甲骨字网络及其特性初步探索

焦清局◎著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

甲骨字网络及其特性初步探索

焦清局 著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS
·北京·

图书在版编目（CIP）数据

甲骨字网络及其特性初步探索 / 焦清局著. —北京：科学技术文献出版社，2018. 12

ISBN 978-7-5189-4777-5

Ⅰ. ①甲… Ⅱ. ①焦… Ⅲ. ①互联网络—应用—甲骨文—研究 Ⅳ. ① K877.1-39

中国版本图书馆 CIP 数据核字（2018）第 196757 号

甲骨字网络及其特性初步探索

策划编辑：张 丹 责任编辑：王瑞瑞 责任校对：文 浩 责任出版：张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路15号 邮编 100038
编 务 部 （010）58882938，58882087（传真）
发 行 部 （010）58882868，58882870（传真）
邮 购 部 （010）58882873
官方网址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 北京虎彩文化传播有限公司
版 次 2018 年 12 月第 1 版 2018 年 12 月第 1 次印刷
开 本 710×1000 1/16
字 数 151千
印 张 9
书 号 ISBN 978-7-5189-4777-5
定 价 39.00元

前 言

早在18世纪，人们就已对“网络”进行了初步的探索：欧拉将“七桥问题”抽象为图问题，对其进行研究。图是网络在数学上的表现形式，但当时的“网络”研究只局限于极小规模的情形，因此，并没有网络概念的产生，它只是数学的一个分支。随着文章“Collective Dynamics of ‘small-World’ Networks”“Emergence of Scaling in Random Networks”的发表，以及互联网上大规模网络数据的产生，科学界掀起一股研究复杂网络的热潮，并产生了一门学科：网络科学。随后复杂网络的一些特性被人们广泛研究，如连通性、稀疏性、结点的度及其分布、平均路径长度与直径、聚类系数、度相关性、社团结构、结点重要性、网络模型等，为人们认识网络及复杂网络的应用提供了重要的理论基础。在应用方面，生物领域、物理领域、计算机领域等都取得了卓有成效的成果。

社团（或称模块）结构作为复杂网络的一个重要特性，其理论和应用都得到了广泛的研究。社团是网络的一个子网络，它要求社团内的结点紧密相连，而不同社团间的结点连接稀疏。随着模块度概念的出现，社团结构的研究逐步开展，大量的研究文献涌现。模块度不仅是挖掘社团结构的一种方法，而且是衡量不同算法好坏的标准，即模块度的值越高，表明算法的性能越好，网络分割的社团越好。因此，社团结构的挖掘成为一个交叉学科的研究内容。物理学家、计算机学家、数学家都可以设计不同的算法优化模块度的值，社团结构的挖掘问题转化为优化问题。随后，

不同的社团结构衡量标准相继出现，社团结构的研究呈现出了多元化的状态，社团结构的理论研究和应用研究并行发展。

甲骨文是迄今为止中国所发现的最早古文字，是汉字的鼻祖，传承着中华民族基因。甲骨学的研究可以极大地提高中国的文化自信。甲骨文已入选“世界记忆名录”，表明甲骨文的价值得到了全世界的公认。然而，甲骨文的研究还存在很多问题，其中最大的问题是约2/3的甲骨文未知其语义，并且利用传统的方法很难对其破译。因此，设计不同于传统的考释方法迫在眉睫。甲骨文是相对比较成形的文字系统，具有一定的复杂性。传统孤立地考释甲骨文字的方法已无法取得突破性的进展，而复杂网络作为一种描述和解决复杂系统的有力工具，用于破译未识甲骨字的语义成为必然趋势。

本书首先简单回顾了复杂网络的基本概念，以及社团结构的研究方法和进展。然后着重介绍了甲骨字网络的构建与其特性，以及社团结构挖掘的新方法。因此，本书的组织结构如下：第一章概述地介绍了复杂网络的基本概念和社团结构；第二章详细介绍了甲骨字网络的构建及其特性，并给出以后网络甲骨学的研究内容；第三章阐述了一种基于已知社团个数的社团结构识别算法；第四章详细介绍了多尺度的社团结构概念、算法和在生物网络方面的应用；第五章提出了一种基于新结点相似性的链路预测算法；第六章给出了本书的总结和未来展望。

与其他书籍相比，本书的特色在于：①首次提出网络甲骨学的概念，并利用计算机技术首次构建了甲骨字网络，对甲骨字网络进行详细的分析，为用社团结构特性预测未识甲骨字的语义提供坚实的理论保障；②详细介绍了社团结构的多尺度概念，与单尺度的社团结构相比，多尺度社团结构更能反映现实网络的特征，本书还阐述了多尺度结构在生物网络方面的应用，以及在甲骨字网络方面的应用；③提出了一种基于已知社团个数的社团结构识别算法；④笔者在本书中给出了相应算法的主要代码。这些内容

可以帮助读者尽快地进入社团结构研究的领域。

本书得到了国家语委科研规划项目（YB135-50）、河南省科技攻关项目（182102310920）、国家自然科学基金项目（61806007、U1504612）、国家社会科学基金重大委托项目（16@ZH017A3）教育部甲骨文信息处理重点实验室及教育部“甲骨文信息处理”创新团队、河南省甲骨文信息处理重点实验室的大力支持，在此表示衷心的感谢。书中每章内容后列举了参考的主要文献，在此对所引参考文献中的作者和出版机构表示感谢。

虽然本书尽可能地介绍了复杂网络和甲骨字网络各个方面的内容，但由于笔者水平有限，书中难免存在疏漏和不足之处，欢迎各位专家和读者批评指正。

目 录

第一章

绪 论

当今时代是大数据的时代，各个领域都产生了大量的数据。大量数据的产生为深刻理解社会、经济、服务等领域提供了坚实的理论数据基础。大数据代表着人类认知过程的进步。利用大数据描述事物，并通过数据分析方法挖掘有效信息，可为人们提供辅助决策，实现大数据的价值。然而，如何利用大数据对事物进行抽象和描述是大数据时代人们关注的问题。复杂网络为人们利用大数据抽象事物提供了强有力的工具，它可以使复杂的事物简单化，也是理解复杂现象的一种基本方式。复杂网络作为一门新的学科——网络科学，已引起不同领域学者的广泛关注，其研究内容涉及计算机科学、数学、物理、生命科学、社会学等众多学科。

1.1 复杂网络的定义

复杂网络（complex network）一般可以抽象为由结点（node）集 V 和边（edge）集 E 构成的图 $G=(V, E)$[1]。图 1-1 表示的是一个含有 8 个结点和 10 条边的网络示意。结点是现实世界中某一种具体事物或者人的抽象，如在社会网络（social network）中结点代表不同类型的人，蛋白质相互作用网络（protein protein interaction network）中结点代表蛋白质。边对应现实世界中事物与事物或人与人之间的联系。网络的研究是图论中研究的重点内容，而最早的图论研究可以追溯到 18 世纪著名的数学家欧拉对大家熟知的“七桥问题”的解决[2]。欧拉利用数学抽象法把 4 块陆地抽象为 4 个结点，而 7 座桥抽象为连接陆地的 7 条线，进而“七桥问题”的研究就转化为图论的研究。复杂网络的研究和图论的发展紧密相连、一脉相承，最为主要的表现

是复杂网络中的研究方法和研究思路等大部分都来自图论。

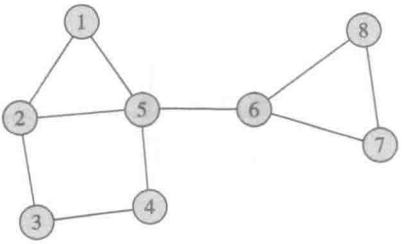

图 1-1　一个包含 8 个结点和 10 条边的网络示意

在我们认识的所有领域中都有复杂网络的存在。例如，我们生活的社会由无数个网络构成，并形成稳定的社会状态。再如，每个人身边都围绕这几个网络：亲戚网络、同学网络、同事网络。在这 3 个网络中，结点代表人、边是亲情（或血缘）、共同学习、共同工作关系的抽象。其实，每个人都在这 3 个网络中相互转换，并扮演着不同类型的角色。在家庭中，我们生活在亲戚网络中；在学习中，我们处在共同的学校中；在工作中，我们时刻与同事交流。因此，我们时刻是复杂网络中的一个结点，并与网络中的其他结点共同生活、学习和工作。

在我们看到的计算机上时刻存在着复杂网络的身影。我们上网浏览网页时，其实这个网页是因特网中的一个小小结点。当我们看完这个网页，根据网页下端的链接跳转到另外一个网页时，复杂万维网中的边就产生了。因此，在万维网（图 1-2，来源：https：//wenku. baidu. com/view/66a9a63b0166f5335a8102d276a20029bd6463d4. html）中，每一个结点代表一个网页，每一条边表示网页和网页之间的链接关系。万维网是人们构建的最大的虚拟网络，而且这个网络的大小还在增加。万维网络的产生为人们了解外部世界、获取资源提供了最快的方式。

随着高通量分子生物学实验技术的发展，产生了多水平、多层面的生物多组学数据。研究生物组分间的复杂关系，对于解析疾病的信号转导和调控过程，从系统层面了解疾病的发生、发展机制，发现复杂疾病的潜在治疗靶标及诊治生物标志物，进行系统模式的药物发现，均具有重要的研究意义。多组学为数据挖掘提供了坚实的数据基础[3]。生物组分间的复杂关系研究和多组学数据的抽象是分子生物学研究的重要内容，也迫切需要强有力的工具。复杂网络作为研究复杂系统的有效手段，已被广泛应用于生物学领域，

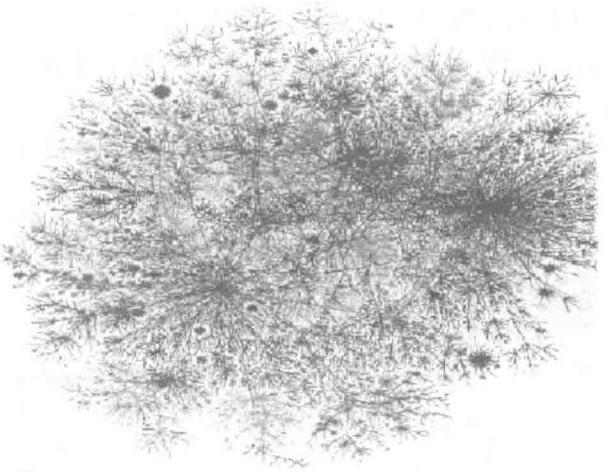

图 1-2　万维网示意

并产生了海量的生物网络数据，如蛋白质相互作用网络、代谢网络（metabolic network）、基因调控网络（gene regulatory network）（图 1-3）[4]、基因共表达网络（gene co-expressed network）（图 1-4）[5]。

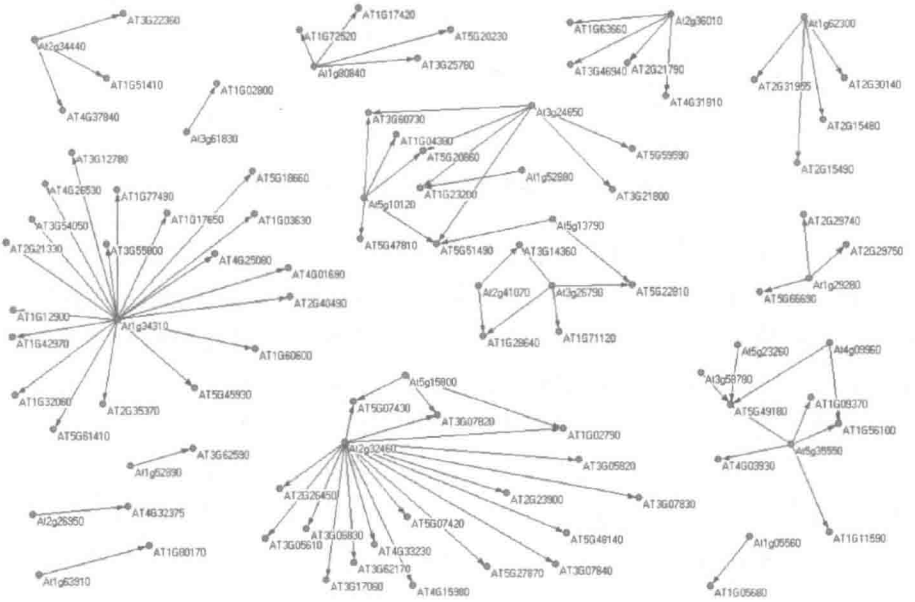

图 1-3　基因调控网络示意

大量生物网络的涌现，导致新研究方向的产生——网络生物学（network biology）。网络生物学包含了以生物网络为基础数据的各种生物学研究，也包含了以复杂网络为思维方式的生物学思维研究。网络生物学的产生为人们从系统的角度研究生物学提供了数据和理论基础，为揭示复杂疾病的病因提供了可能。

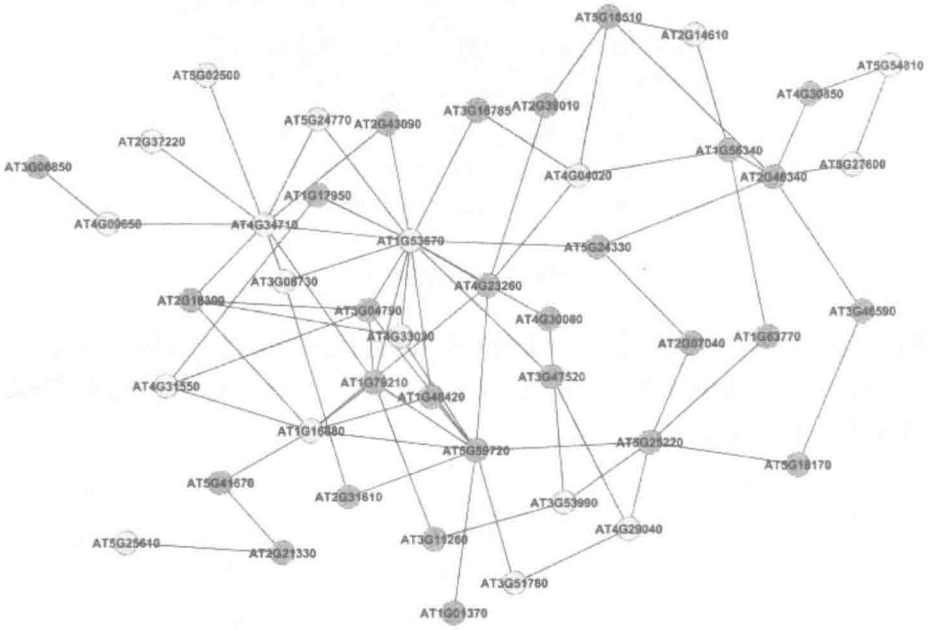

图 1-4　有关拟南芥的基因共表达网络

即使在中国最古老的文字系统——甲骨文系统中也有网络的存在。甲骨文字是一种刻在龟甲上的文字，是中国最古老且相对比较成熟的文字系统。例如，以收集的 72 151 片甲骨文拓片为研究对象，通过建模构建甲骨字网络（图 1-5），并在此甲骨字网络之上，分析网络的度分布、局部连接比率、聚类系数、模块度等相关特性。结果表明，构建的甲骨字网络不仅能充分反映甲骨文系统的单音节词多和复音节词少的古文字特征，而且能捕捉甲骨文拓片的语义单元，并具有很强的模块特性。甲骨字网络及其特性为历史学家和网络甲骨学家揭示未知甲骨字的语义提供了新的数据与理论基础。

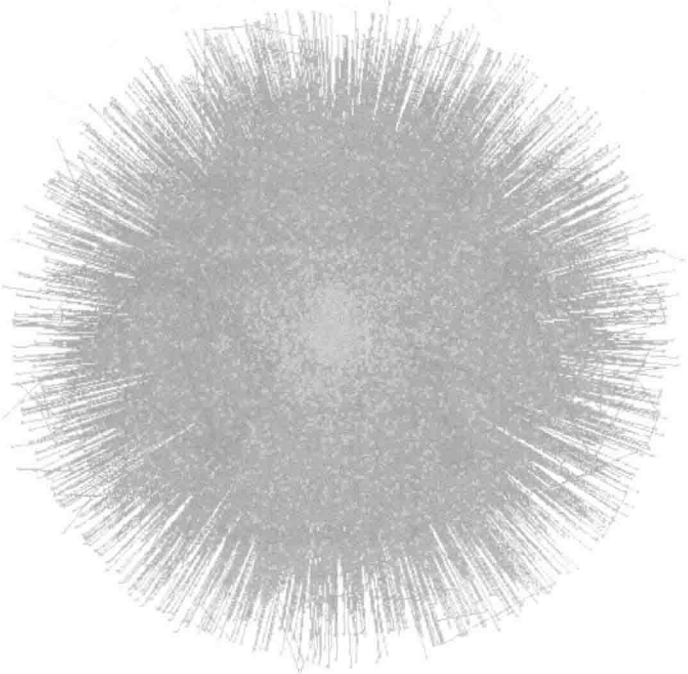

图 1-5　甲骨字网络

1.2　网络的类型

按照网络中边的类型，可以把网络分为 4 类（图 1-6）[1]：无权重无方向网络、无权重有方向网络、加权重无方向网络、加权重有方向网络。边的权重代表结点之间联系的强度，无权重网络表示的是结点之间的联系强度平等。边的方向表示结点之间的单向关系，边的方向性是现实世界的具体描述。例如，在人与人之间的认识网络中，方向性表示一个人认识另外一个人，而他们之间并不是相互认识的。

不同类型的网络并不是固定不变的，可随着时间或地点的变化而相互转变。例如，在某个时间，一个人单向认识某一个人，随着时间的推移，他们可能就相互认识了，产生的网络也会发生变化。这 4 种类型网络（或图）之间的详细转换关系如图 1-7 所示[1]。

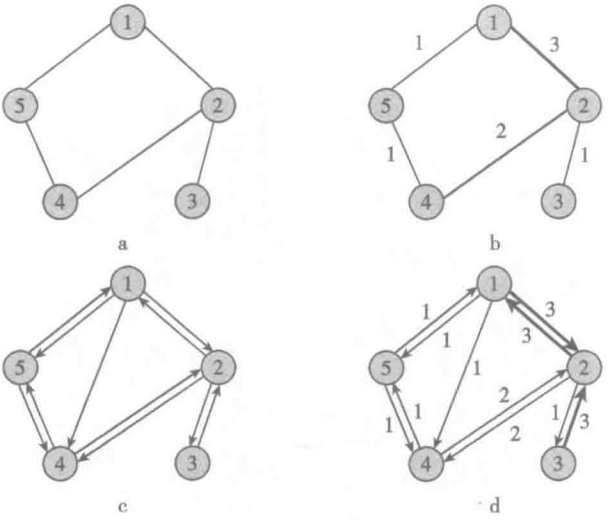

图 1-6　不同类型的网络

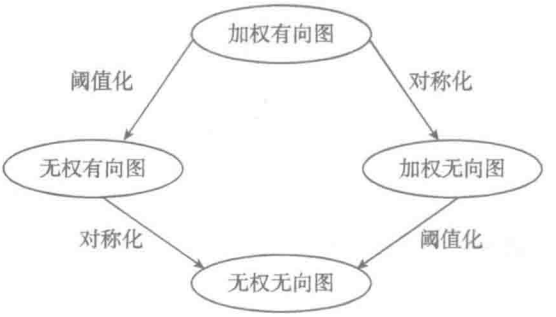

图 1-7　不同类型网络之间的关系

1.3　网络的计算机表示

网络的计算机表示是用计算机处理网络的首要任务，也是处理大规模网络的前提。最常见的计算机表示网络的方法主要有两种[1]：邻接矩阵（adjacency matrix）和邻接表（adjacency list）。

1.3.1　网络的邻接矩阵表示

假设网络（或图）$G=(V, E)$ 含有 N 个结点，其可以表示为一个 $N \times N$

的矩阵。邻接矩阵 $\boldsymbol{A}$ 可以表示为：$\boldsymbol{A}=(a_{ij})_{N\times N}$，矩阵中的元素 a_{ij} 为结点 i 和结点 j 之间的关联值，对于不同的网络，其定义如下。

（1）无权重无方向网络

a_{ij}表示为：

$$a_{ij}=\begin{cases}1, & \text{结点 } i \text{ 和结点 } j \text{ 之间有边相连}\\ 0, & \text{结点 } i \text{ 和结点 } j \text{ 之间无边相连}\end{cases}。$$

图 1-6 中 a 网络的邻接矩阵可以表示为：

$$\begin{bmatrix}0&1&0&0&1\\1&0&1&1&0\\0&1&0&0&0\\0&1&0&0&1\\1&0&0&1&0\end{bmatrix}。$$

（2）有权重无方向网络

a_{ij}表示为：

$$a_{ij}=\begin{cases}w_{ij}, & \text{结点 } i \text{ 和结点 } j \text{ 之间有边相连且权重为 } w_{ij}\\ 0, & \text{结点 } i \text{ 和结点 } j \text{ 之间无边相连}\end{cases}。$$

图 1-6 中 b 网络的邻接矩阵可以表示为：

$$\begin{bmatrix}0&3&0&0&1\\3&0&1&2&0\\0&1&0&0&0\\0&2&0&0&1\\1&0&0&1&0\end{bmatrix}。$$

（3）无权重有方向网络

a_{ij}表示为：

$$a_{ij}=\begin{cases}1, & \text{有结点 } i \text{ 指向结点 } j \text{ 的边}\\ 0, & \text{无结点 } i \text{ 指向结点 } j \text{ 的边}\end{cases}。$$

图 1-6 中 c 网络的邻接矩阵可以表示为：

$$\begin{bmatrix}0&1&0&1&1\\1&0&1&1&0\\0&1&0&0&0\\0&1&0&0&1\\1&0&0&1&0\end{bmatrix}。$$

（4）有权重有方向网络

a_{ij}表示为：

$$a_{ij}=\begin{cases}w_{ij}, & \text{有结点 } i \text{ 指向结点 } j \text{ 的边且权重为 } w_{ij}\\ 0, & \text{无结点 } i \text{ 指向结点 } j \text{ 的边}\end{cases}。$$

图 1-6 中 d 网络的邻接矩阵可以表示为：

$$\begin{bmatrix}0 & 3 & 0 & 1 & 1\\ 3 & 0 & 1 & 2 & 0\\ 0 & 3 & 0 & 0 & 0\\ 0 & 2 & 0 & 0 & 1\\ 1 & 0 & 0 & 1 & 0\end{bmatrix}。$$

1.3.2 网络的邻接表表示

网络的邻接表是用一个三元组表示，每一行用 3 个数字表示，第一个数字表示第一个结点的标识，第二个数字表示第二个结点的标识，第三个数字表示两个结点之间的权重。图 1-8 给出了图 1-6 中 d 网络的邻接表表示。

1 2 3
1 4 1
1 5 1
2 1 3
2 3 1
2 4 2
3 2 3
4 2 2
4 5 1
5 1 1
5 4 1

图 1-8 网络的邻接表表示示意

1.4 复杂网络中的主要参数概述

复杂网络中的参数是对复杂系统抽象使用的最重要工具，也是人们理论上研究复杂网络的驱动力。例如，复杂网络中的度（degree）表示的是结点