与金老师负责的实验室所承接的项目的工作中，我积累到了宝贵的项目经验。本论文的工作也包含了金老师悉心的监督和指导，在论文的撰写上提出了很多宝贵意见。感谢两位导师的帮助和指导。

感谢陪伴我两年的舍友张俊逸、谢辰和刘哲湘。两年的朝夕相处与你们建立了深厚的情谊，不论是学习还是生活中都少不了你们的帮助和支持。看着你们如今都找到满意的工作，有了很好的归宿，真心替你们开心，这一毕业就是各奔东西了，祝大家都前程似锦吧。

感谢我的同窗们李璟明、蔡天倚、王宇侠等，感谢帮助过我的学长学姐龚思胜、朱冲冲、孙晨光等，感谢你们对我学习生活上的帮助和支持。

感谢我的父母和家人，是他们无私的关怀和奉献支撑我完成了学业。

最后，感谢各位参加论文评审和论文答辩的老师们的批评与指导！