普通高等教育“十二五”规划教材

高等院校重点推荐教材

复杂网络基础理论

郭世泽　陆哲明　编著

科学出版社

普通高等教育“十二五”规划教材
高等院校重点推荐教材

复杂网络基础理论

郭世泽　陆哲明　编著

科学出版社
北京

内 容 简 介

本书是复杂网络基础理论的入门教材。本书取材广泛，内容全面、新颖，反映了十多年来复杂网络理论的最新研究动态和作者的部分研究成果。全书共分八章。第一章主要介绍与复杂网络有关的基本知识。第二章详细介绍了网络的拓扑结构和静态特征。第三章从机制模型的不同类型入手分别介绍了规则网络、随机网络、小世界网络、无标度网络、层次网络、确定性网络和自相似网络。第四章介绍复杂网络上的流行病传播、舆论传播和知识传播以及数据包传递和拥塞控制机理。第五章介绍复杂网络的混沌同步。第六章介绍复杂网络中的搜索算法与策略、社会网络的分散式搜索以及 P2P 网络和 WWW 网络中的搜索。第七章介绍复杂网络中的重要节点挖掘和社团挖掘原理和算法。第八章介绍复杂网络中的演化博弈、抗毁性分析以及抗毁性优化和修复策略等。为便于消化和理解书中内容，每章末附有习题，书末给出大量参考文献。

本书可作为高等院校计算机科学与技术、通信工程、应用数学、生物科学、应用物理学、社会学等专业本科生和研究生的教材，也可供从事复杂性科学和网络科学等领域工作的教学、科研人员参考。

图书在版编目（CIP）数据

复杂网络基础理论/郭世泽，陆哲明编著．—北京：科学出版社，2012
（普通高等教育“十二五”规划教材·高等院校重点推荐教材）
ISBN 978-7-03-034599-8

Ⅰ.①复… Ⅱ.①郭…②陆… Ⅲ.①计算机网络-高等学校-教材
Ⅳ.①TP393

中国版本图书馆 CIP 数据核字（2011）第 115630 号

责任编辑：鞠丽娜 / 责任校对：耿 耘
责任印制：吕春珉 / 封面设计：三函设计

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

印刷
科学出版社发行 各地新华书店经销

*

2012 年 6 月第 一 版 开本：787×1092 1/16
2012 年 6 月第一次印刷 印张：22
字数：504 000

定价：40.00 元

（如有印装质量问题，我社负责调换〈环伟〉）
销售部电话 010-62134988 编辑部电话 010-62135763-8002

前　言

科学根植在理论基础之上，理论是对经验现象或事实的科学解说和系统解释。顾名思义，复杂网络作为一门新兴科学，是对存在的网络现象及其复杂性进行解释的学科。首先，它研究的是网络现象。网络在自然界和人类社会中普遍存在，包括自然界中天然存在的星系、食物链网络、神经网络、蛋白质网络；人类社会中存在的社交网络、传染病传播网络、知识传播网络；人类创造的交通网络、通信网络、计算机网络等。网络科学作为一门交叉学科，主要研究利用网络特性描述物理、生物和社会等现象，进而建立这些现象的预测模型或分析模型，并利用网络的静态特性和动力学特性来解释这些现象。

其次，它代表着对网络自身复杂性的深化认识。从1736年到1958年漫长的400多年里，人们一直用基于图论的规则网络理论来研究与网络有关的问题。1959年，匈牙利数学家Erdös和Rényi首次将随机性引入到网络建模中，创立了随机网络理论，提出了ER随机图模型，使网络科学迈入了第二个重要的发展阶段。1959年到1998年的近40年时间里，随机网络理论曾一度被公认为是正确认识真实网络的基本理论。20世纪80年代，复杂性科学（complexity science）兴起，引发了自然科学界方法论的变革，并且日益渗透到哲学、人文社会科学领域。著名物理学家霍金称："21世纪是复杂性科学的世纪。"1999年，美国"Science"杂志出版了一期以"复杂系统"为主题的专辑，分别就化学、生物学、神经学、动物学、自然地理、气候学、经济学等学科领域中的复杂性研究进行了报道。概括起来，复杂系统都有一些共同的特点，就是在变化无常的活动背后呈现出某种捉摸不定的秩序，其中演化、涌现、自组织、自适应、自相似被认为是复杂系统的共同特征。在这个研究背景下，应用复杂性科学解释网络现象的复杂网络理论应蕴而生。美国康奈尔大学理论和应用力学系博士生Watts及其导师Strogatz于1998年在"Nature"杂志上发表题为"小世界网络的群体动力行为"的论文，提出了一种介于规则网络和随机网络之间的网络模型——小世界网络模型。随之，1999年美国圣母大学物理系Barabási教授及其博士生Albert在"Science"杂志上发表题为"随机网络中标度的涌现"的论文，提出了一个无标度网络模型，引起了全世界的高度重视。小世界网络和无标度网络模型的提出标志着网络科学进入了一个新时代。随后的许多真实网络实证研究表明，真实世界网络既不是规则网络，也不是随机网络，而是兼具小世界和无标度特性，具有与规则网络和随机网络完全不同的统计特性。面对快速发展的Internet和WWW网络，还有其他各种社会、生物、物理网络，科学家们发现已无法用规则网络和随机网络理论来解释它们的结构和演化的许多新问题，因而粗略地称这类网络为复杂网络。

复杂网络至今还没有一个统一公认的定义。人们通常采用复杂网络表现出来的不同于规则网络和随机网络的特性来表征复杂网络，如小世界特性、无标度特性、层次特性、自相似特性、自组织特性等。目前，复杂网络的研究工作集中在以下几个方面：①复杂网络拓扑结构的静态统计分析，包括更广泛的实证研究和更深入的理论刻画；

②复杂网络的演化和机制模型，实证上可以研究实际网络演化的统计规律，如检验 BA 模型的择优连接假设；理论上则可以发展完善的具有形成特定几何性质的网络机制模型；③复杂网络上的动力学研究，包括网络容错性和鲁棒性以及网络上的搜索、传播、演化博弈、同步与共振等各种动力学过程；④有关复杂网络的分析方法与应用研究。总的来说，网络的结构与功能及其相互关系是网络研究的主要内容，结构与功能的相互作用特别是其对网络演化的影响是复杂网络研究需要解决的重要问题。

自 2000 年以来，有关复杂网络的研究论著不断涌现。本书作为基础理论教材，旨在介绍十多年来复杂网络领域公认的广泛提及和深入研究的一些基础理论，尽量以浅显易懂的方式为来自不同学科领域的本科生和研究生学习复杂网络理论提供指导。在学习本书之前，读者需要具备微积分、概率论和图论的基础知识，了解信息论、控制论、博弈论的相关概念，掌握必要的计算机编程语言和仿真验证技能。本书共八章，分别是绪论、网络拓扑结构与静态特征、网络机制模型、复杂网络上的传播动力学、复杂网络的混沌同步、复杂网络上的搜索、复杂网络中的挖掘以及复杂网络中的博弈。为了让读者更好地理解课程内容，各章都配有例题和习题。不管针对何种读者对象，第一章、第二章和第三章是必修课程，每章需要 8 学时课堂授课和 2 学时上机实验。后面的章节针对不同的读者可以有所侧重，例如，对于计算机网络安全人员，第四章、第七章和第八章需要重点学习。针对本科生而言，除了前三章，建议学习第四章、第六章和第七章，每章需要 8 学时课堂授课和 2 学时上机实验。针对研究生而言，建议学习所有章节，一共 64 学时课堂授课和 16 学时上机实验。各章内容简述如下：

第一章：图论是研究网络结构与特性的最有效的理论工具之一，复杂网络中的大多数概念和特性可以从图论中找到依据。其次，数理统计和概率论是进行复杂网络建模和特性分析的一种有力的数学工具。因此，第一章绪论主要介绍与复杂网络有关的基本知识，内容包括：网络科学理论发展的历史、复杂网络的概念与特性、数理统计基础、图论的基本概念、复杂网络的研究内容和意义。

第二章：复杂网络具有很多与规则网络和随机网络不同的统计特征。因此，究竟用哪些统计特性来描述一个给定网络呢？刻画一个网络最基本的三种特征量是度分布、集聚系数和平均距离，其中度分布描述了每个节点的邻居节点个数的分布情况，如幂律度分布是无标度网络的特性；集聚系数描述了同一个节点相连接的两节点之间也相互连接的概率，如对社交网络而言，代表着如下情况出现的程度：你朋友的朋友也是你的朋友或者你的两个朋友彼此也是朋友；平均距离描述了平均意义上网络中两个节点之间的最短路径所经历的边的条数，如小世界网络具有较小的平均距离。本书的第二章详细介绍了网络的拓扑结构和静态特征，首先介绍网络的三个基本静态特征，接着按无向网络、有向网络和加权网络分别讨论网络的各种基本静态特征，然后简要介绍近些年文献中提出的其他一些新的静态特征，最后简要介绍了复杂网络分析软件 Pajek。

第三章：在现实生活中，大家经常会感慨世界很小，尤其在当今发达的通信技术和网络技术背景下，地球上的任何两个人平均经由 6 个朋友的牵线搭桥，就可以连通，这就是一种著名的称为“六度分离”的小世界现象。在 Internet 中，通常是越著名的

网站，链接它的网站就越多，从而形成许多“hub”网站，整个网络呈现无标度特性和“富者愈富”的效应。那么，如何能够构造具有小世界特性的人际关系网络模型和具有无标度特性的网络模型，以便人们进行后续的特性分析和动力学行为分析，这就是网络机制模型问题。本书第三章从机制模型的不同类型入手分别介绍了规则网络、随机网络、小世界网络、无标度网络、层次网络、确定性网络、自相似网络，并在每个小节中比较详细的介绍了与这些模型有关的知识。

第四章：在人类社会中，传染病夺去了无数人们的生命，为什么只要将感染率控制在一定的阈值之下，传染病的传播就可以得到有效控制呢？而在计算机网络中，病毒种类越来越多，对 Internet 安全构成极大威胁，为什么没有一种杀毒软件可以查杀所有可能出现的病毒，没有一种防护措施可将所有病毒拒之门外呢？实际上，病毒在计算机网络上的蔓延，传染病在人群中的流行，谣言在社会中的扩散，舆论和知识的传播，网络中的数据包传递等都可以看作是服从某种规律的网络传播行为。本书第四章首先介绍复杂网络上的流行病传播机理，接着介绍复杂网络的免疫策略，然后介绍复杂网络上的舆论传播和知识传播，最后介绍复杂网络上的数据包传递机理和拥塞控制等。

第五章：大家在日常生活中可能都有过如下的发现：夏天的萤火虫会有规律的同时发光，或者同时不发光；窗外青蛙和蟋蟀的叫声一段时间后就会同时叫，或同时不叫；剧场中观众掌声频率逐渐趋于一致等，这就是同步化现象。本书的第五章介绍了复杂网络上的混沌同步，首先简要介绍混沌理论，然后概述混沌同步的概念和方法，接着引出一般意义上的复杂网络完全同步问题及其稳定性分析方法，最后讨论典型复杂动态网络在线性耗散耦合条件下的混沌同步问题。

第六章：在信息爆炸的当今网络时代，如何能够在网络中快速搜索到所需要的信息或文件？在交通网络中，如何快速确定任意两个城市之间的最短路径？在社会网络中，如何快速有效地搜寻罪犯、恐怖分子或失散亲人？这些都是复杂网络中的搜索问题。本书第六章首先介绍三种经典的搜索策略，即广度优先搜索算法、随机行走搜索算法和最大度搜索算法，然后介绍社会网络的快速分散式搜索问题，最后介绍 P2P 网络和 WWW 网络的搜索问题。

第七章：首先，在罪犯关系网络中，发掘网络中的重要节点可以迅速定位犯罪团伙的头目，集中警力进行布控；在电力网络中，对重要的断路器、发电单元等进行保护，可以有效防止由相继故障引起的大范围停电；在谣言传播网络中，通过发掘出始作俑者来避免“蝴蝶效应”，这些都是重要节点挖掘问题。其次，在许多现实网络中，社团结构是主要特征之一，如在人际关系网中，社团可能按照人的职业、年龄等因素来划分；在新陈代谢网和神经网络中，社团可能反映了功能单元；在食物链网中，社团可能反映了生态系统中的子系统，这些可以归结为社团挖掘问题。本书第七章从节点挖掘和社团挖掘两个方面入手，详细介绍这两方面的相关知识，并对典型的分析算法进行细致讲解。

第八章：群体合作现象的涌现和稳定维持可以说是自然界中最令人兴趣盎然、也是最令人疑惑的问题。如何理解和解释合作行为的广泛存在和稳定维持是人们面临的

最大的挑战之一，复杂网络的演化博弈论扮演了重要角色，并提供了强有力的理论框架。另外，随着复杂网络理论研究和应用研究的深入，人们开始关注：复杂网络到底有多可靠？在网络遭受损失后如何高效快速的修复网络？例如，生物领域基因网络中的一些核心基因的故障会带来灾难性的后果，军事领域中网络中心战将成为未来战争的主要样式，网络系统的抗毁性将直接关系到整个战争的成败，这些可以归结为复杂网络中的博弈问题。本书第八章首先概述博弈论，接着重点介绍复杂网络的演化博弈理论，然后重点介绍复杂网络中的抗毁性分析，最后简要介绍复杂网络的抗毁性优化和修复策略。

本书可供从事复杂性科学、网络科学、图论、系统工程、计算机网络、统计物理学、生命网络分析、社会网络分析、传播动力学、演化博弈论等研究领域的科技人员与教学人员参考，并可以作为上述专业本科生和研究生的教材。本书的每一章都是由郭世泽研究员提供总体思路和初稿框架，由陆哲明教授完善细化和修改审定。本书广泛参考了国内外复杂网络研究领域的学术论文、学位论文和学术著作，尤其从如下三本专著中得到很多帮助：2006 年清华大学出版社出版由汪小帆、李翔和陈关荣合作编著的《复杂网络理论及其应用》、2006 年上海科技教育出版社出版由郭雷和许晓鸣编著的《复杂网络》以及 2009 年高等教育出版社出版由何大韧、刘宗华、汪秉宏等编著的《复杂系统与复杂网络》。本书作者及其研究团队不仅要感谢国际学术大师 Strogatz、Watts、Barabási、Albert、Newman 等为我们开创了复杂网络研究领域，也非常感谢国内学者们在学术界开展的研究工作，包括陈关荣、汪小帆、方锦清、李翔、刘宗华、汪秉宏、章忠志等。本书得到了多个国家自然科学基金项目（项目编号：61171150；项目编号：61003255；项目编号：61070208；项目编号：61071128）和浙江省杰出青年基金项目（项目编号：R1110006）的资助，在此致以深深的谢意。在本书的撰写过程中还得到了总参第五十四研究所、北京邮电大学、浙江大学航天电子工程研究所各位教师、博士生和硕士生的协助，在此表示衷心地感谢。

“放眼何能一叶休，登攀绝顶作身谋。梦中借我春秋笔，挥却丹书始探究。”可以说，我们现在所看到的复杂网络科学只是冰山一角，尚有大量认识不清、理论不足、验证不够的研究领域，亟待我们抱着科学的态度去研究、探索、发现。此书是我们团队关于复杂网络科学的第一本粗习之作，相信随着研究的开展和深入，我们还将为业界奉献更有分量的论著。我们团队的座右铭是：前面的高山是如此巍峨美丽，让我们一起去攀登吧！

限于水平，书中难免有错误与不妥之处，恳请读者批评指正。

郭世泽
于北京总参第五十四研究所
陆哲明
于杭州浙江大学航空航天学院航天电子工程研究所
2012 年 5 月

目　　录

第一章　绪　　论

1.1　引　　言

人类从远古走来，很早就构造出山间小路和林中小径，并且把小路连成网络；在农业社会，人又构造出各种水利网络，便于灌溉；通过航海网络，资本主义才走遍全世界；在工业社会，普通的小路被公路、铁路和高速公路所淹没，公路网和铁路网极大方便了人们的交流和贸易；在今天的信息时代，各个国家致力于建设自己的信息高速公路，即新型的信息网路，才会使如今的 Internet/WWW 网络基本覆盖整个世界。有太多的网络与人们生活息息相关，除了上面介绍的网络外，还有通信网络、电力网络、航空网络、银行网络、商业网络等。人类把自己生存的世界变成了网络世界，网络越发达、越有效，世界就越小，人的社会性就越得到强化。现如今，网络显得如此广泛、如此重要，人类已处在网络的重重包围之中。如何开辟出一条研究思路，揭示网络拓扑结构的形成机制，探索网络的演化规律和整体行为，认识网络内部深奥的动力学特性，挖掘网络展现出的广泛、潜在的应用价值等问题，正引起国内外学术界的高度重视[1]。

21 世纪是复杂性和网络化的世纪。从 20 世纪七八十年代开始，复杂性问题的研究已引起国内外关注，并与非线性科学及其混沌动力学的复杂性研究交错在一起，也因此在国际上形成了非线性科学和复杂性问题的研究热潮[2]。来自各国的不同学科的科学家，包括物理学家、生物学家、计算机科学家和经济学家等都开始不约而同地讨论和研究各自领域的复杂性问题，例如自组织现象和自组织临界性、自适应问题、计算机与智能问题、生命与生物的演化、全球经济的演化、人类文化和语言的演变等。另一方面，随着人类认识能力的进一步提高，人们发现自然界、人类社会、生物群体中的许多复杂系统都可以通过网络的概念加以描述。尤其是 20 世纪 90 年代以来，随着计算机技术和 Internet 技术的迅猛发展，标志着人类迈入了网络时代。人类已经生活在一个充满各种各样复杂网络的世界中[3]：从 Internet 到 WWW，从大型电力网络到全球交通网络，从生物体中的大脑到各种新陈代谢网络，从科研合作网络到各种经济、政治、社会关系网络等。甚至像语言和软件等许多在常人眼里并非网络的东西也可以从复杂网络的角度去研究。由于复杂性问题研究与网络的复杂性关系有密切联系，所以近年来它们之间的交叉研究引起了人们的高度重视。在上述背景下，网络科学成为 21 世纪兴起的多学科交叉的研究领域，它关注的是复杂网络的共性和处理它们的有效方法，从而增进人类对各种自然和人工复杂网络的科学理解。从网络观点重新认识事物有可能带来革命性变化的一个典型例子就是 Google 的诞生。与之前的搜索引擎相比，Google 的一个主要突破是它的 PageRank 算法利用了 WWW 的网络结构[3]。

实际上，随着生命科学的发展、网络时代的到来以及人们交流和经济活动的全球化，人们早就开始观察和思考生命网络、技术网络、交通网络、社会网络等呈现的一些普遍现象或问题。例如，计算机病毒如何在 Internet 网中传播？艾滋病、禽流感、SARS 等疾病如何在人类和动物中传染？为什么美国的次贷危机会引发全球的经济危机？为什么各大城市堵车现象那么严重？生物体的神经系统和新陈代谢系统是如何开展工作的？是否有时你会因为“朋友认识的人刚好也是你的朋友”而感慨这个世界太小？为什么鼓掌时大家的节奏会趋于一致？所有这些问题看上去互不相关，实际上这些都是复杂网络所反映的普遍规律和复杂网络领域学者们所要研究的课题。

近 10 年来，复杂网络的研究正渗透到从物理学到生物学的众多不同的学科，对复杂网络的定性特征与定量规律的深入探索、科学理解以及可能的应用已成为网络时代复杂性科学研究中一个极其重要的挑战性课题。人们对网络的复杂性奥妙的研究只是揭示了冰山一角，大量复杂网络的奥秘有待于深入探索和研究。推进复杂性科学的交叉研究，深入探索和科学理解复杂网络的定性特征与定量规律，使它获得广泛的应用，对全球科学和社会的发展具有十分重大的长远意义[2]。

1.2 网络科学理论发展的三个时期

1.2.1 规则网络理论阶段

规则网络理论的发展得益于图论和拓扑学等应用数学的发展[4]。历史上，多位杰出数学家各自独立地建立和研究过图论，其中，欧拉于 1736 年首先开创了图论这门新的数学分支，因此他被誉为“图论之父”。直到两百年后的 1936 年，图论的第一本专著《有限图与无限图的理论》才由匈牙利数学家 König 发表出来。实际上，图论的研究对象就是由一些节点按照一定方式连线组成的图（集合）。用图论的语言和符号可以精确简洁地描述各种网络，为物理学家和数学家提供了共同描述语言和平台，图论的许多研究思想、技巧、成果和结论（如解决网络最短路问题、最大流问题、最小费用最大流问题的算法）能够自然地根植到复杂网络的研究中来，因此图论是一种强有力的研究工具和研究方法。以下是历史上著名的四个图论问题。

1. 哥尼斯堡七桥问题

关于图论的文字记载最早出现在 1736 年瑞士数学家欧拉的论著中，他所考虑的原始问题具有很强的实际背景，那就是著名的哥尼斯堡七桥问题[4]。哥尼斯堡是当时东普鲁士的首都，今俄罗斯加里宁格勒市，普莱格尔河横贯其中，这条河上建有七座桥，将河中间的两个岛和河岸联结起来，如图 1.1 所示。有人在闲暇散步时提出：能不能每座桥都只走一遍，最后又回到原来的位置。这个看起来很简单却很有趣的问题吸引了大家，很多人在尝试各种各样的走法，然而无数次的尝试都没有成功。1736 年，有人带着这

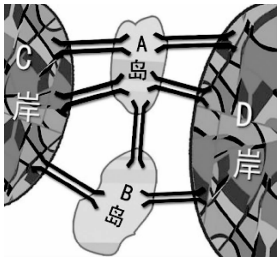

图 1.1 哥尼斯堡七桥问题示意图

个问题找到了当时的大数学家欧拉，欧拉经过一番思考，很快就用一种独特的方法给出了解答。他把两座小岛和河的两岸分别看作四个点，而把七座桥看作这四个点之间的连线，于是这个问题就简化成：能不能用一笔就把这个图形画出来？经过进一步的分析，欧拉得出结论：不可能每座桥都走一遍，最后回到原来的位置，并且给出了所有能够一笔画出来的图形所应具有的条件。这项工作使欧拉成为图论（及拓扑学）的创始人。

2. 哈密顿问题

哈密顿问题也是图论中的著名问题之一。英国数学家哈密顿于 1859 年以游戏的形式提出：把一个正十二面体的二十个节点看成二十个城市，要求找出一条经过每个城市恰好一次而回到出发点的路线，如图 1.2 所示。这条路线就称“哈密顿圈”。换一种说法，对于一个给定的网络，在确定起点和终点后，如果存在一条路径能够穿过该网络，就称该网络存在“哈密顿路径”。一百多年来，对哈密顿问题的研究，促进了图论的发展。哈密顿路径问题在 20 世纪 70 年代初，终于被证明是“NP 完备”的，也就是说具有这样性质的问题，难于找到一个有效的算法。实际上对于某些节点数不到 100 的网络，利用现有最好的算法和计算机可能也需要几百年才能确定是否存在一条这样的路径。

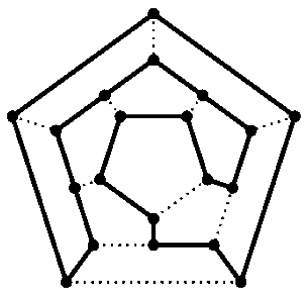

图 1.2 哈密顿问题示意图

3. 四色猜想

四色猜想又称四色问题、四色定理，是世界近代三大数学难题之一[4]。1852 年，毕业于伦敦大学的格思里来到一家科研单位做地图着色工作时，发现了一个有趣的现象：“看来，每幅地图都可以用四种颜色着色，使得有共同边界的国家着上不同的颜色。”这个结论能不能从数学上加以严格证明呢？他和在大学读书的弟弟格里斯决心试一试，可是研究工作没有进展。他们求教著名数学家摩尔根和哈密顿也均没能解决该问题。1872 年，英国当时最著名的数学家凯利正式向伦敦数学学会提出了该问题，于是四色猜想成了世界数学界普遍关注的问题。1878～1880 年两年间，著名的律师兼数学家肯普和泰勒两人分别提交了证明四色猜想的论文，宣布证明了四色定理。但是到了 1890 年，数学家赫伍德以自己的精确计算指出了肯普的证明存在漏洞，而不久之后，泰勒的证明也被人们否定了。进入 20 世纪以来，随着电子计算机的问世，由于演算速度迅速提高，加之人机对话的出现，大大加快了对四色猜想证明的进程。1976 年，美国伊利诺伊大学哈肯和阿佩尔在大学里的两台不同的电子计算机上，用了 1200 个小时，作了 100 亿判断，终于完成了四色定理的证明。四色猜想的计算机证明，轰动了世界。它不仅解决了一个历时 100 多年的难题，而且成为了数学史上一系列新思维的起点。不过也有不少数学家并不满足于计算机取得的成就，他们还在寻找一种简捷明快的书面证明方法。

4. 旅行商问题

旅行商问题（traveling salesman problem，TSP）也叫货担郎问题或中国邮路问题等，是计算机算法理论历史上的经典问题。在过去几十年中，它成为许多重要算法思想的测试平台，同时也促使一些新的理论领域的产生，比如多面体理论和复杂性理论。该问题描述如下：给定 N 个节点和任意一对节点 $\{v_i, v_j\}$ 之间的距离为 dist（v_i，v_j），要求找出一条闭合的回路，该回路经过每个节点有且仅有一次，并且该回路的费用最小（这里的费用是指每段路径的距离和）。实际上，旅行商问题就是加权的哈密顿路径问题，因此求解旅行商问题的精确解是 NP 难的。若将问题限定在欧氏平面上，就称为欧几里德旅行商问题，但是它也是 NP 难的。因此，通常用来解决 TSP 问题的解法都是近似算法。第一个欧几里德旅行商问题的多项式近似算法是由 Arora 于 1998 年使用随机平面分割和动态规划方法给出的（发表在期刊“Journal of the ACM”的 1998 年 45 卷第五期）。

1.2.2 随机网络理论阶段

图论的第一本专著的出版使得图论研究进入了快速发展和突破期。1959 年，两个匈牙利著名的数学家 Erdös 和 Rényi 又一次对图论作出了第二个里程碑式的贡献。他们建立了著名的随机图理论，用相对简单的随机图来描述网络，简称 ER 随机图理论[4]。ER 随机图理论对图论理论研究的影响长达近 40 年，以至于在随后的近半个世纪，随机图一直是科学家研究真实网络最有力的武器。随机网络是指在由 N 个节点构成的图中以概率 p 随机连接任意两个节点而成的网络，即两个节点之间连边与否不再是确定的事，而是由概率 p 决定。或简单地说，在由 N 个节点构成的图中，可以存在 $N(N-1)/2$ 条边，从中随机连接 M 条边所构成的网络就叫随机网络。如果选择 $M=pN(N-1)/2$，则这两种构造随机网络模型的方法就可以联系起来。随机图和经典图之间最大的区别在于引入了随机的方法，使得图的空间变得更大，其数学性质也发生了巨大的变化。Erdös 和 Rényi 系统研究了当 $N\to\infty$ 时随机图性质与概率 p 的关系，他们发现：随机网络的许多重要的性质都是随着网络规模的扩大而突然出现的，也就是说对于给定概率 p，随着网络规模的扩大，要么几乎所有的随机图具有某种性质，要么几乎每一个图都不具有该性质。

Erdös 被称为 20 世纪的欧拉，于 1984 年获得沃尔夫奖[4]。他善于与人合作，打破了数学领域的喜欢个人独立研究的传统，一生中同 511 位合作者发表过约 1500 篇数学论文，涉及数学的许多领域，与那些伟大的物理学家和数学家，如爱因斯坦、哥德尔、奥本海姆等有都密切学术交往。Erdös 是一个数学“苦行僧”，他从来没有固定的职位，从来不定居在一个地方，也没有结婚，但更是一个流浪学者，随时带着手提箱，穿梭于学术研讨会，浪迹天涯，颇富传奇色彩。几乎每一个当代数学家都有一个有限的 Erdös 数，而且这个数往往非常小，小得出乎本人的预料。

1.2.3 复杂网络理论阶段

1. 小世界效应的发现

1998 年，网络科学又一次取得突破性进展，出现了第三个里程碑，这在很大程度上要归功于计算机技术和 Internet 技术的迅猛发展[4]。美国的瓦茨和斯特罗加茨首先冲破了随机网络理论的框框，于 1998 年在“Nature”上发表了题为《“小世界”网络的群体动力行为》的论文[5]，他们推广了“六度分离”的科学假设，提出了小世界网络模型。“六度分离”最早来自于 20 世纪 60 年代美国哈佛大学心理学家 Milgram 对社会调查的推断，是指在大多数人中，任意两个素不相识的人通过朋友的朋友，平均最多通过 6 个人就能够彼此认识。为了进一步验证六度分离假设，1997 年三个美国大学生发明了一个名叫“Kevin Bacon”的游戏（他们认为美国电影演员 Kevin Bacon 是电影界的中心）。世界上每位电影演员，通过跟他共同演过电影的演员，可以最终联系到 Kevin Bacon。如果一个人跟 Kevin Bacon 演过电影，他的 Bacon 数就是 1，如果一个人跟 Kevin Bacon 演过电影的人演过电影，他的 Bacon 数就是 2，以此类推。比如成龙演了“Around the World in 80 Days”（2004 年），其中有 Luke Wilson，而 Luke Wilson 演了“My Dog Skip”（2000 年），其中就有 Kevin Bacon。所以成龙的 Bacon 数是 2。人们可以访问网站 http://oracleofbacon.org/去查任何一个演员的 Bacon 数，到 2010 年 5 月数据库里存有 3 585 458 个演员信息以及 1 604 556 部电影电视信息。表 1.1 是对所有这些近 360 万个演员所做的统计，左边是 Bacon 数，右边是拥有这个 Bacon 数的演员个数，可以看到最大的 Bacon 数仅仅为 8，而近 360 万个演员的平均 Bacon 数仅为 2.98。

表 1.1 电影演员的 Bacon 数（截至 2010 年 5 月）

Bacon 数	演员数	Bacon 数	演员数	Bacon 数	演员数
0	1	3	719 767	6	1 040
1	2 251	4	178 784	7	165
2	225 506	5	12 205	8	17

在数学界也有类似的游戏，他们建立关联的方式是看他们是否合作发表过论文。其中当代最伟大的数学家之一 Erdös 成为了数学界的中心。比如说证明 Fermat 大定理的 Andrew Wiles，他的研究方向与 Erdös 相去甚远，但他的 Erdös 数只有 3，是通过这个途径实现的：Erdös—Andrew Odlyzko—Chris M. Skinner—Andrew Wiles。借助迅猛发展的 Internet，瓦茨领导的研究小组在 2003 年又发表一个实验报告。他们在全世界范围内检验了上述惊人的“六度分离”假说，有 6 万多志愿者参与利用电子邮件通信实验，确实不到 6 步就实现了他们的假设，从而利用 Internet 初步验证了复杂网络的小世界效应。

2. 社会网络中弱连接优势的发现

在 20 世纪 60 年代晚期，哈佛大学 Granovetter 通过寻访麻省牛顿镇的居民如何找

工作来研究社会网络。他非常惊讶地发现那些紧密的朋友反倒没有那些关系一般的朋友甚至只有一面之缘的朋友更能发挥作用。事实上，紧密的朋友根本帮不上忙。Granovetter 撰写的题为《弱连接的强度》的论文被当年的《美国社会学评论》拒之门外而无人问津，直到 1973 年之后才得到认可，并被认为是现代社会学最有影响力的论文之一（发表在“The American Journal of Sociology”的第 78 卷第六期）。Granovetter 指出（见百度百科词条“弱连接”）：在传统社会，每个人接触最频繁的是自己的亲人、同学、朋友、同事，这是一种十分稳定的然而传播范围有限的社会关系，是一种“强连接”（strong ties）；同时，还存在另外一类相对于前一种社会关系更为广泛的、联系却很少的社会关系，例如一个在朋友聚会中认识的人或者打开收音机偶然听到的一个人，这是一种“弱连接”（weak ties）。Granovetter 的弱连接优势理论指出：与一个人的工作和事业关系最密切的社会关系并不是“强连接”，而常常是“弱连接”。“弱连接”虽然不如“强连接”那样坚固，却有着极快的、可能具有低成本和高效能的传播效率。而在强连接关系下，成员彼此之间具有相似的态度，他们高度的互动频率通常会强化原本认知的观点而降低了与其他观点的融合，故强连接网络通常不能提供创新机会。相对于强连接关系，弱连接则能够在不同的团体间传递非冗余性的讯息，使得网络中的成员能够增加修正原先观点的机会。因此，拥有更多弱连接的人拥有信息流通的优势，往往可以得到更多的工作机会和业务选择机会。

3. 无标度性质的发现

紧接小世界效应和弱连接优势之后的另一个重要发现是：无标度性质。1998 年，Barabási 等开展一项描绘 WWW 的研究，他们原本以为会发现一个随机网络的钟形图，但他们意外地发现：WWW 基本上是由少数高连通性的页面串联起来的，80％以上页面的连接数不到 4 个，而占节点总数不到万分之一的极少数节点，却和 1000 个以上的节点连接，随机网络所具有的有特征意义（多数节点有大致相同的连接数）的连接数——“平均数”不见了。根据这一发现，1999 年他们在“Science”上发表了题为《随机网络中标度的涌现》的论文[6]，提出了一个无标度网络模型，指出在复杂网络中节点的度分布具有幂指数函数的规律（节点的度是指与该节点连接的边数，而度分布是指网络中所有节点的度的分布情况），其度分布可以用幂律形式 $P(k)\propto k^{-\gamma}$进行描述。在一个度分布为具有适当幂指数的幂律形式的大规模无标度网络中，绝大部分的节点的度相对较低，而存在少量的度相对很高的节点，因此这类网络也成为非均匀网络。经过许多的实证研究发现，大量复杂系统，诸如 Internet、细胞代谢系统以及好莱坞的演员合演网络都存在这种少数但高连通的遵循幂律分布的节点，可称为“集散节点”。许多不同的复杂系统，其网络结构都是无标度网络，都是由少数集散节点主控的系统。

4. 复杂网络研究的新时代

ER 随机图理论在提出之后一直对复杂网络研究具有重大影响，直到 20 世纪末随着“Nature”和“Science”上两篇开创性论文的发表，才宣告复杂网络研究新纪元的到来，并树起了网络科学的第三个里程碑。现在人们已经认识到：规则网络和随机网

络是两种极端的情况，对于大量真实的网络系统而言，它们既不是规则网络也不是随机网络，而是介于两者之间的某种网络。复杂网络研究在过去 10 年才得到迅速发展，其原因有以下几个方面[7]：①计算机技术的迅猛发展。近些年来，由于计算机技术的发展才使我们有可能获得各种网络的数据库和各种大规模网络的统计性质，并有可能对大规模的网络进行实证研究。②普适性的发现。实证分析表明，从 WWW 到新陈代谢网，许多领域的各种复杂网络展现了某些共同的统计性质，如幂律度分布，表明其中存在一些普适性的概念和规律，而先前的理论已经无法解释这些性质。③理论研究也有了突破，主要表现在：Watts 和 Strogatz 给出了小世界网络的构造方式，Barabási 和 Albert 则指出，增长和偏好连接是形成无标度网络的根本原因，而且统计物理学的研究方法也在复杂网络研究中得到广泛应用。理论研究和实证分析的相互促进在复杂网络的研究中得到了充分体现。2006 年，由 Barabási 等人编著的《网络的结构与动力学》专著，在国际上产生了广泛而深刻的影响。由于 Barabási 在网络科学方面的杰出贡献，他于 2006 年获得了美国冯·诺依曼计算机金奖。此后，复杂网络的文章铺天盖地，有关复杂网络的各种形式的学术研讨和会议越来越多，复杂网络的综述和专著不断涌现，从物理学到生物学，从社会科学到技术网络，从工程技术到经济管理等众多领域，受到了人们的空前关注和广泛重视。

1.3 复杂网络的概念和特性

1.3.1 复杂网络的概念

1. 系统和网络

系统（见百度百科词条“系统”）是由相互作用和相互依赖的若干组成部分结合的具有特定功能的有机整体。系统是普遍存在的，从基本粒子到河外星系，从人类社会到人的思维，从无机界到有机界，从自然科学到社会科学，系统无所不在。按宏观层面分类，它大致可以分为自然系统、人工系统、复合系统。自然系统内的个体按自然法则存在或演变，产生或形成一种群体的自然现象与特征，如生态平衡系统、生命机体系统、天体系统以及社会系统等。人工系统内的个体根据人为的、预先编排好的规则或计划好的方向运作，以实现或完成系统内各个体不能单独实现的功能、性能与结果，如生产系统、电力系统、教育系统、医疗系统等。复合系统是自然系统和人工系统的组合，如导航系统、交通管理系统等。我们可以从三个方面理解系统的概念：①系统是由若干要素（部分）组成的。这些要素可能是一些个体、元件、零件，也可能其本身就是一个系统（或称之为子系统）。②系统有一定的结构。一个系统是其构成要素的集合，这些要素相互联系、相互制约。系统内部各要素之间相对稳定的联系方式、组织秩序及失控关系的内在表现形式，就是系统的结构。③系统有一定的功能，或者说系统要有一定的目的性。系统的功能是指系统与外部环境相互联系和相互作用中表现出来的性质、能力和功能。系统有如下的属性：集合性、相关性、层次性、整体性、涌现性、系统对环境的适应性。系统是作为一个整体出现并且作为整体存在于

环境之中、与环境发生相互作用的，系统的任何组成要素或者局部都不能离开整体去研究。系统的局部问题必须放在系统的全局之中才能有效地解决，而系统的全局问题必须放在系统的环境之中才能有效地解决。系统的功能和特性，必须从系统的整体或总体来加以理解和要求，使之实现并且优化。系统的整体观念是系统概念的精髓，但系统整体的性质和功能不等于各个要素的性质和功能的简单加和，而是整体独有的、孤立的部分及其总和不具有的特性，我们称之为整体的涌现性。

在汉语中，“网络”一词最早用于电学，指的是电路或电路的一部分。《现代汉语典》(1993 年版）做出这样的解释：“在电的系统中，由若干元件组成的用来使电信号按一定要求传输的电路或这种电路的部分，叫网络。”但在这里，我们是从图论意义上理解网络的，也就是说，网络是指由节点和连线构成的图。有时用带箭头的连线表示从一个节点到另一个节点存在的某种顺序关系。有时在节点或连线旁标出数值，称为点权或线权，有时不标任何数。网络除了数学定义外，还有具体的物理含义，即网络是从某种相同类型的实际问题中抽象出来的模型，习惯上就称其为什么类型网络，如开关网络、运输网络、通信网络、计划网络等。总之，网络是从同类问题中抽象出来的用数学中的图论来表达并研究的一种模型。网络和系统通常是密切联系的，如果用节点表示系统的各个组成部分即系统的元素，两个节点之间的连线表示系统元素之间的相互作用，那么网络就为研究系统提供了一种新的描述方式。网络可以用来描述人与人之间的社会关系，物种之间的捕食关系，词与词之间的语义联系，计算机之间的网络连接，网页之间的超链接，科研文章之间的引用关系，以及科学家之间的合作关系，甚至产品的生产与被生产关系[8]。网络还可以作为现象的背景舞台，例如在社会关系网络上讨论舆论的传播，在科学家网络上研究科学家之间的相互影响等。网络与现象的结合还可以用来讨论网络的稳定性等结构与功能关系，例如在食物链网络上讨论个别或部分物种灭绝对整体生态系统的影响，在不同的网络上讨论传染病传播的控制。此外，网络本身的演化过程也是一个有趣问题，例如 Internet 网络的形成被认为是无限定原则的，但是它却展现了一些重要而普适的结构特征与稳定性。

2. 复杂性

在汉语中，“复杂”一词是由“复”和“杂”两个字组合而成的[9]。“复”的主要含义指多样、重复、反复，形成某种层次嵌套的自相似结构。“杂”的主要含义指多样、破碎、纷乱，形成某种不规则的、无序的结构。但是“复而不杂”和“杂而不复”还不是真正的完全的复杂性，只有“既复且杂”才是真正的完全的复杂性，它把层次嵌套的自相似性与无规则性、破碎性、混乱性有机地结合起来。这种事物的部分与整体之间既是相似的、又不严格相似的对象之所以出现，是因为在反复迭代（即生成演化）过程中不时有随机因素侵入，但又是不可预料的，才导致严格自相似性的破缺，因而不能用确定论的方法描述。这种对象也不能用统计方法描述，因为它们的生成演化过程毕竟有某些规则在不断重复，具有明显的尺度（层次）变换下的不变性，即规律性。

复杂性涉及面很宽，从国内外自然科学、工程技术科学、管理科学和人文社会科

学等领域关于“复杂性”、“非线性”的研究状况来看，复杂性是涉及不同学科领域的共同问题。在不同学科领域中，研究对象和方法不同，因而对复杂性概念的定义也不同。复杂性是建立在多样性、差异性之上的，我们应当承认不同意义上的复杂性，承认不同层次有不同的复杂性，允许使用不同的复杂性定义。据劳埃德统计，西方学者已提出 45 种复杂性定义。总的来看，复杂性还不能算作一个严格的科学概念，人们也没有给出一个公认的复杂性定义。

复杂性是相对于简单性而存在的，它是在客观事物的联系、运动和变化中表现出来的一种状态，表达了一种不可还原的特征，而不是孤立、静止和显而易见的特性。复杂性科学打破了线性、均衡、简单还原的传统范式，极大地促进了科学的发展。

3. 复杂系统

随着科学技术的发展，人类社会出现了许多大型、复杂的工程技术和社会经济问题，它们都是以系统的面貌出现，客观地要求从整体上加以优化解决[9]。因此，系统科学的出现是历史的必然。自组织是系统科学的重要概念，它是复杂系统演化时出现的一种现象。复杂系统是相对牛顿时代以来构成科学研究重点的简单系统相比较而言的，它是复杂性科学研究的基本对象，它与简单系统的最大区别在于系统的整体涌现性。目前关于复杂系统的定义还很不统一，但是对于复杂系统的基本特性的认识却比较一致。一般认为复杂系统具有以下特征：非线性与动态性、非周期性和开放性、奇怪吸引性、结构自相似性（分形）。另外，复杂系统还具有突现性、不稳性、不确定性、不可预测性等特征。值得注意的是，1999 年 4 月美国“Science”杂志出版了《复杂系统》的专辑，在其以“超越还原论”为标题的导言中，对其所指的“复杂系统”作了如下简单描述：通过对一个系统的分量部分（子系统）性能的了解，不能对系统的性能作出完全的解释，这样的系统称为“复杂系统”。

4. 复杂网络

在自然界和人类社会中，个体总是和周围的环境紧密联系的，无论是否情愿，个体总是或多或少被环境所影响，也不停地影响着环境。所以，对系统的分析不仅要分析个体自身的内容和特征，其他个体对它的影响也必须关注，剥离环境孤立的分析已不合时宜，而应该采用整体的系统的分析方法。复杂网络是一种很好地描述自然科学、社会和科学工程技术上的相互关联的系统的模型，它应用了数学上图的概念：复杂网络可以看作由一些具有独立特征的又与其他个体相互连接的节点的集合，每个个体可视为图中一个节点，节点间的相互连接视为图中的边。复杂网络包括两个层面：作为其连接拓扑结构的图和作为其状态和功能的系统。顾名思义，复杂网络就是呈现高度复杂性的网络。钱学森给出了复杂网络的一个较严格的定义：具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络称为复杂网络。原则上说，任何包含大量组成单元（或子系统）的复杂系统，当我们把构成单元抽象成节点，单元之间的相互作用抽象为边时，都可以当作复杂网络来研究。复杂网络可以用来描述物种之

间的捕食关系，人与人之间的社会关系，词与词之间的语义联系，计算机之间的网络链接，神经元之间的通信反馈作用，蛋白质之间的相互关系等。所以，复杂网络为研究复杂系统提供了一种新的描述方式，可以加深我们对系统结构的深入了解；反过来，复杂网络的研究成果对探索复杂性又具有一定的启发和借鉴意义。

1.3.2 复杂网络的特性

复杂网络的优美结构和新奇的规律，越来越吸引着人们去探索更多的奥妙。复杂网络第一个显而易见的特性是其复杂性，概括地说，绝大多数实际的复杂网络的复杂性主要表现在以下几个方面[7,10]：

1）网络规模庞大。网络节点数可以有成百上千万，甚至更多，但大规模性的网络行为具有统计特性。

2）连接结构的复杂性。网络连接结构既非完全规则也非完全随机，但却具有其内在的自组织规律，网络结构可呈现多种不同特征，如图 1.3 所示。网络连接结构也可能随时变化，表现在节点或连接的产生与消失，例如 WWW 网页或链接随时可能出现或断开，导致网络结构不断发生变化。此外，节点之间的连接权重可能存在差异，且有可能存在方向性。如神经网络中的突触有强有弱，可以是抑制的也可以是兴奋的。

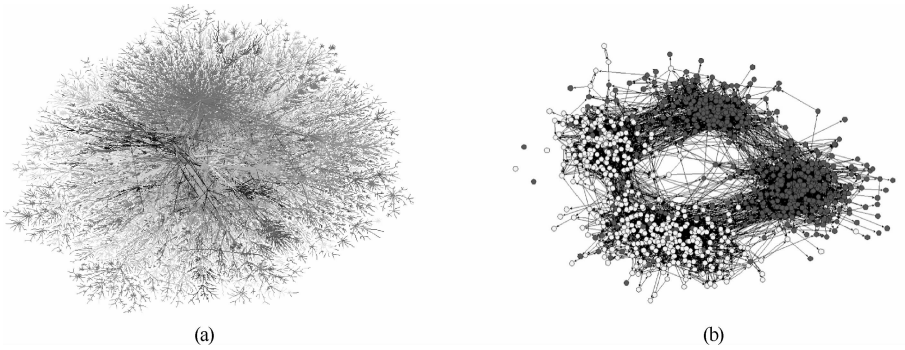

图 1.3 两个典型的实际复杂网络结构示意图

（a）部分 IP 地址连接示意图；（b）朋友关系网

3）节点的复杂性。首先表现在节点的动力学复杂性，即各个节点本身可以是各非线性系统（可以有离散的和连续微分方程描述），具有分岔和混沌等非线性动力学行为。例如，节点状态随时间发生复杂变化；其次表现在节点的多样性，复杂网络中的节点可以代表任何事物，而且一个复杂网络中可能出现各种不同类型的节点。例如，人际关系构成的复杂网络节点代表单独个体，WWW 组成的复杂网络节点可以表示不同网页，一个生物网络通常包含各种不同性质的酶和基质。

4）网络时空演化过程复杂。复杂网络具有空间和时间的演化复杂性，展示出丰富的复杂行为，特别是网络节点之间的不同类型的同步化运动（包括出现周期、非周期、混沌和阵发行为等运动）。

5）网络连接的稀疏性：一个有 N 个节点的具有全局耦合结构的网络的连接数目为

$O(N^2)$，而实际大型网络的连接数目通常为 $O(N)$。

6）多重复杂性融合：若以上多重复杂性相互影响，将导致更为难以预料的结果。例如，设计一个电力供应网络需要考虑此网络的进化过程，其进化过程决定网络的拓扑结构。当两个节点之间频繁进行能量传输时，它们之间的连接权重会随之增加，通过不断的学习与记忆逐步改善网络性能。

除了复杂性，复杂网络一般还具有以下三个特性[11]：

1）小世界特性。大多数网络尽管规模很大，但任意两个节点间却有一条相当短的路径。简单地说，单个节点拥有的相互关系的数目可以很小但却能够连接世界的事实，例如，在社会网络中，人与人相互认识的关系很少，但是却可以找到很远的无关系的其他人。正如麦克卢汉所说，地球变得越来越小，变成一个地球村，也就是说，变成一个小世界。

2）无标度特性。人们发现一些复杂网络（如演员合作网、WWW 和电力网）的节点的度分布具有幂指数函数的规律。因为幂指数函数在双对数坐标中是一条直线，这个分布与系统特征长度无关，所以该特性被称为无标度性质。无标度特性反映了网络中度分布的不均匀性，只有很少数的节点与其他节点有很多的连接，成为“中心节点”，而大多数节点度很小。

3）超家族特性。2004 年 Sheffer 和 Alon 等在“Science”上发表文章[12]，比较了许多已有网络的局部结构和拓扑特性，观察到有一些不同类型的网络的特性在一定条件下具有相似性。尽管网络不同，只要组成网络的基本单元（最小子图）相同，它们的拓扑性质的重大轮廓外形就可能具有相似性，这种现象被他们称为超家族特性。顾名思义，不同网络之间存在与某个家族的“血缘”相近联系，而出现与该家族相似的特性，究其原因在于它们拥有相同的或相似的网络“基因”，但问题是网络“基因”是不是找准了？是否存在网络“基因”排序等更深层次的问题。目前，对于超家族特性在研究理论方法和技术上都有待进一步改进和发展，需要更多的不同网络的实证研究和严格的理论证明。

综上所述，复杂非线性动态网络不仅节点具有非线性和复杂性，其内容和形式可以多种多样，而且复杂动态网络系统的连接结构和时空演化更是错综复杂、丰富多彩。这就向复杂性科学、非线性动力学和复杂网络理论等交叉科学提出了一系列极富挑战性的新课题。特别需要强调的是：复杂动态网络系统时空演化中出现的复杂性，尤其是各类同步问题，包括广义同步的物理机制及其控制方法等问题，是迄今尚未解决的一类难题，也是众多领域中都存在的值得共同研究的复杂性课题。这些研究将在复杂网络系统中占有头等重要的位置。

1.4 数理统计基础

为了便于对本书内容的理解，从本节开始的两节介绍一些预备知识，分别是数理统计基础[13]和图论[14]的基本概念。本节的主要目的是概括性地介绍概率论基础、数理统计基础、假设检验和回归分析这四方面的基础知识。

1.4.1 概率论基础

1. 随机试验

在自然界和人类生活中普遍存在着两类现象[13]：一类是在一定条件下必然出现的现象，称为确定性现象；另一类是事先无法准确预知结果的现象，称为随机现象。虽然随机现象的每次结果无法预知，但当随机现象大量出现时，每种可能的结果出现的频率却具有稳定性，通常把随机现象在大量重复出现时所表现出来的规律性称为随机现象的统计规律性。对随机现象的统计规律性进行的观察称为随机试验（简称试验），一个随机试验有以下特点：①可重复性：试验原则上可在相同条件下重复进行；②可观察性：试验结果是可观察的，所有可能的结果是明确的；③随机性：每次试验将要出现的结果是不确定的，事先无法准确预知。

2. 样本空间、随机事件

一个随机试验的任何一个不可再分解的结果称为一个样本点，所有样本点的集合称该试验的样本空间，记为 $\Omega=(\omega_1,\omega_2,\cdots,\omega_n)$。

随机试验的样本空间 Ω 的子集称为该试验的随机事件，简称为事件。在每次试验中，当且仅当该子集中的一个样本点出现时，称该随机事件发生了。样本点本身不能作为随机事件，只能作为组成子集的元素。样本空间 Ω 本身也是 Ω 的一个事件，称为“必然事件”，而不包含样本点的空集 $\varnothing$ 称为“不可能事件”。

事件也是一个集合，因此事件间的关系及运算可以按照集合论中集合之间的关系及运算来处理。假设 A、B、C 各表示一个事件，它们之间的各种关系分别表示如下：

1）$A\supset B$ 表示 B 发生则 A 必发生，称 B 为 A 的子事件；

2）$A=B$ 表示 A 发生则 B 必发生，且 B 发生则 A 必发生，称 A 与 B 相等；

3）$A\cup B$ 表示 A、B 两个事件中至少有一个事件发生，也可理解为 A 发生或 B 发生；

4）$A\cap B$ 表示 A、B 两个事件同时发生，可简记为 AB；

5）$B-A$ 表示 B 发生 A 不发生，称为 B 与 A 之间的差事件，若 $B=\Omega$，则 $B-A$ 可用 $\overline{A}$ 表示，称为 A 的对立事件；

6）$A\cap B=\varnothing$ 表示 A、B 两个事件不可能同时发生，称为互斥事件或互不相容事件；

7）$A\cap B=\varnothing$ 且 $A\cup B=\Omega$ 表示 A、B 两个事件互为对立事件，称 A、B 为互逆事件，或 B 是 A 的逆事件，即 $B=\overline{A}$。

3. 频率、概率

频率描述了事件发生的频繁程度，它的定义为：在相同的条件下，进行了 n 次试验，在这 n 次试验中，事件 A 发生的次数 n_A 称为事件 A 发生的频数，比值 n_A/n 称为事件 A 发生的频率，记为 $f_n(A)$。由于频率值的大小表示 A 发生的频繁程度，因此频

率值越大，事件 A 发生就越频繁。

频率描述是某件事件基于多次重复试验得出的观察结果，而概率表征该事件在一次试验中发生的可能性大小的数值。假设 E 是随机试验，Ω 是该试验的样本空间。假设 P 是以事件为自变量的实值集合函数，如果 P 还满足：

1）非负性：对任何事件 A，有 $P(A)\geqslant 0$；

2）规范性：对于必然事件 Ω，有 $P(\Omega)=1$；

3）可列可加性：假设 $A_iA_j=\varnothing$（$i\neq j$，i，$j=1$，2，…），事件间互不相容，有

$$P\Big(\sum_{i=1}^{\infty}A_i\Big)=\sum_{i=1}^{\infty}P(A_i) \tag{1.1}$$

就称 P 为概率。概率的统计定义可以表示为：当不断重复同一试验时，事件 A 发生的频率的极限称为事件 A 的概率。

概率具有以下性质：

1）对任何事件 A，有 $1\geqslant P(A)\geqslant 0$，且 $P(\varnothing)=0$；

2）若 A、B 为互斥事件，则 $P(A\cup B)=P(A)+P(B)$；

3）$P(\overline{A})=1-P(A)$。

4. 古典概型与几何概型

如果一种随机试验只有有限个样本点，而且每个基本结果的出现都是等可能的，这类随机试验模型称为古典概率模型。古典型事件的概率可以由 $P(A)=k/n$ 直接计算得出，其中 n 为样本空间的样本点总数，k 为事件 A 包含的样本点数。

若随机试验有无穷不可数个样本点，但每个样本点在一次试验中出现的机会是均等的，这类随机试验模型称为几何概型。可由 $P(A)=S_A/S_\Omega$ 计算得出，其中 S_Ω 为样本空间的几何度量值，S_A 为事件 A 的几何度量值（几何度量值一般指长度、面积、体积等）。

5. 条件概率、全概率公式、贝叶斯公式、独立事件判定

条件概率是在一个事件已经发生的情况下另一个事件发生的概率。假设 A、B 是两个事件，且 $P(A)>0$，则 $P(B\mid A)=P(AB)/P(A)$ 称为 A 发生的条件下 B 发生的条件概率。假设试验 E 的样本空间为 Ω，B_1，B_2，…，B_n 为 E 的一组事件。若 $B_iB_j=\varnothing$，（$i\neq j$；i，$j=1$，2，…，n），且 $B_1\cup B_2\cup\cdots\cup B_n=\Omega$，则称 B_1，B_2，…，B_n 为样本空间 Ω 的一个划分。由此可以引出如下全概率公式：

$$P(A)=\sum_{i=1}^{n}P(A\mid B_i)P(B_i) \tag{1.2}$$

以及贝叶斯公式：

$$P(B_i\mid A)=\frac{P(A\mid B_i)P(B_i)}{\sum_{j=1}^{n}P(A\mid B_j)P(B_j)} \tag{1.3}$$

假设 A_1，A_2，…，A_n 是 n 个事件，若对于任意 $k(1<k\leqslant n)$，均满足等式

$$P(A_1A_2\cdots A_k)=P(A_1)P(A_2)\cdots P(A_k) \tag{1.4}$$

则称 A_1，A_2，…，A_n 为相互独立的事件。

6. 随机变量和随机向量

设 $X(\omega)$为定义于样本空间 Ω 上的单值实函数，若对于任意实数 x，使 $X(\omega)<x$ 成立的样本点组成的集合均是事件，则称 $X(\omega)$ 为一个随机变量，并简称 $X(\omega)$ 为 X，简记 $\{\omega: X(\omega)<x, \omega\in\Omega\}$ 为 $\{X<x\}$。随机变量的取值由试验的结果所决定，X 随着试验的不同结果而取不同的值。若 X_1，X_2，…，X_n 是样本空间 Ω 上的 n 个随机变量，则称 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)^{\mathrm{T}}$ 构成一个 n 维随机向量，称之为 n 维随机变量。常见的随机变量有两类：离散型随机变量和连续型随机变量。

对于离散型随机变量，设 x_1，x_2，…，x_n 是 n 个实数，则称 n 元函数

$$F(x_1, x_2, \cdots, x_n) = P\{X_1 \leqslant x_1, X_2 \leqslant x_2, \cdots, X_n \leqslant x_n\} \tag{1.5}$$

为随机向量 $\boldsymbol{X}=(X_1, X_2, \cdots, X_n)^{\mathrm{T}}$ 的联合概率分布函数。假如已知离散型随机向量 $\boldsymbol{X}$ 的分布律，则可求出该随机试验的任何事件出现的概率。

对于连续型随机向量 $\boldsymbol{X}$，若存在非负可积函数 $f(x_1, x_2, \cdots, x_n)$，使

$$F(x_1, x_2, \cdots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, y_2, \cdots, y_n) \mathrm{d}y_1 \mathrm{d}y_2 \cdots \mathrm{d}y_n \tag{1.6}$$

则称 $f(x_1, x_2, \cdots, x_n)$ 为 $\boldsymbol{X}$ 的概率密度函数，简称概率密度。

设 $F(x_1, x_2, \cdots, x_n)$ 是一个随机试验的 n 元概率分布函数，任意保留 k 个 $x_i(1\leqslant k\leqslant n)$，而使其他的 x_j 都趋向于$+\infty$，则 k 元函数

$$F(x_1, x_2, \cdots, x_k, +\infty, \cdots, +\infty) = \lim_{x_{k+1}, \cdots, x_n \to \infty} F(x_1, x_2, \cdots, x_n) \tag{1.7}$$

称为 F 的 k 元边缘概率分布函数。

7. 随机变量的数字特征

一旦知道了随机变量的分布，计算所有与其相关的事件的概率都是不难的。但是在处理许多实际问题时，却并不要求将随机变量所有的概率进行计算，仅用几个能够反映该随机变量主要特征的参数就足够了。常用来表征随机变量主要特征的参数，又称为随机变量的数字特征，包括数学期望、方差、标准差、协方差和相关系数。

随机变量的数学期望是随机变量以其概率为权的加权平均值，也可以说是随机变量取值在加权意义下的重心。对于一个确定的随机变量，数学期望是一个确定的数值。

若 X 是离散型随机变量，$P\{X=x_i\}=p_i(i=1, 2, \cdots, n)$ 为其已知的分布律，称

$$\mathrm{E}(x) = \sum_{i=1}^{n} x_i p_i \tag{1.8}$$

为 X 的数学期望，简称期望，记为 $\mathrm{E}(x)$ 或 $\mathrm{E}X$。

若 X 是连续型随机变量，$\varphi(x)$为其已知密度，如果 $\int_{-\infty}^{+\infty} |x| \varphi(x)\mathrm{d}x$ 可积，则称

$$\mathrm{E}(x) = \int_{-\infty}^{+\infty} x\varphi(x)\mathrm{d}x \tag{1.9}$$

为 X 的数学期望。

随机变量的方差是随机变量与其重心的偏差平方的平均值，这个数值反映了随机变量的分散程度。方差越大，表明该随机变量取值越分散。

若 X 是离散型随机变量，$P\{X=x_i\}=p_i(i=1,2,\cdots,n)$ 为其已知的分布律，称

$$\mathrm{D}(x)=\sum_{i=1}^{n}(x_i-\mathrm{E}X)^2p_i \tag{1.10}$$

为 X 的方差，记为 $\mathrm{D}(x)$ 或 $\mathrm{D}X$。

若 X 是连续型随机变量，$\varphi(x)$为其已知密度，如果 $\int_{-\infty}^{+\infty}[x-\mathrm{E}(x)]^2\varphi(x)\mathrm{d}x$ 可积，则称

$$\mathrm{D}(x)=\int_{-\infty}^{+\infty}(x-\mathrm{E}\,X)^2\varphi(x)\mathrm{d}x \tag{1.11}$$

为 X 的方差。方差的算术平方根称为标准差，即

$$\sigma(x)=\sqrt{\mathrm{D}X} \tag{1.12}$$

设 X 和 Y 是两个随机变量，如果

$$\mathrm{COV}(X,Y)=\mathrm{E}[(X-\mathrm{E}X)(Y-\mathrm{E}Y)] \tag{1.13}$$

存在，则上式称为 X 和 Y 的协方差。协方差 $\mathrm{COV}(X,Y)$ 的绝对值大小反映了相关的密切程度。当 X 和 Y 相互独立时，$\mathrm{COV}(X,Y)=0$。

随机变量 X 和 Y 之间的相关系数定义为

$$\rho(X,Y)=\frac{\mathrm{COV}(X,Y)}{\sqrt{\mathrm{D}X}\sqrt{\mathrm{D}Y}} \tag{1.14}$$

显然，$0\leqslant|\rho(X,Y)|\leqslant1$，相关系数绝对值的大小直接反映了 X 和 Y 线性相关的程度。当 $|\rho(X,Y)|=1$ 时，X 和 Y 有完全的线性关系。当 $\rho(X,Y)=1$ 时，有 $Y=aX+b$ $(a>0)$；当 $\rho(X,Y)=-1$ 时，有 $Y=aX+b$ $(a<0)$。当 $|\rho(X,Y)|<1$，表明 X 和 Y 存在某种程度的线性关系。

8. 大数定理、中心极限定理

如果仅仅知道随机变量的期望和方差，可以用下面的不等式对随机变量的概率分布提供大致估计。

假设随机变量 X 的期望为 $\mathrm{E}X$，方差为 $\mathrm{D}X$，则契比雪夫不等式

$$P[|X-\mathrm{E}X|\geqslant\varepsilon]\leqslant\frac{\mathrm{D}X}{\varepsilon^2} \tag{1.15}$$

成立，其中 ε 是任意给定的一个小的正数。契比雪夫不等式显示了随机变量的各种取值和期望的差与随机变量的方差之间的关系。这个不等式以数量化方式来描述，究竟有多少比例的变量取值会以多大的差值接近期望。比如，与期望相差 2 个标准差的值，数目不多于 1/4；与期望相差 3 个标准差的值，数目不多于 1/9；…；与期望相差 k 个标准差的值，数目不多于 $1/k^2$。举例说，若一班有 36 个学生，而在一次考试中，平均分是 80 分，标准差是 10 分，我们便可得出结论：少于 50 分（与平均分相差 3 个标准差以上）的人，数目不多于 36/9=4 个。

大数定理揭示了大量重复试验结果的平均值的统计特征。根据大数定理，大量重复试验出现的结果的平均值却几乎总是接近于某个确定值，即大量随机变量的平均值表现出了可以预测的规律性。设 X_1，X_2，…，X_n，…为随机变量序列，令 $Y_n=\frac{1}{n}\sum_{i=1}^{n}X_i$，$n=1$，2，…，若存在常数列 a_1，a_2，…，a_n，…使得对任意的 $\varepsilon>0$ 均有

$$\lim_{n\to\infty}P[\,|\,Y_n-a_n\,|<\varepsilon]=1 \tag{1.16}$$

则称随机变量序列服从大数定律。下面介绍三个著名的大数定理。

（1）契比雪夫大数定理

设 X_1，X_2，…，X_n，…相互独立，它们的期望、方差都存在且方差一致有界，即 $E(X_i)=\mu_i$，$D(X_i)=\sigma_i^2\leqslant C$（常数），$i=1$，2，…，则对任意的 $\varepsilon>0$ 均有

$$\lim_{n\to\infty}P[\,|\,Y_n-\mathrm{E}(Y_n)\,|<\varepsilon]=1 \tag{1.17}$$

换言之，在该定理条件下，当 n 无限变大时，n 个随机变量的算术平均将变成一个常数。

（2）伯努利大数定理

设在伯努利试验中，事件 A 发生的概率均为 p，m 为 n 重伯努利试验中事件 A 发生的次数，则对任意的 $\varepsilon>0$ 均有

$$\lim_{n\to\infty}P\left[\left|\frac{m}{n}-p\right|<\varepsilon\right]=1 \tag{1.18}$$

该定理表明事件发生的频率依概率收敛于事件的概率。该定理以严格的数学形式表达了频率的稳定性。就是说当 n 很大时，事件发生的频率与概率有较大偏差的可能性很小。

（3）辛欣大数定理

若 X_1，X_2，…，X_n，…相互独立又服从相同的分布，且它们的期望都存在，$\mathrm{E}(X_i)=\mu$，$i=1$，2，…，则对任意的 $\varepsilon>0$ 均有

$$\lim_{n\to\infty}P\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right|<\varepsilon\right]=1 \tag{1.19}$$

换言之，在该定理条件下，当 n 无限变大时，n 个独立同分布随机变量的算术平均将会依概率收敛于随机变量的期望值。

中心极限定理描述了一列相互独立的随机变量 X_1，X_2，…，X_n，…的和 $Z_n=\sum_{i=1}^{n}X_i$ 的极限分布情况。在随机变量满足一定条件的情况下，一列随机变量的和的极限分布为正态分布。最著名的两个中心极限定理如下。

（1）列维-林德伯格中心极限定理

设 X_1，X_2，…，X_n，…是独立同分布的随机变量序列，且 $\mathrm{E}(X_i)=\mathrm{E}X$，$\mathrm{D}(X_i)=\mathrm{D}X>0$，（$i=1$，2，…）均存在，则对任意的 $x\in\mathbf{R}$ 有

$$\lim_{n\to\infty}P\left[\frac{\frac{1}{n}\sum_{i=1}^{n}X_i-\mathrm{E}X}{\sqrt{\frac{\mathrm{D}X}{n}}}<x\right]=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}x=\Phi(x) \tag{1.20}$$

其中，$\Phi(x)$表示标准正态分布的分布函数。该独立同分布的中心极限定理表明：无论

$\{X_i\}$是服从什么分布的随机变量序列，当 n 趋近于无穷大时，随机变量 Z_n 都服从正态分布，而随机变量 Z_n 标准化后都趋于标准正态分布。需要强调的是：上述中心极限定理对随机变量的形状和性质没有基本要求，只要它们的期望和方差存在并且有限即可；无论是连续随机变量还是离散随机变量或者混合随机变量，一列独立的随机变量的和近似满足正态分布。

（2）棣莫弗-拉普拉斯中心极限定理

棣莫弗-拉普拉斯中心极限定理实际上是列维-林德伯格中心极限定理针对贝努利分布的特例。设 X_1，X_2，…，X_n，…独立同分布，均服从参数为 $0<p<1$ 的伯努利分布 B（1，p），则 $Z_n=\sum_{i=1}^{n} X_i$ 在 $n\to\infty$时满足正态分布，即

$$\lim_{n\to\infty} P\left[\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - p}{\sqrt{\frac{p(1-p)}{n}}} < x\right] = \lim_{n\to\infty} P\left[\frac{Z_n - np}{\sqrt{np(1-p)}} < x\right] = \Phi(x) \tag{1.21}$$

1.4.2 数理统计基础

1. 总体、样本

在数理统计中，许多时候需要研究有关对象的某一项数量指标（例如某个班级男生的身高指标），那就必须考虑与该指标相联系的随机试验，对该数量指标进行试验或者观察。通常将试验的全部可能的观察值称为总体，每一个可能的观察值称为个体。总体中包含个体的数目称为总体的容量，容量有限的总体称为有限总体，容量无限的总体称为无限总体。由于总体中的每个个体都是随机试验的一个观察值，同时它是某一个随机变量 X 的值，因此一个总体可看作一个随机变量 X。对总体的研究就可以看作是对一个随机变量 X 的研究，X 的分布函数和数字特征称为总体的分布函数和数字特征。总体和相应的随机变量可统称为总体 X。

在数理统计中，人们都是通过从总体中抽取一部分个体，根据获得的数据来对总体分布得出推断，通常将被抽出的部分个体称为总体的一个样本。假设 X 是具有分布函数 F 的随机变量，若 X_1，X_2，…，X_n 是具有同一分布函数 F 的、相互独立的随机变量，则 X_1，X_2，…，X_n 称为从分布函数 F（或总体 F、总体 X）得到的容量为 n 的简单随机样本，简称样本。它们的观察值 x_1，x_2，…，x_n 称为样本值，又称为 X 的 n 个独立观察值。

2. 统计量

设 X_1，X_2，…，X_n 是来自总体的一个样本，$g(X_1, X_2, \cdots, X_n)$ 是样本的某一函数。若 g 中不含未知参数，则 $g(X_1, X_2, \cdots, X_n)$ 称为一个统计量。常用的统计量有如下五种：

（1）样本平均值

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.22}$$