

复杂网络分析与应用

赵正旭　郭　阳　王　威
钟　谦　彭育贵　赵卫华　著

科学出版社

需要全本请访问 www.ertongbook.com

复杂网络分析与应用

赵正旭　郭　阳　王　威
钟　谦　彭育贵　赵卫华　著

学院图书馆
藏书专用章

科学出版社
北京

内 容 简 介

复杂网络是指具有高度复杂性的网络系统，其具有很深的物理学和数学理论基础。本书没有过多地引入相关的物理学理论解释和数学推导，而是通俗和系统地介绍复杂网络的基本概念、分析方法和应用领域。本书的重点是帮助读者了解复杂网络的分析方法，从而利用计算机工具来解决实际工程中的复杂网络技术问题。

本书适合作为工程技术类研究生和高年级本科生学习复杂网络技术及其应用的教材，也可供自然科学、工程技术科学和社会科学领域的研究人员参考。

图书在版编目(CIP)数据

复杂网络分析与应用/赵正旭等著. —北京：科学出版社，2018.11

ISBN 978-7-03-057570-8

Ⅰ. ①复… Ⅱ. ①赵… Ⅲ. ①计算机网络-网络分析-研究 Ⅳ. ①TP393.021

中国版本图书馆 CIP 数据核字（2018）第 112508 号

责任编辑：赵丽欣 张星 常晓敏 / 责任校对：赵丽杰
责任印制：吕春珉 / 封面设计：东方人华平面设计部

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

北京中科印刷有限公司印刷
科学出版社发行 各地新华书店经销
*
2018 年 11 月第 一 版 开本：787×1092 1/16
2018 年 11 月第一次印刷 印张：7 1/4 插页：2
字数：155 000

定价：59.00 元

（如有印装质量问题，我社负责调换〈中科〉）

销售部电话 010-62136230 编辑部电话 010-62134021

版权所有，侵权必究

举报电话：010-64030229；010-64034315；13501151303

前　言

复杂网络（complex networks）具有自组织、自相似、容错性等特点。随着数据科学和信息工程等新兴技术的兴起和发展，复杂网络理论分析与应用研究引起了诸多领域的广泛关注，并迅速形成了一门贯穿计算机科学与技术、控制理论与工程、生物科学、医学、材料科学、天文科学、社会科学、管理科学等领域的交叉学科，并且极大地推动了这些相关学科的理论研究、方法设计、系统架构、应用推广等多个方面的发展与深化。特别是在网络信息技术背景下的大数据时代，复杂网络的研究和发展已经超出了物理学和数学的范畴，各种复杂网络的分析和应用已经为诸多学科提出了新的具有挑战性的难题，如复杂网络基本理论、复杂网络在实际系统中的应用、复杂网络算法、信息组织与挖掘和预测、复杂网络的信息动力学、复杂网络的结构和演化等问题。为了给复杂网络相关学科和相关应用领域感兴趣的研究人员和工程技术人员提供具有参考价值的研究方向、分析方法、应用案例，本书系统地介绍了复杂网络的基本概念、分析方法和应用领域，其重点在于帮助读者了解复杂网络的分析方法，提高利用复杂网络来解决实际工程和技术问题的能力。为此，本书没有过多地引入相关的物理学理论解释和数学推导，而是更加注重对复杂网络的组织、结构及数据模型的分析。

本书的主要内容如下。

第 1 章主要介绍信息数据长期保存研究现状，以及复杂网络和本体的研究现状。

第 2 章介绍复杂网络的基本概念、典型特征和网络模型等，先讲解小世界效应、度分布和无标度性等特征，然后详述包括 ER 模型、WS 模型、BA 模型及权重网络模型在内的 4 种典型网络模型，并对现实世界中的复杂网络进行介绍。

第 3 章针对凭借网页之间形成的链接结构构成的超大规模的复杂网络，通过结点度、平均路径长度及集群系数等相关指标的计算对比，得出万维网链接结构网络符合幂律分布，属于无标度网络，同时具有小世界效应的结论。这表明万维网上的海量信息资源是由少量的以信息集散结点为核心的链接形式构成的，包含相似主题的信息资源组成若干小集群。

第 4 章采用复杂网络的方法来研究软件源程序中的函数调用关系。针对大型开源太空仿真软件源程序中的函数调用关系图进行了复杂网络的实例验证，发现其表现出小世界及无标度特性。

第 5 章通过对已收集到的软件的数据格式进行分析建模，得到初始的 WS 小世界网络模型。根据现实中存在的影响数据格式转换关系的部分因素，分别构建了软件生命周期模型、信息数据生命周期模型、数据格式生命周期模型和数据失真转换模型，最终得出上述模型属于小世界网络的结论。

第 6 章针对工程信息具有形式复杂和内容分散的特性，以及结构的关联性和数据的兼容性等特点，通过对工程信息进行概念体系的划分，定义了类、属性和实例之间的关

系。根据本体的构建语言、描述规则、构建方法及领域本体的构建理念，用 Protégé 构建了工程信息本体，局部实现了领域知识的共享和重用，为解决信息移存、转换的时间和效率问题提供了可靠的科学依据。

本书从万维网的链接结构、软件系统函数调用关系及其结构、数据格式转换关系模型、工程信息本体论分析 4 个方面详细介绍了实际工程和技术应用中典型复杂网络的分析方法和应用实例，为广大读者提供了一个全面、系统、详尽、通俗的学习读物和参考工具。书的结尾向读者提供了主要参考文献，以便于对复杂网络物理学和数学感兴趣的读者作进一步的探究。我们在撰写本书的过程中，对广大的复杂网络理论研究者及相关物理学和数学科学前辈深感敬佩。衷心希望读者为复杂网络的研究、发展、应用和普及贡献力量。

在本书临近付印之际，向国家自然科学基金项目“工程信息中小世界效应研究”（项目编号：60873208）、河北省首批百人计划项目（项目编号：410003）、河北省高等学校高层次人才科学研究项目（项目编号：GCC2014010）、河北省国防科技工业局军民融合产业专项“河北省国家重大国防可视化系列系统及软件产品的研发与实际推广应用”（冀财预复[2016]300 号）、石家庄铁道大学与钱学森空间技术实验室联合实验室建设项目（项目编号：Z67802）的资助和支持致以诚挚的谢意。

特别感谢北京航天飞行控制中心的领导和专家及同行过去十几年中在载人航天和深空探测实战工程任务中所给予的合作和关怀，特别是在复杂网络在航天工程信息化建设的应用及相关科研合作过程中所给予的支持和帮助。另外，还要感谢石家庄铁道大学复杂网络与可视化研究所的所有研究生和北斗卫星导航科技邢台有限公司石家庄铁道大学可视化技术研究院的全体员工。自 2009 年以来，他们怀着对复杂网络及其应用技术和实践技能的渴求，积极参与了研究所的各项科研和实践活动，参加并见证了复杂网络在研究所承担的十几次国家载人航天工程、探月工程、深空探测工程的三维可视化任务中的应用，其中包括 2009 年 10 月至 2010 年 8 月火星探测可视化、2010 年 9 月至 11 月嫦娥二号绕月探测飞行控制指挥任务、2011 年 6 月 9 日嫦娥二号驶向拉格朗日 L2 点的可视化飞行控制指挥任务、2011 年 9 月 29 日天宫一号的发射飞行控制指挥任务、2011 年 11 月 3 日天宫一号和神舟八号的首次交会对接指挥任务、2011 年 11 月 14 日至 15 日天宫一号和神舟八号的分离和第二次交会对接指挥任务、2011 年 11 月 16 日至 17 日神舟八号飞船返回指挥任务、2012 年 6 月 16 日至 29 日神舟九号载人航天可视化飞行控制与指挥任务、2012 年 12 月 13 日至 16 日嫦娥二号对 4179 小行星的探测过程的实时可视化、2013 年 6 月 11 日神舟十号载人航天飞船发射可视化飞行控制与指挥任务、2013 年 6 月 13 日天宫一号和神舟十号飞船的自动交会对接可视化飞行控制与指挥任务、2013 年 6 月 23 日天宫一号和神舟十号飞船的手动交会对接可视化飞行控制与指挥任务、2013 年 6 月 24 日神舟十号飞船对天宫一号的绕飞可视化飞行控制与指挥任务、2013 年 6 月 26 日神舟十号飞船返回过程的可视化飞行控制与指挥任务、2013 年 12 月 2 日至 15 日嫦娥三号发射-绕月-落月着陆互拍可视化飞行控制指挥任务、2013 年 12 月 15 日以后嫦娥三号和玉兔月面巡视遥操作信息服务可视化控制指挥任务、2017 年 4 月 20 日至 9 月 22 日天舟一号货运飞船发射与飞行任务可视化控制指挥任务。正是这些实战工程使广大

需要全本请访问 www.ertongbook.com

师生深刻体会和清醒地认识到复杂网络作为一门交叉门类技术在信息时代和重大工程中的发展需求。同时也使他们积累了相关学科领域的宝贵实战经验。没有这些工程的实战经验，也就没有本书的真正意义和实际价值。在本书撰写之前，刘贾贾、龙瑞、王中义对本书的内容作了初步的调研和试验，并对相关内容提出了许多有价值的观点和有用的素材。在校稿过程中，白英杰、申跃杰、左宗成和石家庄铁道大学复杂网络与可视化研究所的全体师生对字里行间中的数百处勘误付出了辛勤的劳动，北斗卫星导航科技邢台有限公司石家庄铁道大学可视化技术研究院的领导和专家对本书的终稿内容和出版准备工作提出了宝贵的意见。在此对他们的工作一同表示真挚的感谢。

目　录

第 1 章　绪论 …… 1

1.1　信息数据长期保存研究现状 …… 1
1.2　复杂网络的研究现状 …… 2
1.2.1　六度理论在维基百科的应用 …… 3
1.2.2　六度理论在 MSN 的应用 …… 3
1.2.3　六度理论在数学界的应用 …… 4
1.2.4　六度理论在演艺界的应用 …… 4
1.2.5　六度理论在人立方关系搜索的应用 …… 4
1.2.6　哥伦比亚大学的小世界项目 …… 4
1.3　本体的研究现状 …… 5

第 2 章　复杂网络概述 …… 7

2.1　复杂网络的基本概念 …… 7
2.2　复杂网络的典型特征 …… 8
2.2.1　小世界效应 …… 8
2.2.2　度分布 …… 9
2.2.3　无标度性 …… 9
2.2.4　群落结构 …… 9
2.3　几种典型的网络模型 …… 10
2.3.1　ER 模型 …… 10
2.3.2　WS 模型 …… 12
2.3.3　BA 模型 …… 13
2.3.4　权重网络模型 …… 14
2.4　现实中的复杂网络 …… 15
2.4.1　信息网络 …… 15
2.4.2　技术网络 …… 16
2.4.3　生物网络 …… 17
2.5　本章小结 …… 17

第 3 章　万维网及其链接结构复杂性分析 …… 19

3.1　万维网概述 …… 19
3.2　网络链接分析 …… 21
3.2.1　数据样本的选择 …… 22

3.2.2　网页超链接的提取 ……22
3.2.3　网络分析工具 Pajek ……24
3.3　数据集的选取及预处理 ……25
3.3.1　研究假设 ……26
3.3.2　数据的预处理 ……26
3.4　网络相关指标的计算 ……27
3.4.1　结点度 ……27
3.4.2　平均路径长度 ……28
3.4.3　集群系数 ……29
3.5　网络图的绘制 ……30
3.5.1　ER 随机网络图的绘制 ……31
3.5.2　BA 无标度网络图的绘制 ……31
3.6　万维网的小世界效应 ……33
3.6.1　相关指标的计算 ……33
3.6.2　两结点间的最短路径 ……34
3.6.3　度分布 ……35
3.7　本章小结 ……37

第 4 章　软件系统函数调用关系图及其结构复杂性分析 ……39

4.1　函数调用关系图概述 ……39
4.2　函数调用关系图构造算法 ……40
4.3　数据集的选取及预处理 ……41
4.3.1　研究假设 ……41
4.3.2　数据的预处理 ……42
4.4　源程序中函数调用关系图的绘制 ……42
4.5　函数调用关系图相关指标的计算 ……43
4.5.1　函数调用关系图的结点度 ……43
4.5.2　函数调用关系图的结点度分布 ……43
4.5.3　平均路径长度和直径 ……44
4.5.4　函数调用关系图的集群系数 ……44
4.6　ER 随机网络图和 BA 无标度网络图的对比 ……44
4.6.1　ER 随机网络图和 BA 无标度网络图的绘制 ……44
4.6.2　相关指标的计算及对比 ……45
4.7　本章小结 ……45

第 5 章　数据格式转换关系模型研究 ……47

5.1　现实数据格式转换关系研究 ……47
5.2　构建小世界网络模型的方法 ……49

5.2.1 WS 小世界网络模型 MATLAB 程序 ····· 49
5.2.2 NW 小世界网络模型构建过程中的加边程序 ····· 51
5.2.3 WS 小世界网络模型和 NW 小世界网络模型的比较 ····· 51
5.2.4 构建小世界网络模型 ····· 54
5.3 属性模型构建与分析 ····· 54
5.3.1 软件生命周期模型 ····· 55
5.3.2 信息数据生命周期模型 ····· 57
5.3.3 数据格式生命周期模型 ····· 59
5.3.4 数据失真转换模型 ····· 61
5.4 属性模型对 WS 小世界网络模型的影响 ····· 62
5.5 最佳小世界网络模型构建过程 ····· 63
5.5.1 类型定义 ····· 63
5.5.2 WS 小世界网络模型和 NW 小世界网络模型的构建 Z 描述 ····· 63
5.5.3 属性模型的构建 ····· 66
5.5.4 网络模型的加权 ····· 67
5.5.5 网络模型的叠加 ····· 68
5.6 本章小结 ····· 70

第 6 章 工程信息本体的构建及复杂性分析 ····· 71

6.1 工程信息本体的理论分析 ····· 71
6.1.1 本体和工程信息本体的概念 ····· 71
6.1.2 工程信息本体的概念分类 ····· 73
6.1.3 工程信息本体的概念体系 ····· 76
6.1.4 工程信息本体的语义关系及其表达 ····· 77
6.2 工程信息本体的构建 ····· 79
6.2.1 工程信息本体的构建工具 ····· 79
6.2.2 工程信息本体的描述语言 ····· 79
6.2.3 工程信息本体的构建规则 ····· 82
6.2.4 工程信息本体的构建方法 ····· 82
6.3 工程信息本体的构建过程——以软件本体的构建为例 ····· 83
6.4 工程信息的采集和规范处理 ····· 86
6.4.1 工程信息采集系统 ····· 86
6.4.2 工程信息的规范处理 ····· 88
6.5 基于本体的工程信息保存和查询 ····· 89
6.5.1 信息输入 ····· 89
6.5.2 信息更新 ····· 89
6.5.3 信息查询 ····· 89
6.5.4 数据格式转换的路径检索 ····· 90

6.6　工程信息的复杂性分析 …… 92
6.6.1　工程信息数据格式转换网络的描述 …… 92
6.6.2　工程信息数据格式转换网络的小世界网络特性 …… 93
6.6.3　工程信息数据格式转换网络的结点度分布 …… 97
6.6.4　用小世界网络特性规范工程信息管理 …… 100
6.7　本章小结 …… 101
参考文献 …… 103

第1章 绪 论

1.1 信息数据长期保存研究现状

随着科学技术的迅速发展、信息的海量增加、信息的长期利用，以及计算机技术的不断发展更新，人们开始研究数字信息的长期保存和利用。数字信息长期保存的研究经历了两个发展阶段[1]：一是数字信息存取与保护调查研究阶段；二是数字信息长期保存措施研究阶段。

1）数字信息存取与保护调查研究阶段。在20世纪90年代初，欧美等发达国家和地区就对数字信息长期保存进行了大量的研究，并且发布了《电子文件的存取与保护》《保护数字信息：数字信息归档特别工作组报告》《保护数字信息：总结报告与建议》等研究成果。

2）数字信息长期保存措施研究阶段。在20世纪末和21世纪初，数字信息长期保存过程中的问题开始引起一些组织机构的关注，它们开始致力于研究数字信息长期保存过程中遇到的问题。目前国际上颇具影响力的项目有Electronic Records Archives（ERA）项目[2]、Open Archival Information System（OAIS）项目[3]、The International Research on Permanent Authentic Records in Electronic Systems（interPARES）项目[4]、CURL Exemplars in Digital Archives（CEDARS）项目[5]。ERA项目主要研究的是保存任何形式的电子记录，提供给现在及未来存取使用；OAIS项目为具有数字信息长期保存和利用责任的档案机构提供数字信息长期保存和利用的基本概念、术语、基本信息模型、功能模型和指南；interPARES项目主要是为了“建立一整套的理论和方法用于切实保障在数字形态中创建或维护的可信文件，这些理论和方法为永久性保护电子文件及其真实性提供模型构建方针、策略和元数据标准，从而确保电子文件及其真实性的永久保存”；CEDARS项目主要研究确认数字信息收藏体系长期保护的战略框架和具体方法。

数字信息长期保存过程中遇到的主要问题有数字信息的可读性、可用性、完整性、可理解性和真实性等，这些问题主要涉及数字信息长期保存过程中的介质问题、软硬件兼容问题、格式兼容问题、数据完整真实及不丢失等问题。为了解决这些问题，目前研究的技术措施主要有更新、技术典藏、迁移和仿真4种。

1）更新。更新是将数字信息从一个存储媒介复制到另一个存储媒介上的方法，是为了保护数据不受所在存储媒介性能的影响。这种方法只能解决存储媒介的老化问题，并不能解决数字信息长期保存过程中的可用性、完整性和真实性等问题。

2）技术典藏。技术典藏是将计算机硬件本身进行典藏，将过时的计算机软硬件系统进行保存，建立计算机硬件博物馆，利用过时的计算机硬件读取存储的数字信息。典

需要全本请访问 www.ertongbook.com

藏的内容应该包括原始的应用程序、操作系统软件及硬件平台。只有保存原始数据的原始运作环境，才能实现数字信息的长期保存。

3）迁移。迁移是根据软硬件的发展情况，将数字信息从旧的软硬件环境转换到新的软硬件环境，从而保证数字信息可以长期保存和利用，它是数字信息长期保存的主要技术手段。现在采用的迁移方式主要有两种：第一种是将数字信息从稳定性比较低的媒体迁移到稳定性较高的媒体上；第二种是将数字信息从各种不同的数据格式迁移到当前广泛使用的、易于管理的几个标准格式中，也就是将数字信息的存储数据格式进行转换，在转换过程中会出现一定程度的数据丢失。有效的迁移必须保证数字信息的真实性、完整性、可用性和可理解性，而当数据量很大时，数据迁移所用的人力、物力及时间费用都会很高，这都是现在所面临的问题。

4）仿真。仿真是指利用仿真软件在新的系统环境中建立一个兼容数字信息产生时的软硬件运行环境，通过仿真程序仿效与过时的软硬件相关的操作系统，从而形成数字信息所依赖的原始运行环境，是延迟技术的一种方法。仿真被视为理想的保存策略，但是由于其技术的复杂性，目前仿真还处于研究和实验阶段，没能真正地投入使用。

数据长期存取过程中的数据转换包括存储媒介的转换和存储信息文件的数据格式转换：存储媒介转换是指将数据从原有媒介上转移到适合长期保存的媒介上存储，而数据格式转换则是尽可能地使用理想的文件保存格式代替原有的文件保存格式，也就是通过转换数据的保存格式，来保证数据信息的长期有效可用。

我国在数字信息长期保存方面尚处于起步阶段，数字信息资源相关部门已经意识到信息长期保存的重要意义，并开始在国家科技部的支持下研究数字信息资源的长期保存。2004 年 7 月，“中欧数字资源长期保存国际研讨会”在北京召开，这对我国在数字信息长期保存方面的研究具有重要的意义。

1.2　复杂网络的研究现状

将复杂系统中的个体定义为网络的结点，个体之间的关系定义为网络中的边，这样由大量的结点及结点间的连接所构成的网络称为复杂网络。对复杂网络的研究最早可追溯到 18 世纪“七桥问题”的研究[6,7]。随后，在 20 世纪的后 40 年中，随机图理论一直是研究复杂网络的基本理论，但现实中绝大多数的复杂系统结构并不是完全随机的。直到 20 世纪末，Watts 和 Strogatz 发表了《“小世界”网络的集体动力学》[8]，Barabàsi 和 Albert 发表了《随机网络中的标度涌现》[9]。Watts 和 Strogatz 提出了小世界网络模型，即 WS 网络模型，描述了从完全规则网络到完全随机网络的转变；Barabàsi 和 Albert 提出了无标度网络模型，即 BA 网络模型，该网络的连接度分布具有幂律形式，其幂律分布没有明显的特征长度。随着这两大理论的提出，复杂网络的研究进入新的阶段，在现实系统中得到广泛应用，如科研合作网[10]、万维网[11,12]、科研引用网[13]、语言网[14-16]、

因特网[17]、电力网[18-20]、航空网[21,22]、生物代谢网[23]与蛋白质网[24]等。在国内，多所高校设有复杂网络研究中心，最近几年也设置了关于复杂网络的基金项目[25]，对复杂网络的研究已成极其重要而且富有挑战性的前沿科研课题。

匈牙利作家 Karinthy 在 1929 年提出了“小世界现象”的设想，他认为世界上随意选择两个人都可以通过 6 个熟人而联系起来。在 20 世纪 60 年代，美国哈佛大学的社会心理学教授 Milgram 通过设计一个送信实验，提出了著名的“六度分离”假说[26,27]，大意为任何两个欲取得联系的陌生人之间最多只隔着 6 个人，便可完成两人间的联系。当年，Milgram 给内布拉斯加州奥马哈市随意选择的 300 多人发信，要求他们把这封信寄给波士顿市一个独一无二的目标人，分别由每个人独自联系。Milgram 告诉每个发信人有关目标人的信息，包括姓名、所在地、职业，如果发信人不认识这个目标人，他们可以把这封信寄给他们认为有可能认识目标人的熟人。以此类推，形成了发信人的链条，链上的每个成员都力图把这封信寄给他们的朋友、家庭成员或同事熟人，以便使信件尽快到达目标人。最后 Milgram 发现，有 60 个链条最终到达目标人，链条中平均长度大约为 6，Milgram 便由此得出结论：任意两个人都可平均通过 6 个熟人联系起来[28]。从严格的科学角度来讲，“六度分离”理论猜想的因素居多。Milgram 的实验事实上并不十分理想，实验的结果是有 4/5 的信件没有到达目标人。一直以来，虽然“六度分离”理论一再被研究提及，但一直没有获得颇具说服力的佐证。因此，“六度分离”与其说是一种理论，不如说是一种假说。

“六度分离”现象在学术上称为小世界效应（small world effect）。小世界效应的定义如下：如果网络中的任意两结点之间的平均距离 L 随网络结点数 N 的增加呈对数增长，即 L-ln$N$，并且网络的局部结构仍具有较为明显的集群性质，那么该网络就具有小世界效应。这里的平均距离 L 具有广泛的含义，如在上述信件传递实验中，平均距离 L 就是平均传递次数 6。小世界效应表现了一个普遍存在的客观规律：在如今的信息化时代，人们之间的关系已经完全社会化，任意两个素不相识的人都可能通过“六度分离”现象产生联系。在规模庞大的网络中各元素之间的“距离”实际上是非常“近”的[29-31]。

自从 Milgram 教授在 20 世纪 60 年代提出“六度分离”理论之后，“六度分离”理论已经在互联网络、数学界、演艺界及科学合作网络等领域成功应用。

1.2.1　六度理论在维基百科的应用

爱尔兰都柏林大学的学生 Stephen Dolan 设计了维基百科六度理论查询器。该软件的原理是利用维基百科每篇条目内包含的链接，计算从一篇条目到另一篇条目的最少单击次数。软件在测试英文维基百科时发现，条目 2007 是英文维基百科的中心条目，通过这个条目链接到达其余的 2 111 479 篇条目平均只需要单击 3.45 次鼠标就能够实现。由此可以得出，英文维基百科的平均路径长度是非常短的。

1.2.2　六度理论在 MSN 的应用

微软的研究人员通过过滤 2006 年某个月的 MSN 信息，对 1.8 亿名用户的 300 亿通

信信息进行匹配比较，最后发现任意用户与整个数据库的 1 800 亿组配对建立联系平均只需要通过 6.6 个人，高达 87%的用户可以在 7 次内完成关联。这就说明 MSN 通信网络同样也符合六度理论。

1.2.3　六度理论在数学界的应用

埃尔德什数，简称为埃数，是根据发表论文高达 1 475 篇的匈牙利数学家保罗·埃尔德什而命名的。埃数描述的是数学论文中某个作者与保尔·埃尔德什之间的“合作距离”。埃数的定义如下：保罗·埃尔德什的埃数是 0，与其合写过论文的人的埃数是 1，如果一个人至少要经过 k 个有论文合作关系的中间人才能与保罗·埃尔德什建立联系，那么这个人的埃数是 k+1。例如，保罗·埃尔德什与 M 合写过论文，M 与 N 合写过论文，而保罗·埃尔德什没有与 N 合写过论文，则 M 的埃数是 1，N 的埃数是 2。诺贝尔物理学奖得主爱因斯坦的埃数是 2，德国著名数学家大卫·希尔伯特的埃数是 4。

1.2.4　六度理论在演艺界的应用

贝肯数（Bacon number）是以演员凯文·贝肯为中心定义的，将某个演员与凯文·贝肯的距离描述为是否与其有过演出合作。著名的游戏《六度空间》也由此产生，游戏的规则描述如下：选定一名演员，如果他（她）曾经与凯文·贝肯合作演出过一部电影，那么他（她）的贝肯数为 1；如果他（她）未曾直接与凯文·贝肯合作过电影，不过他（她）与曾经和凯文·贝肯合作演出过的人合作过，那么他（她）的贝肯数为 2，以此类推。弗吉尼亚大学的计算科学家恰登曾表示，参演过美国电影的所有演员中，任意一个演员的贝肯数都不超过 4。

1.2.5　六度理论在人立方关系搜索的应用

人立方关系搜索是微软亚洲研究院设计开发的一款新型社会化搜索引擎，它的创建理念来自于六度理论。在该搜索引擎中，只要任意输入一个人名，人立方关系搜索就会给出所有可能与该人名有关的简介、人际关系、网页报道及资讯等内容。人立方关系搜索凭借它能从逾 10 亿的中文网页中提取出人名、地名等相关中文短语的能力，利用算法自动计算出相关短语之间存在关系的可能性；另外，人立方关系搜索对于支持相关短语之间关系的网页文字建立索引。此外，人立方关系搜索还能够自动查询人名之间最可能存在的关系描述词、最可能与人名相匹配的称呼和事迹等。人立方关系搜索从这些中文网页中提取出的与人名相关的人物简介等文字，并将文字按照与人名的相关性进行排序后向用户提供搜索结果。

1.2.6　哥伦比亚大学的小世界项目

2001 年秋天，美国哥伦比亚大学的 Watts 成立的研究小组开始利用电子邮件进行大规模的实验，以期验证“六度分离”理论的正确性[32]。

Watts 等创建了一个小世界项目的网站，在世界范围内进行在线实验。每一个参与

实验的志愿者都在网站上注册，从来自 13 个不同国家的 18 个目标人中随机选择一个分配给志愿者。志愿者的任务就是用电子邮件将一条信息传达给目标对象。与 Milgram 的小世界实验类似，如果志愿者不认识指定的目标对象，则可以通过这个网站把消息传达给他认为最有可能完成任务的一个熟人或朋友。

最初有近 10 万人在网站上成功注册为志愿者，其中大约 25%的人提供了他们的个人信息，并且进行了电子邮件传递实验。从 2001 年秋天开始，来自 166 个国家和地区的 61 168 名志愿者在 1 年多的时间里陆续在网站上注册了个人信息，参与了这项实验，利用这些志愿者的个人信息最终构成了 24 163 条关系链，其中有 384 条消息成功传达给目标对象，实验成功率为 384/98 847≈0.4%，实验结果并不理想。Watts 的研究小组准备进行第二次实验，不再限制志愿者只能发一封邮件给自己的熟人或朋友，以期能够提高实验的成功率。

通过对成功传达给目标对象的 384 条关系链进行分析，Watts 等发现链的平均长度为 L=4.05，但是这仅仅是完整链的平均长度，显然越短的路径就越有可能是完整的。因此 Watts 等计算了假设在零磨损（即所有收到消息的人都将消息继续传递下去，而没有人退出实验）情况下链的长度的中值，并得出根据消息的初始发送者与目标人的地理位置的远近，一封邮件平均转发 5～7 次，即可传达给目标对象，完全符合 Milgram 所描述的“六度分离”现象。当然这项实验也存在一些不可控因素，如志愿者的朋友可能没有兴趣参与该实验。另外，人们通常非常警惕来自陌生人的电子邮件，这些因素都会影响整个实验的顺利进行。

1.3 本体的研究现状

本体论（ontology）的研究起源于哲学领域，定义为“对世界上客观存在物的系统地描述，即存在论”[33]。目前被广泛接受的是 Gruber[34,35]在 1993 年给出的定义，他认为“本体是概念模型的形式化规范说明”。随着计算机和网络技术的发展和海量数字信息的生成，如何表示和组织这些信息，让人们在信息的海洋中避免迷航，能快速准确地查询所需要的信息备受关注，本体逐渐成为研究的热点。

国外对本体的研究主要集中在本体语言、本体工程和信息检索等方面。

1）本体语言。本体语言用来表示本体，具备描述和推理能力，一般情况下要基于某种逻辑语言。现有的本体语言主要分为以一阶逻辑[36]为基础和以描述逻辑[37]为基础两类。

2）本体工程。本体工程主要针对特定领域本体和通用本体构建，如建模元语、构建方法、具体构成和评价原则等。比较著名的领域本体有爱丁堡大学的企业项目和多伦多大学的虚拟企业项目，通用本体有美国微电子与计算机技术有限公司研究的 CYC 项目。

3）信息检索。基于本体的分布式半结构化信息获取（ontobroker）[38]、基于网络代理搜索本体的本体［(onto)^{2}agent］[39]和可升级知识合成（The Scalable Knowledge Composition，SKC）[40]，都是应用于信息检索方面的本体项目。

国内关于本体的研究还处于起步阶段，没有成熟的理论体系，目前主要研究方向是知识工程、信息管理和语义 Web。具体应用项目有中国科学院数学与系统科学研究院应用数学研究所的常识知识库研究[41]、中国科学院计算技术研究所的大规模知识系统研究[42]、中国农业科学院科技文献信息中心的 AOS（Agriculture Ontology Service）项目[43]及浙江大学人工智能研究所基于本体论的产品配置研究[44]。

第 2 章　复杂网络概述

如今的人类社会处处存在各种各样的复杂网络：从因特网到万维网，从科研合作网络到各种政治、经济、社会关系网络，从生物体的大脑到各种新陈代谢网络，从大型电力网络到全球交通网络等，甚至于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。任意一个有着相互作用关系的复杂系统都可以看作一个网络。例如，万维网是凭借网页之间所形成的链接结构所构成的网络[45]；科学家合作网是凭借不同科学家之间的合作关系构成的网络；基因调控网络是凭借不同的基因通过调控与被调控的关系构成的网络[46]等。

2.1　复杂网络的基本概念

18 世纪伟大的数学家欧拉对著名的“七桥问题”的研究，标志着图论研究复杂网络的开端。在图论中，网络描述为点集 $V(G)$和边集 $E(G)$组成的图。$E(G)$中的每条边 e 均有 $V(G)$中的一个结点对(u,v)与之对应。网络中结点数 N 即为网络的大小，结点间相互的连接数即为边数 $L(E)$。图还可以用矩阵表示，每一个图都可以定义一个邻接矩阵（adjacency matrix）来表示。对于网络图的研究，研究者通常关注结点之间是否存在连接边，而并不关心图中结点的位置、边的粗细长短及边是否相交等因素。网络所表现出来的不依赖于结点的实际位置和边的形态的性质称为拓扑性质，网络图的结构称为拓扑结构（topology structure）。网络中所有的拓扑结构信息都包含在与其对应的图的邻接矩阵 $\boldsymbol{A}_{i,j}$ 中。对于含有 N 个结点的图 $G(v,E)$来说，其邻接矩阵是 N 阶方阵，其矩阵元 $a_{i,j}$ 表示网络中结点 i 与结点 j 的连接关系：若结点 i 与结点 j 之间有连接边，则相应的矩阵元 $a_{i,j}$ 为 1，否则为零。在此，首先对复杂网络的一些基本概念进行简单介绍。

1）度分布（degree distribution）。度是结点的一种极为重要的属性，其定义为与结点 i 连接的其他结点的数目。网络中结点的度的大小在某种意义上体现了该结点的重要性。网络的平均度定义为所有结点的度 k 的平均值，表示为$\langle k\rangle$。网络中结点的度分布表示为 k 的分布函数 $P(k)$，代表任意选择的结点的度值正好等于 k 的概率，同时也可表示为网络中度值为 k 的结点数所占总结点数的比例。

2）平均路径长度（average path length）。网络中连接结点 i 和 j 的最短路径上的边数称为这两个结点之间的距离，表示为 L_{ij}。网络的平均路径长度 L 又称为特征路径长度（characteristic path length），定义为 L_{ij} 的平均值，也就是任意两个结点之间距离的平均

需要全本请访问 www.ertongbook.com