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

摘要 I
Abstract II
第 1 章 绪论 1
1.1 研究背景和意义 1
1.2 国内外研究现状 2
1.3 论文主要工作 5
1.4 论文组织结构 6
第 2 章 相关工作 7
2.1 社交网络 7
2.1.1 社交网络概述 7
2.1.2 社交网络相关概念定义 9
2.1.3 社交网络的典型特征 11
2.2 社区发现 12
2.2.1 社区结构定义 12
2.2.2 社区网络模型描述 14
2.3 社区结构评价指标 14
2.3.1 模块度 14
2.3.2 NMI 15
结论 17
参考文献 18
致谢 21

插图

图 2.1　一个简单的社交网络抽象图 8
图 2.2　社交网络中度的幂律分布曲线 12
图 2.3　一个具有社区结构的网络示意图 13
图 2.4　网络示例 15

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　社交网络

2.1.1　社交网络概述

在维基百科中，社交网络（Social Network）被定义为“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（又称社团），社交网络代表各种社会关系”。对社交网络的分析在早期只是针对现实生活中切实的方便调查的关系进行分析。比方说：早期在国外曾有研究人员在研究如何减少政府机构的冗余行政人员以提高办事效率和降低政府开销时，就使用到了社交网络分析这一手段。他们采用私下采访和调查的手段获取了某一政府机关几乎全部工作人员之间的来往接触关系，建立了一张交际网络。通过对这张交际网络的分析发现，其中有些节点在业务流程线上是属于多余的，其功能只是交接两边的节点。对于提高效率而言，分析此网络并减少这样无谓的节点即可有效的降低开销。不像早期的社交网络主要是通过合作关系建立起来的职业网络，如今随着互联网社交媒体的诞生和飞速发展，社交网络逐渐线上化。

本文所指的社交网络特指在线社交网络（下文统称社交网络）。直白简洁的说，在线社交媒介其实就是在互联网上与其他人产生联系的一个平台。在线社交媒体主要有即时通讯类软件（比如微信、QQ）、在线社交类软件（比如 Facebook、人人网）、微博类软件（比如新浪微博、Twitter）、贴吧类软件（比如百度贴吧、悟空问答、知乎）、博客分享类软件（比如 CSDN、简书）、职场关系类软件（比如领英、脉脉）和短视频分享类软件（比如抖音、快手）等等。而社交网络就是在这些社交媒介中抽象虚拟出来的一张网络图，在这张图中，每个个人或者组织抽象为一个节点，而人与人之间的关系或者互动则抽象为边。每个账号在社交媒体上填写的个人信息就是其节点属性，同样节点彼此之间的边上也有着相应的边的属性。这一切就构成了一个社交网络结构。网络虽然都是抽象出来的，但是这些关系却又都是真实的。图2.1展示了一个简单的社交网络抽象图。

在种类繁多各色各样的在线社交媒体中，人们的参与度也越来越高。2017 年 Q3 微博财报数据显示，截至 2017 年 9 月，微博月活跃用户共 3.76 亿；2018 年 1 月 15 日，在广州举行的微信公开课上微信创始人、腾讯高级副总裁张小龙指出微信用户量

图 2.1　一个简单的社交网络抽象图

已超 10 亿；而国外的 Facebook 更是早在 2017 年就已超 20 亿用户。社交网络具有传播迅速、传播广泛、自发性和言论相对自由等特点。面对这么巨大的用户量，不仅仅是普通个人用户，我们可以发现在诸多知名在线平台上各类官方媒体也都已入驻，以借助传播更加快捷和广泛的在线社交媒介来达到进行宣传等目的。当然，因为用户量的越发增加，在线社交媒介的广泛使用带来的问题也越来越多。这也对社交网络的规范化和整治不断提出挑战。

在现在这样的背景下，已经越发明显的出现线上影响线下的这种趋势。在人们享受社交网络带来的乐趣和便利之时，同样也有不法分子为了金钱或其他利益利用社交网络缺乏规范又利于传播等特点进行违法犯罪，包括诈骗、散布暴力恐怖信息或谣言等等。最近国家广电总局也整治了一大批社交媒体，封杀了一系列严重违规的软件，即使是今日头条、抖音、快手这样的大公司也面临着很大的危机，被勒令整改。因此，为了更好的利用社交网络给人们带来的便捷，同时又能避免产生危害，就产生了社交网络分析（Social Network Analysis）这一研究领域。它是一门横跨信息学、数学、计算机技术、社会学、管理学和心理学等学科的交叉科学，主要研究的是社交网络的网络结构和其演化、社交网络中的群体及其互动、社交网络中的信息及其传播。

2.1.2 社交网络相关概念定义

因为社交网络的本质其实就是一个由节点（人或组织）和边（社会关系）组成的图结构，所以说社交网络模型之中的很多概念都是来源于图论。在这一小节中，将会简单介绍社交网络中常用的几种统计概念，包括节点的度及其分布、网络密度、平均路径长度、边的介数和聚类系数。这些统计概念或多或少都旨在反映社交网络的一些特性，比如疏密程度、信息传播开销等等。

（1）节点的度（Degree）。在无向图中，任意节点的度即是与其相连的边的数目。而在有向图中，又可以细分为入度和出度。任意节点的入度就是以该节点为终点的边的数目；同样的，任意节点的出度就是以该节点为起点的边的数目。在社交网络之中，一个节点的度越大，就表示其在这个网络中扮演着越重要的角色。影响力越大的人，在网络中节点的度就越大。比如说在微博上，拥有众多粉丝的明星们，他们在社交网络中抽象出的节点度就很大，而普通用户往往只有很少的粉丝，其度就很小。网络的平均节点度就是网络中所有节点的度的平均值，它可以反映网络的疏密程度。此外，还可以通过节点度的分布来刻画描述不同节点的重要性。

（2）网络密度（Density）。在社交网络之中，网络密度被定义为网络中实际存在的边数与最多可容纳边数的比值。通常被用来测量社交网络中社交关系的紧密程度及其演变趋势，其计算方式详见公式2.1。如果一个社交网络的网络密度还很小，则说明该网络还尚且处在起步阶段；而若一个社交网络的网络密度已经比较大了，那么说明该网络已经比较成熟，网络之中几乎所有节点之间都有联系。

$$Density = \frac{2m}{n(n-1))} \tag{2.1}$$

公式2.1中的 n 和 m 分别为社交网络中边的数目和节点的数目，且 $Density \in [0,1]$。其中，当整个网络中没有一条边，即所有节点都独立存在时，$Density$ 取 0；而当网络中所有节点之间都有边相连时，即网络处于全连接状态时，$Density$ 取 1。一般而言，大规模的社交网络的密度会比中小规模的小一些，因此，不同规模之间的网络也就不具有可比性了。这也不难理解，举个简单的例子，以学校为规模建立一个社交网络和以一个家庭为规模建立一个社交网络，显然以一个家庭社交网络的网络密度会大很多。

（3）平均路径长度（Average Path Length）。一个社交网络的平均路径长度被定义

为任意两个节点之间的最短路径的平均长度，也就是任意两个节点之间的最短关系路径上节点个数的平均值。其计算方法详见公式2.2。

$$APL = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} \tag{2.2}$$

公式2.2中的 n 为当前网络节点个数，d_{ij} 为网络中节点 v_i 和 v_j 之间的最短路径。平均路径长度 APL 通常也是用于反映网络间的紧密程度的，一般也可叫做网络的平均距离。如果社交网络的平均路径长度比较大，则代表网络比较稀疏，节点之间进行信息传播的开销比较大；相反，若社交网络的平均路径长度小，则代表网络稠密，节点之间可以比较迅速快捷的进行传递消息。

（4）聚类系数（Clustering Coefficient）。根据图论，聚类系数表示的是一个图中节点汇聚程度的系数。在很多社交网络中，若节点 v_i 与节点 v_j 相连接，而节点 v_j 与节点 v_k 相连接, 那么很大概率上节点 v_i 和节点 v_k 也会相连。这种现象也表明了社交网络中部分节点之间存在着密集连接的这一特性。在无向图中，节点 v_j 的聚类系数 CC_{v_j} 的计算方式详见公式2.3。

$$CC_{v_j} = \frac{n}{C_k^2} = \frac{2n}{k(k-1)} \tag{2.3}$$

公式2.3中 k 表示节点 v_j 所拥有的邻居节点数目，n 表示节点 v_j 的所有相邻节点之间互相连接的边的数目。简单来说，聚类系数可以用来描绘社交网络中一个用户朋友们之间也是朋友的概率，反映的也就是社交网络的聚集性。具体的，它还可以分为全局聚类系数和局部聚类系数，这里不再赘述。

（5）介数（Betweeness）。介数可以分为节点介数和边介数，表示的是网络图中某一节点或者某一条边被整个图中所有节点间的最短路径经过的概率之和。通常是用来评价节点的重要程度的。比方说在连接不同社区之间的中间节点（或者边）的介数就会比其他节点（或者边）的介数要大很多，这也反映了这类节点在社交网站中作为消息传播的核心地位及其重要程度。对于网络中任意节点 v，其介数的计算方式详见公式2.4。

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.4}$$

在公式2.4中，$\sigma_{st}(v)$ 表示经过节点 v 的 $s \to t$ 的最短路径条数，σ_{st} 表示 $s \to t$ 的最短路径条数。直观上来看，节点 v 的介数 $C_B(v)$ 反映的是节点 v 作为“桥梁”或者“枢纽”的重要程度。

2.1.3 社交网络的典型特征

在社交网络之中普遍存在着两个典型的特征：小世界效应和无标度特性。

（1）小世界效应（Small-world Effect）。在 1929 年匈牙利作家 F.Karinthy 率先提出了“小世界现象”的论断。他认为，地球上的任何两个人都可以平均通过一条由 5 位联系人组成的链条而联系起来。在 1967 年，美国哈佛大学的社会心理学教授 Stanley Milgram 提出了著名的“六度分隔（Six Degrees of Separation）假说”，大意同样为任何两个想要取得联系的陌生人之间最多只隔着 5 个人，便可完成两人之间的联系。他通过设计了一个信件实验来证明他的猜想，实验大致经过为：他随机选择了 300 多人，每人分发了一封信并指定了各不相同的收信人；要求如果寄信人认识收信人，则直接寄出，否则就寄给一个自己认识的并且可能认识收信人的人，直至收信人收到信为止；实验最终共有约 60 人收到了信，而这些信平均经手了 6 次就到达了收信人手中。在 1998 年的时候，Duncan Watts 和 Steven Strogatz 正式提出了小世界网络的概念并建立了小世界模型[33]。文中将小世界效应定义为：若网络中任意两个节点之间的平均距离（即平均路径长度 APL）随网络中节点数 n 的增加呈对数增长，即 $APL \sim ln(n)$，且网络的局部结构上仍然具有较明显的集团化特征。

小世界效应反映的是社交网络中任何用户之间都近在咫尺的现象，简单来说就是社交网络的平均路径长度都很短。小世界现象在在线社交网络中得到了很好地验证，根据 2011 年 Facebook 数据分析小组的报告，Facebook 约 7.2 亿用户中任意两个用户间的平均路径长度仅为 4.74，而这一指标在 Twitter 中为 4.67。因此可以说，在五步之内，任何两个网络上的个体都可以互相连接。

（2）无标度特性。大多数社交网络都存在着少数节点的度极大，而大部分节点都只有较小的度这一现象。其网络缺乏一个统一的衡量尺度而呈现出异质性，我们将这种节点度分布不存在有限衡量分布范围的性质称为无标度。这其实体现的是社交网络中用户的度呈现出幂律分布的规律。其实幂律分布广泛存在于各个领域，其核心就是绝大部分事件的规模其实很小，但是极少数事件的规模却表现的相当大，直观上就像幂函数2.2的函数曲线一样。举几个简单的例子，比如说世界上绝大部分的财富都被

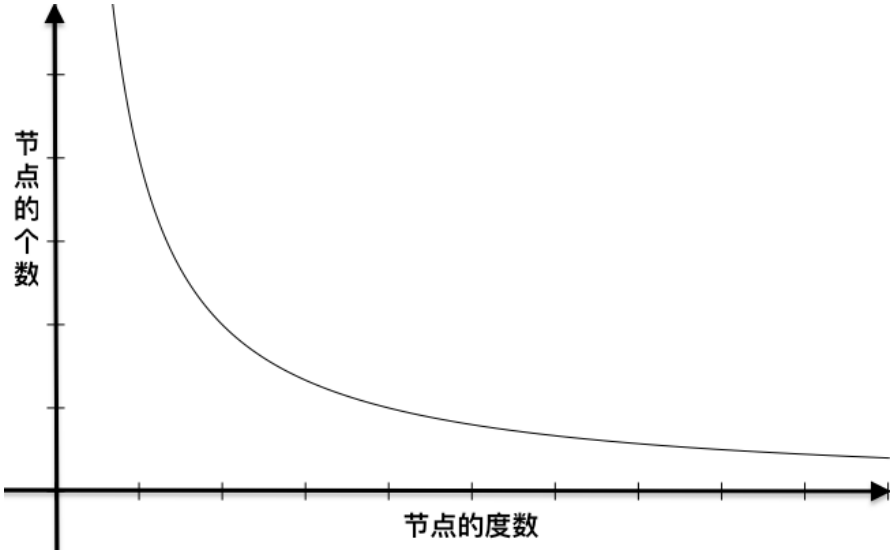

图 2.2　社交网络中度的幂律分布曲线

掌握在极少数的超级富豪们的手中；再拿网站的访问量来说，尽管互联网之上为广大网民提供了无数的网页，但是每天大家访问量最多的也就是那么几个熟悉的网页；又比方说在社交网络中，例如微博上，一个明星的粉丝可能有上百万上千万，但是大部分人也只有寥寥无几的粉丝关注量。幂律分布其实体现的是一种极端的不平衡性。

2.2　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。

2.2.1　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

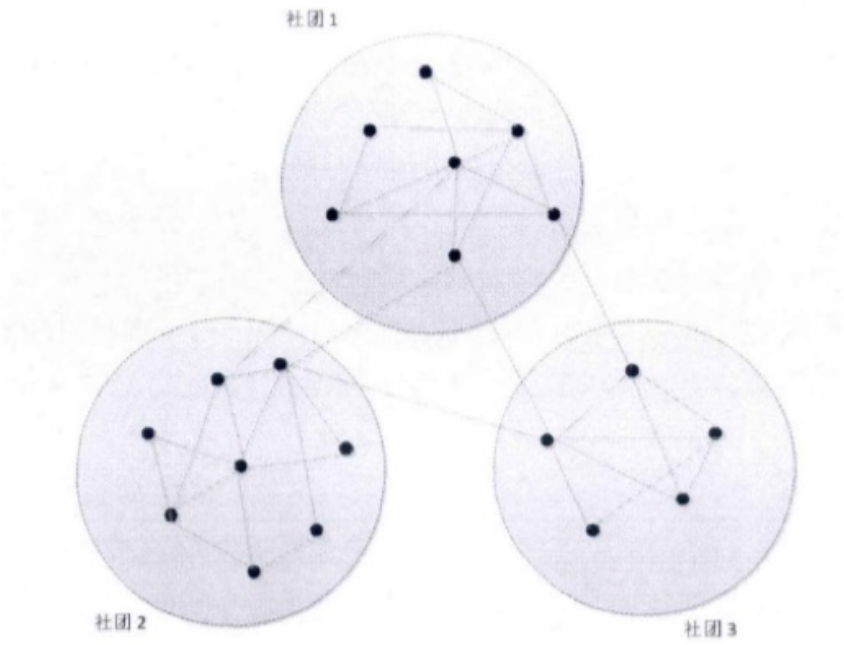

图 2.3　一个具有社区结构的网络示意图

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2.3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

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。

2.2.2 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。

2.3 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。

2.3.1 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

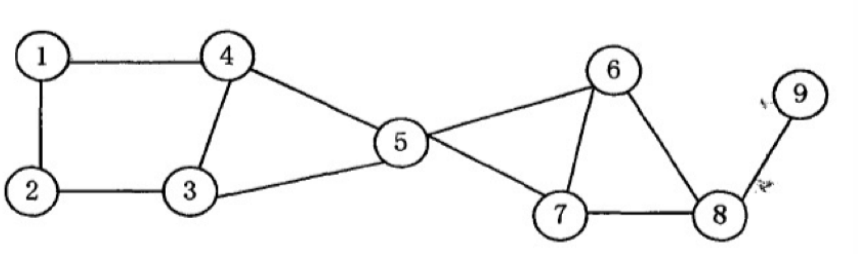

图 2.4　网络示例

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。模块度计算方式详见公式2.5。

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{2.5}$$

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 $\delta(c_i, c_j) = 0$

.....

2.3.2　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 计算方式详见公式2.6。

$$NMI = \frac{-2\sum_{i,j} N_{ij} ln\frac{N_{ij}}{N_i N_j}}{\sum_i N_i ln\frac{N_i}{n} + \sum_j N_j ln\frac{N_j}{n}} \tag{2.6}$$

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2.4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2.7。

$$N = \begin{bmatrix} 0 & 0 & 4 \\ 0 & 1 & 0 \\ 4 & 0 & 0 \end{bmatrix} \tag{2.7}$$

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.

[33] DJ W, SH S. Collectivedynamics of ’small-world’networks[C]. Nature. [S.l.: s.n.], 1998: 440–442.

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！