在线社交网络分析与信息传播丛书

虚拟社区发现与演化

吴 斌 张云雷 著

科学出版社

在线社交网络分析与信息传播丛书

虚拟社区发现与演化

吴 斌 张云雷 著

科学出版社

北京

内 容 简 介

本书系统介绍虚拟社区发现与演化相关的基本概念，重点介绍近 10 年来社区发现算法的研究进展；并介绍与其相关的社区演化分析算法；同时对社区发现与演化分析相关算法在其他研究方向如社会化推荐、知识图谱构建、链接预测等问题的应用进行梳理和总结；针对超大规模社交网络分析问题，专门介绍基于当前主流大数据图计算平台的并行社区分析算法；最后，针对如何简单、快捷地评价社区发现算法的优劣问题，从不同角度介绍社区分析算法评测平台的设计思路，并演示相关示例，方便用户理解。

本书是一部关于网络社区结构分析的、内容全面的参考书，可以作为网络科学等相关专业高年级本科生和研究生的教材，也可供社交网络、复杂网络结构等相关问题的科研、技术人员参考。

图书在版编目 (CIP) 数据

虚拟社区发现与演化 / 吴斌, 张云雷著. —北京：科学出版社，2018.9
（在线社交网络分析与信息传播丛书）
ISBN 978-7-03-058475-5

Ⅰ. ①虚… Ⅱ. ①吴… ②张… Ⅲ. ①互联网络－应用－心理交往－研究 Ⅳ. ①C912.11-39

中国版本图书馆 CIP 数据核字(2018)第 180521 号

责任编辑：赵艳春 / 责任校对：郭瑞芝
责任印制：张 伟 / 封面设计：迷底书装

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

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 9 月第 一 版 开本：720×1 000 B5
2018 年 9 月第一次印刷 印张：19 1/2
字数：420 000

定价：108.00 元

（如有印装质量问题，我社负责调换）

《在线社交网络分析与信息传播丛书》编委会

前　言

社交网络和社交媒体已经成为人们生活的一部分，社交网络的本身结构、网络上的行为以及信息传播规律也成为值得研究的重要科学问题。社交网络是一种复杂网络。复杂网络是网络科学研究的基本对象。而社区结构的分析是网络科学研究的基础问题。它的研究随着20世纪末网络科学的兴起而受到众多学者的关注。社区结构分析对深入理解社交网络的结构特征、进一步分析社交网络中的群体行为、认识和建模社交网络上信息的传播过程有着重要意义。本书就是针对这个问题，从问题分类、方法及技术多方面总结社区结构发现与演化分析的研究成果。

本书包括四个方面的内容：第一，基础知识，是全书基础性章节，对相关章节涉及的基础知识和核心技术进行初步介绍与总结，包括第1、2、9章，分别是引言、基础知识和总结。第二，相关算法，以时间为序介绍社区发现算法和演化分析方法的研究成果，包括第3、4、5章。第三，相关具体技术，介绍社区结构分析算法在大数据技术框架下的快速实现技术以及相关的评测技术，包括第7、8章。第四，扩展性介绍，介绍社区结构分析算法在社交网络分析其他相关问题的扩展与应用，包括第6章。

基础知识：第1章介绍在网络科学兴起的大背景下，虚拟社区发现与演化问题研究的简要发展过程，介绍虚拟社区发现与演化主要涉及的研究问题以及研究方法。并解析本书的各章结构与关联。第2章介绍本书涉及的一些基础知识，主要是图论中的基础概念和网络科学的一些基础知识，包括图的分类与表示、图的性质与路径、复杂网络模型、社区发现基本概念、算法分类、算法评估方法与标准数据集等。第9章针对社区发现与演化问题总结全书内容，并展望未来的相关研究。

相关算法：第3章和第4章介绍各类社区发现算法。第3章介绍的是早期的算法，从图聚类、模块度目标优化、概率论与信息论、物理模型等角度归纳社区发现的研究成果。第4章介绍近5年的算法，分别从重叠社区发现、异质网络社区发现、属性网络社区发现等角度总结近年来社区发现的研究进展。第5章介绍社区演化分析方法，包括问题定义、典型分析算法与框架以及评估方法。

相关具体技术：第7章在介绍目前主流的图计算并行框架的基础上，总结大数据技术背景下复杂网络数据快速社区发现计算方法。第8章是应用与开发章节，介绍专门用于社区发现算法评价的评测平台，涉及平台的框架、主要功能模块以及应用方法等。

扩展性介绍：第6章介绍与社区分析相关的网络科学领域其他问题求解的研究

进展，即引入社区分析角度对社会化推荐问题（排名）、知识图谱构建问题、链接预测、网络视频数据分析等问题的研究成果。

本书系统全面地总结复杂网络研究兴起以来社区发现与演化问题的研究进展。本书不仅详细介绍经典的分析算法，而且综述近年来新的研究进展。一方面介绍传统的单机运行算法，另一方面介绍基于当前主流图并行计算框架的快速大规模图分析方法。本书以社区分析算法为核心，扩展介绍相关方法的交叉应用以及评测系统平台。本书是一本兼顾社区分析算法基础理论和方法实际应用的、内容丰富的参考书。

国家重点基础研究发展计划（973 计划）于 2013 年设立了“社交网络分析与网络信息传播的基础研究”项目。项目从社交网络本身的结构特性、社交网络中的群体及其行为、社交网络中的信息及其传播三方面展开研究。虚拟社区发现与演化是其中一个重要研究课题。本书的写作主要由承担此项目课题“虚拟社区发现与演化”研究任务的相关单位教师与学生完成。该课题由北京邮电大学、国防科技大学、中国科学院计算技术研究所、北京大学四个单位共同承担。书中介绍了此课题研究过程中的一些成果。

由衷感谢为本书写作及出版提供实质性贡献的诸位老师和同学。北京邮电大学的数据科学与服务中心的老师和同学为本书的出版提供了基础材料和初稿。他们是：石川教授、贾丙静、吕金娜、郑玉艳、陈晓纪、张孟昊、王琳、孟琳、尹丁艺、郭谦、吴心宇、张子兴、佟雪松、戴唯、彭程程、曹桢、庄楠、周尧棋、王耀和林文鼎。特别感谢国防科技大学的周斌教授、黄久鸣老师，中国科学院计算技术研究所的余智华高级工程师、刘盛华副研究员，北京大学的许进教授、蒋飞博士对本书提供的指导与帮助。本书得到了国家重点基础研究发展计划（项目名称：社交网络分析与网络信息传播的基础研究）的支持，在此特别感谢项目首席科学家方滨兴院士和各位专家的指导与帮助。

由于作者水平有限，书中难免存在不足之处，恳请广大读者批评指正。

作　者

2018 年 8 月

目　录

第1章 引 言

20世纪末，网络科学悄然兴起。由于网络科学可能将成为物理、生物、社会等众多科学的基础，有专家将其称为21世纪的元科学[1]。在网络科学研究的众多问题中，虚拟社区发现与演化是一个典型问题。随着互联网、移动网络、物联网、社交网等技术的迅猛发展，人们有机会获得庞大的、真实的网络数据，也称为复杂网络，无论它是以机器、物品为节点，还是以用户为节点，这些复杂网络真实地出现在我们面前。如何分析这些网络，在人们发现这些网络共有的、宏观的基本统计特征之后，更深入的结构分析方法之一就是划分网络，从中观层面理解网络组成。而虚拟社区发现与演化就是从不同粒度和时间角度进一步认识网络。显然，对复杂网络结构的认识有助于人们理解网络上事物发生的过程，从而指导人们设计网络、控制网络，并最终服务于网络无处不在的社会。

严格地讲，虚拟社区发现与演化的研究并非起步于20世纪末。从图论的观点看，网络基本表示方式就是图。网络结构的分析在图论中总是能找到更早的起源。社区发现问题形式化以后可以认为是图的一种分割。主要的目标是分割在同一簇中的节点间连接的边数远大于节点在簇间连接的边数。在社会网络中，则体现在同一社区内的用户间的联系紧密程度远高于用户在社区间的联系。由于在线社交网络中用户的身份并不是与现实生活中一一对应的，针对在线社交网络的社区分析也称为虚拟社区分析，本书采用了虚拟社区发现与演化的名称，以在线社交网络分析为主，由于图结构的通用性，本书同时也涵盖了更广泛的复杂网络场景中社区分析的方法与技术。

社区发现(community detection)是网络科学的典型问题。众多学者运用各自不同学科的理论与方法针对这个问题展开了研究。Fortunato 于2010年对此做了一个很好的综述[2]。社区发现的历史可以追溯到1927年，Rice基于投票模式的相似性发现小的政治团体中的社群[3]。20年后，Homans 证明社交团体可以通过社交关系矩阵交换行列产生大致块主对角的形式而显露出来[4]。最早提出与现代的社区发现算法思路相似方法的学者应该是 Weiss 和 Jacobson。他们通过删除桥接社区间的边来发现社区。这一思路与现代的许多社区发现算法相似[5]。在计算机领域，图分割问题研究也比较早。研究来源于电路设计和计算机内存管理等实际问题。如1970年提出的 Kerighan-Lin 算法[6]。1998年 Watts 和 Strogatz 在 *Nature* 杂志上发表文章，引入了小世界(small-world)网络模型，以描述从完全规则网络到完全随机网络的转变。小世界网络[7]既具有与规则网络类似的聚类特性，又具有与随机网络类似的较小的平均路径长度。1999年，Barabási 和 Albert 在 *Science* 上发表文章指出，许多实际的复杂网络的连接度分布具有幂律形式。由于幂律分布没有明显的特征长度，该类网

络又被称为无标度(scale-free)网络[8]。这两篇论文成为现代复杂网络研究的标志性事件,也被认为是网络科学的诞生。由此,社区发现作为网络科学的一个重要问题,伴随着社交网络这一重要的应用背景兴起和发展,其引起了众多学者的研究兴趣。

网络科学知名学者 Girvan 和 Newman 于 2002 年从每个社区内连边数与期待值之差的角度提出了一种评估网络中社区结构的指标模块度(modularity)[9]。有了评估指标后,众多学者提出了一系列的以目标优化为手段的社区发现方法。虽然 2007 年,Fortunato 等指出了模块度倾向于发现大社区的解析度限制[10],但是模块度还是一种常用的社区评价指标。由于图可以用矩阵表示,使用图拉普拉斯矩阵的性质可以形成谱划分(Spectral partitioning)方法,基于矩阵的图谱分析方法也被用于社区发现。当数据挖掘领域的学者对社区发现产生兴趣时,他们的研究角度有所改变,直接将社区发现问题看作图结构聚类问题。聚类是数据挖掘的一个重要研究方向,从基于划分、基于层次、基于密度、基于模型等方面提出了大量聚类方法。这些思想都被学者用于社区发现,也提出了一系列社区发现方法[11]。随着深度学习的发展,有些学者将深度学习思想与网络结合,产生了很多表示学习的成果[12]。网络表示学习的主要思想是将网络中的节点表示成低维空间中的向量,使得传统的机器学习方法能够应用到网络数据上。DeepWalk[13]使用随机游走策略生成节点序列,通过节点序列生成节点的向量表示。LINE[14]考虑了一阶和二阶相似性生成节点的向量表示。Node2Vec[15]引入更多的节点采样策略,进而生成节点的向量表示。MNMF[16]通过保持社区属性生成节点的向量表示。Metapath2vec[17]结合元路径概念将异质信息网络中的节点生成向量表示。而这些节点的表示学习方法为社区发现提供了新思路。

第 1 章介绍在网络科学兴起的大背景下,虚拟社区发现与演化问题研究的简要发展过程,介绍虚拟社区发现与演化主要涉及的研究问题以及研究方法,并解析本书的各章结构与关联。第 2 章主要介绍社区分析的基础知识,方便读者了解一些基础概念和问题,为理解社区发现的技术和方法打下基础。第 3 章主要介绍社区发现的经典方法,虽然社区发现及演化方法已经过长足的发展,但是经典方法是社区发现的最原始和最基础的方法,掌握经典方法之后,才能更好地理解社区发现的问题,做出进一步的创新。第 4 章介绍社区发现的新兴方法,经过梳理近三年发表的社区发现论文,整理出近期社区发现主要论文的算法思想和发展方向。第 5 章介绍虚拟社区发现与演化的方法,该章节给出了虚拟社区演化的基本概念和方法及最近的演化模型和框架,解决了时序网络中的社区发现及社区演化的问题。第 6 章介绍虚拟社区发现与演化和其他学科的交叉研究,该章介绍虚拟社区发现技术在推荐系统、链路预测、情感分析、异质网络及实体消歧等领域的应用研究。第 7 章介绍社区发现与演化分析快速计算方法,介绍在大数据背景下,如何发现社区及其演化的技术方法和框架,包括基于 Spark、MapReduce 的并行化方法。第 8 章介绍社区分析算法评测平台,给出一种评价社区发现算法的评测平台,使不同算法在同一平台进行比较成为可能,解决了社区发现算法评测问题。第 9 章总结全书。希望本书能帮助读者了解社区发现及演化的背景及已有

的技术，能使读者在“人工智能”的时代背景下获得更深刻的认识。在编写过程中，作者已努力完善每一章节，但也可能存在一些不妥之处，还请读者见谅并予以指正。

参 考 文 献

[1] 李国杰. 网络科学: 21 世纪的元科学[J]. 中国计算机学会通讯, 2016, 12(4): 7.

[2] Fortunato S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174.

[3] Rice S A. The identification of blocs in small political bodies[J]. Am. Polit. Sci. Rev., 1927, 21(3): 619-627.

[4] Homans G C. The human group[J]. Harcourt Brace & World, 1950, 54(2): 261-263.

[5] Weiss R S, Jacobson E. A method for the analysis of the structure of complex organizations[J]. American Sociological Review, 1955, 20(6): 661-668.

[6] Kernighan B W, Lin S. An efficient heuristic for partitioning graphs[J]. Bell System Technical Journal, 1970, 49(2): 291-307.

[7] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393(6684): 440-442.

[8] Barabási A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509.

[9] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[10] Fortunato S, Barthelemy M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

[11] 杨博, 刘大有, Liu J M, 等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66.

[12] Zhang D, Yin J, Zhu X, et al. Network representation learning: A survey[J]. IEEE Transactions on Big Data, 2018.

[13] Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2014: 701-710.

[14] Tang J, Qu M, Wang M, et al. Line: Large-scale information network embedding[C]//Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015: 1067-1077.

[15] Grover A, Leskovec J. Node2vec: Scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016: 855-864.

[16] Wang X, Cui P, Wang J, et al. Community preserving network embedding[C]// Proceedings of the AAAI, 2017: 203-209.

[17] Dong Y, Chawla N V, Swami A. Metapath2vec: Scalable representation learning for heterogeneous networks[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 135-144.

第 2 章　社区分析基本知识

2.1　社区发现的原理

2.1.1　社区的定义

现实中的很多系统都可以用复杂网络来描述。复杂网络中的节点可表示为复杂系统中的个体，节点之间的边则是系统中个体之间按照某种规则而自然形成的一种关系。现实世界中包含着各种类型的复杂网络，如社交网络(朋友关系网络及合作网络等)、技术网络(Internet、万维网及电力网等)、生物网络(神经网络、食物链网络以及新陈代谢网络等)。这些网络都具有一种普遍的特性——社区结构(community structure)。大量实证研究表明，许多网络是异构的，即复杂网络不是一大批性质完全相同的节点随机地连接在一起，而是许多类型的节点的组合。相同类型的节点之间连接紧密，不同类型的节点之间的连接稀疏。把同一类型的节点以及这些节点之间的边所构成的子图称为网络社区(community)[1]。

在复杂网络中搜索或发现社区，有助于人们理解和开发网络，具有重要的社会价值，由此出现了许多社区发现算法。目前的大多算法将一个节点仅归属于一个社区。然而在现实自然界中，事物具有多样性的特点，一种事物往往可归属到不同的类别中，社区间必定存在重叠的现象，即一个节点可属于多个社区。例如，某个体有多种喜好，根据不同的喜好可归属于不同的群体(社区)中。因此，将每个节点仅归属于一个社区的社区称为非重叠社区，而每个节点可能属于多个社区的社区称为重叠社区。非重叠社区发现识别出的社区之间互不重叠，每个节点仅属于一个社区[2]。

下面介绍另外一种社区的概念，也是本书的重点概念，即虚拟社区。虚拟社区，又称电子社区或在线网络社区，是互联网用户交互后，产生的一种社会群体，由各式各样的网络社区所构成。虚拟社区一词在 Rheingold 于 1993 年出版的《虚拟社区》一书中有介绍。Rheingold 在其著作中指出虚拟社区系源自于计算机，基于传播所建构而成的虚拟空间(cyberspace)，是一种社会集合体(social aggregation)，它的发生来自于虚拟空间上有足够的人、足够的情感与人际关系在网络上长期发展。社区发现相关算法、方法，大部分都是基于虚拟社区的。

社区的定义往往依赖于特定的系统或实际应用。从直觉上，社区内部的边必须

比社区之间的边连接得更加稠密。大多数情况，社区是算法上的一个定义，即社区仅是算法的最终结果，不具有一个精确的预定义[3,4]。

假设图 G 的一个子图 C，其中 $|C|=n_C$，$|G|=n$。定义节点 $v\in C$ 的内度和外度分别为 k_v^{int}、k_v^{ext}，分别表示子图 C 内连接节点 v 的边数和其他连接节点 v 的边数。如果 $k_v^{\text{ext}}=0$，该节点的邻居节点只在子图 C 内，其对于节点 v 可能是一个好的群集；如果 $k_v^{\text{int}}=0$，则相反，该节点脱离了 C 且最好把该节点分配到其他群集中。子图 C 的内度 k_{int}^C 是其内部所有节点的内度之和。同样，子图 C 的外度 k_{ext}^C，是其内部所有节点的外度之和。全度 k^C 是 C 中节点的度之和。明显地，$k^C=k_{\text{int}}^C+k_{\text{ext}}^C$。

定义子图 C 群内密度 $\delta_{\text{int}}(C)$ 为 C 的内部边数与所有可能的内部边数的比，即

$$\delta_{\text{int}}(C)=\frac{C\text{的内部边数}}{n_C(n_C-1)/2} \tag{2-1}$$

同样的，群外密度 $\delta_{\text{ext}}(C)$ 是从 C 内节点引出到其余节点边的边数与群外可能的最大边数的比，即

$$\delta_{\text{ext}}(C)=\frac{C\text{内节点与}C\text{外节点相连的边数}}{n_C(n-n_C)} \tag{2-2}$$

对于 C 成为一个社区，期望 $\delta_{\text{int}}(C)$ 明显地大于图 G 的平均连接密度 $\delta(G)$，$\delta(G)$ 为图 G 的边数与可能的最大边数 $n(n-1)/2$ 的比。另外，$\delta_{\text{ext}}(C)$ 应远小于 $\delta(G)$。大多数算法的目标都是寻找到一个大的 $\delta_{\text{int}}(C)$ 和小的 $\delta_{\text{ext}}(C)$ 的最佳平衡点。一个简单的方法是，最大化所有划分的 $\delta_{\text{int}}(C)-\delta_{\text{ext}}(C)$ 之和。

连通性是社区的一个必需属性。对于 C 成为一个社区，期望其内部的每一对节点间都有一条路径相通。该特征简化了非连通图的社区检测，这种情况只需要分析每个连通的部分，除非在结果群集上添加特殊的约束。下面分别给出社区的局部定义、全局定义和基于节点相似度的定义。

局部定义主要包含完全交互度、连接性、节点度数、社区内部跟社区之间边的紧密度的差别。

全局定义主要是将真实的网络图与人工生成的伪随机网络对比，这个人工生成的伪随机网络，满足这样的条件，即其中每个节点的度数与对应的原始网络中每个节点的度数相同，在满足这个限制条件的基础上，每个节点再随机与其他节点连接，最终人工生成一个伪随机网络，而通常用的模块度 Q 这一指标，也是基于这一差异而定义的。

(1)基于节点相似度的主要思想是，如果能将节点映射到 n 维欧氏空间中，则可以用欧氏距离来表示节点间的距离。若网络不能映射到空间中，则使用指标 d_{ij} 作为节点 i 和 j 之间的距离。

$$d_{ij}=\sqrt{\sum_{k\neq i,j}(A_{ik}-A_{jk})^2} \tag{2-3}$$

其中，A是网络对应的邻接矩阵。

(2)另一个节点间的相似度用两个节点间的独立路径的数目来衡量，所谓独立路径，是指两条路径之间没有共同节点。还有一种衡量节点间的相似度的指标是用从一个节点出发，按照随机游走的规则，到达目标节点的平均步数来衡量。

2.1.2　社区发现的方法分类

1. 非重叠社区发现

早期的研究工作大部分都围绕非重叠社区发现展开[5]。近年来，基于对社区结构的不同理解，研究者在对节点集划分时采用的标准和策略不同，产生了许多风格不同的新算法，典型算法分类为模块度优化算法(包含聚类)、谱分析法、信息论方法等。

(1)基于模块度优化的社区发现算法。基于模块度优化的社区发现算法是目前研究最多的一类算法，其思想是将社区发现问题定义为优化问题，然后搜索目标值最优的社区结构。在此基础上，模块度优化算法根据社区发现时的计算顺序大致可分为三类。第一类采用聚合思想，也就是分层聚类中的自底向上的方法。第二类采用分裂思想，也就是分层聚类中自顶向下的方法。第三类为直接寻优法。此外，还有一些基于遗传算法、蚁群算法等智能算法的社区发现算法也可归为此类。

总的来说，模块度优化的社区发现算法是目前应用最为广泛的一类算法，但是在具体分析中，很难确定一种合理的优化目标，这使得分析结果难以反映真实的社区结构，尤其是分析大规模复杂网络时，搜索空间非常大，使得许多模块度优化的社区发现算法的结果变得更不可靠。

(2)基于谱分析的社区发现算法。谱分析法建立在谱图理论基础上，其主要思想是根据特定图矩阵的特征向量导出对象的特征，利用导出特征来推断对象之间的结构关系。通常选用的特定图矩阵有拉普拉斯矩阵和随机矩阵两类。图的拉普拉斯矩阵定义为$L=D-W$，其中D为以每个节点的度为对角元的对角矩阵，W为图的邻接矩阵；随机矩阵则是根据邻接矩阵导出的概率转移矩阵$P=D^{-1}W$。这两类矩阵有一个共同性质，同一社区节点对应的特征分量近似相等，这成为目前谱分析方法实现社区发现的理论基础。

基于谱分析的社区发现算法是将节点对应的矩阵特征分量看作空间坐标，将网络节点映射到多维特征向量空间中，运用传统的聚类方法将节点聚成社区。应用谱分析法不可避免地要计算矩阵特征值，计算开销大，但由于能够通过特征谱将节点映射至欧拉空间，并能够直接应用传统向量聚类的众多研究成果，灵活性较大。

(3)基于信息论的社区发现算法。从信息论的角度出发，网络的模块化描述可以看作对网络拓扑结构的一种有损压缩，从而将社区发现问题转换为信息论中的一个基础问题：寻找拓扑结构的有效压缩方式。以信息论的观点来看，互信息$I(X, Y)$