

在线社交网络分析与信息传播丛书

在线社交网络结构 特性与建模

余智华 刘盛华 张 衍 张宏莉 张 玥 著



科学出版社

在线社交网络分析与信息传播丛书

在线社交网络结构特性与建模

余智华 刘盛华 张 衍 张宏莉 张 玥 著

科学出版社

北 京

内 容 简 介

本书主要围绕在线社交网结构特性与建模展开,从在线社交网络的交友特性分析与发现、在线社交服务的结构特性度量与计算方法、面向在线社交网络多属性融合特征的建模方法、超大规模在线社交网络的高效并行建模及计算方法四个方面展开阐述。

本书总结了参研单位实际的研究经验,同时对各领域研究现状有较好的总结和梳理,可供社交网络结构分析方向的科研人员参考。

图书在版编目(CIP)数据

在线社交网络结构特性与建模/余智华等著. —北京:科学出版社, 2018.5

(在线社交网络分析与信息传播丛书)

ISBN 978-7-03-055264-8

I. ①在… II. ①余… III. ①互联网络-应用-人际关系学-研究 IV. ①C912.11-39

中国版本图书馆 CIP 数据核字(2017)第 275023 号

责任编辑:赵艳春 / 责任校对:郭瑞芝

责任印制:师艳茹 / 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>

保定市中国画美凯印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 5 月第 一 版

开本: 720×1 000 B5

2018 年 5 月第一次印刷

印张: 12 3/4

字数: 240 000

定价: 76.00 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

顾 问:

李国杰 梅 宏

主 编:

方滨兴

副主编:

贾 焰

编 委:

许 进 李建华 黄河燕 齐佳音

张宏莉 吴信东 周 斌 杨善林

胡长军 郭 莉 程学旗 廖湘科

秘 书:

王忠儒

前 言

随着互联网技术的不断发展,以推特(Twitter)、微博、脸书(Facebook)为代表的在线社交网络平台迅猛发展,并逐渐渗入人们的日常生活中,发展成为人类社会社会关系维系和信息传播的重要渠道与载体。在线社交网络的蓬勃发展,一方面极大地改变了人们的生活方式,由简单信息搜索和网页浏览转向网上社会关系的构建与维护、基于社会关系的信息创造、交流和共享;另一方面,也给社会带来了负面影响,例如,微博上流传的各种谣言、虚假信息等。因此,开展社交网络分析及网络信息传播的基础研究,有着重要的意义。

在线社交网络是一种在信息网络上由社会个体集合及个体之间的连接关系构成的社会性结构,包含关系结构、网络群体与网络信息三个要素。本书的主要内容从在线社交网络的关系结构出发,围绕“在线社交网络结构特性与建模”展开,从在线社交网络的交友特性分析与发现、在线社交服务的结构特性度量与计算方法、面向在线社交网络多属性融合特征的建模方法、超大规模在线社交网络的高效并行建模及计算方法四个方面展开阐述。在线社交网络的交友特性分析与发现方面,主要介绍在线社交网络中用户的交友行为与偏好分析,同时对社交关系的强弱程度进行度量和预测;在线社交网络的结构特性度量与计算方法方面,本书重点介绍网络的核与核度的概念及相关理论,并将其应用到社交网络结构分析当中,以分析网络的脆弱性;在面向在线社交网络的多属性融合特征的建模方法方面,本书主要介绍社交网络中消息流行度的动力学建模以及用户影响力建模方面的工作,并对异质社交网络中用户节点的对齐工作进行阐述;在超大规模在线社交网络的高效并行建模与计算方法方面,本书涵盖了当前主流的并行化处理框架,包括 MapReduce、Pregel、GBASE、Pegasus 等。

本书共 9 章,由中国科学院计算技术研究所团队、北京大学团队和哈尔滨工业大学团队合著完成。第 1 章为绪论,对在线社交网络结构特性与建模这一问题的背景和意义进行介绍。第 2 章介绍书中后续涉及的一些基础理论知识。第 9 章对本书进行总结,并对前景进行展望。上面三章的内容由三个团队合著完成。具体内容方面,中国科学院计算技术研究所团队负责第 6 章社交网络的结构动力学模型和第 7 章异质社交网络的结构建模的撰写工作;北京大学团队负责第 3 章社交网络结构特性分析和第 5 章社交网络的关系预测的撰写工作;哈尔滨工业大学团队负责第 4 章在线社交服务的交友特性分析和第 8 章超大规模社交网络的计算方法的撰写工作。

本书得到了方滨兴院士领导的国家重点基础研究发展计划(973 计划)项目“社

交网络分析与网络信息传播的基础研究”的大力支持，在此表示感谢。本书的主要内容来源于该项目中的课题“在线社交网络结构特性与建模”(课题号 2013CB329602)的成果，在此对所有参研人员表示感谢。最后，感谢参与本书资料收集、内容整理以及成果贡献的老师和同学(排名不分先后)：张羽老师，王星老师，以及满彤、王永庆、高金华、靳锐、王晓萌、高阳、李英俊、王金麟、左杨、吴艳蕾等同学。

由于作者水平有限，书中难免有不足之处，恳请广大读者批评指正。

作者

2017 年 11 月

目 录

前言

第 1 章 绪论	1
1.1 社交网络概述及研究意义	1
1.2 社交网络的结构特性分析与建模	2
1.2.1 社交网络的结构建模	2
1.2.2 异质社交网络的结构建模	5
1.3 在线社交服务的交友特性分析与发现	6
1.3.1 级联交友特性分析	6
1.3.2 博弈交友特性分析	7
1.3.3 偏好性分析	7
1.4 社交网络的动力学分析	8
1.5 大规模社交网络计算	9
1.6 本书组织结构	11
参考文献	12
第 2 章 基础理论	14
2.1 概率统计与机器学习	14
2.2 图论的基本知识	17
2.2.1 图的定义	17
2.2.2 途径、迹、路和连通	18
2.2.3 距离和直径	19
2.3 博弈论	19
2.3.1 什么是博弈	19
2.3.2 博弈中的行为推理	20
2.3.3 最佳应对和占优策略	22
2.3.4 纳什均衡	23
第 3 章 社交网络结构特性分析	24
3.1 社交网络统计特性概述	24
3.2 社交网络重要特性	27
3.3 网络的核与核度	28

3.3.1	核度的定义	28
3.3.2	在线社交网络的核度计算方法	29
3.3.3	核度的应用	32
3.3.4	实验验证	33
3.4	结构洞分析	40
3.4.1	结构洞定义	40
3.4.2	结构洞分析研究现状	41
3.4.3	结构洞用户的影响力分析	43
3.4.4	结构洞用户作用分析	46
3.4.5	小结	48
3.5	本章小结	48
	参考文献	49
第 4 章	在线社交服务的交友特性分析	51
4.1	在线社交服务的级联行为分析	51
4.1.1	级联相关研究现状	51
4.1.2	级联可预测性分析	52
4.1.3	级联预测的理论分析	53
4.1.4	实验分析结果	58
4.1.5	小结	59
4.2	网络交友行为的博弈论模型	60
4.2.1	交友行为的博弈思想和博弈行为推理	60
4.2.2	交友行为的博弈模型	60
4.3	交友行为的偏好性分析与发现	66
4.3.1	一维偏好分析模型	67
4.3.2	基于向量的偏好传播模型	68
4.3.3	偏好一致性及偏好传播的收敛性	69
4.3.4	偏好传播模型示例	71
4.3.5	实验结果	72
4.3.6	小结	72
4.4	本章小结	73
	参考文献	73
第 5 章	社交网络的关系预测	78
5.1	社交网络的关系预测概述	78
5.2	社交网络关系强度预测	79

5.2.1	社交网络关系强度预测研究进展	79
5.2.2	社交网络关系强度预测线形模型	80
5.2.3	社交网络关系强度预测概率模型	81
5.3	社交网络关系类型挖掘	83
5.3.1	社交网络关系类型挖掘研究进展	83
5.3.2	社交网络关系类型挖掘 TPGF 模型	83
5.3.3	社交网络关系类型挖掘 PLP-FGM 模型	85
5.4	基于结构特征的社交网络强弱关系的预测	87
5.4.1	问题定义	87
5.4.2	模型概述	88
5.4.3	社交网络的结构特征	94
5.4.4	相关数据集与结果	98
5.5	本章小结	102
	参考文献	103
第 6 章	社交网络的结构动力学模型	105
6.1	社交网络的结构动力学建模的介绍	105
6.1.1	在线信息传播无尺度性	105
6.1.2	爆发现象	105
6.1.3	累积效应	106
6.1.4	用户影响力	106
6.1.5	在线社交媒体信息的建模方法	107
6.2	国内外现状	109
6.3	社交网络的结构动力学模型与方法	110
6.3.1	消息流行度预测	110
6.3.2	用户的转发选择预测	116
6.3.3	用户人际影响力的推断	119
6.4	本章小结	122
	参考文献	123
第 7 章	异质社交网络的结构建模	127
7.1	社交网络节点对齐概述	127
7.2	网络对齐研究现状	128
7.3	网络嵌入学习	131
7.4	异质网络的节点对齐方法	133
7.4.1	Facebook 数据集实验	138

7.4.2 多领域合作者网络对齐实验	142
7.5 本章小结	144
参考文献	144
第 8 章 超大规模社交网络的计算方法	147
8.1 基于内存映射的单机高效图算法 MMap	148
8.2 MapReduce 并行处理方法	152
8.2.1 分布式文件系统	152
8.2.2 MapReduce 流程	153
8.2.3 矩阵向量乘法和矩阵乘	155
8.2.4 MapReduce 上的关系代数操作	156
8.2.5 MapReduce 扩展	157
8.2.6 MapReduce 的通信开销	158
8.2.7 MapReduce 的理论复杂度	159
8.2.8 基于 MapReduce 的开源平台	162
8.2.9 MapReduce 总结	169
8.3 Pregel 并行图计算方法	170
8.3.1 Pregel 设计思想	170
8.3.2 Pregel 工作流程	171
8.3.3 Pregel 中的图计算	173
8.3.4 实验结果	174
8.4 Pegasus 大规模并行图计算平台	175
8.4.1 基于 GIM-V 的图挖掘算法	176
8.4.2 节点相似度度量算法	181
8.5 本章小结	184
参考文献	185
第 9 章 总结	189
9.1 社交网络的结构特性总结	189
9.2 社交网络结构建模的关键问题	189
9.2.1 社交网络关系强度预测	189
9.2.2 基于核度理论的影响力最大化算法	190
9.3 社交网络结构建模的未来研究趋势	191

第1章 绪论

1.1 社交网络概述及研究意义

自 1998 年 Watts 和 Strogatz 在 *Nature* 上发表文献[1]提出小世界网络以及 1999 年 Barabasi 和 Albert 在 *Science* 上发表文献[2]提出无标度网络以后,掀起了一股研究复杂网络的热潮,这股热潮一直持续到今天,不但没有消退,而且吸引了更多学者的加入。在这以后的十几年中,对复杂网络的属性、拓扑结构以及模型等的研究得到了快速的发展,并且对复杂网络的典型,例如,神经网络、蛋白质网络、社交网络、科学家合作网等进行了深入的研究。对复杂网络的研究,能够帮助人们了解其内部结构和机理,从而能够更好地将其应用到人们的日常生活中。

社交网络作为复杂网络的一种典型,主要是为拥有相同兴趣与活动的人提供一个相互交流的平台,因此其中存在着复杂的用户群体互动行为。为了方便对社交网络结构进行研究,一般用一个复杂图来表示社交网络,图中的节点表示构成社交网络的基本要素的人,节点之间的连线(或称为连边)则代表人与人之间的联系。这种联系可以是有向的,就像 Twitter 中的关注关系;也可以是无向的,就像 Facebook 中的好友关系。

近些年来,在线社交网络出现了爆炸式的增长。微博、Facebook、Twitter 等社交网站已经被越来越多的人熟知和使用,这些社交网站变得越来越流行,汇聚了全球几亿的用户。这些社交网络包含海量的信息,对社交网络的研究有着十分重要的实际意义。

对社交网络结构信息的研究,根据已有的结构信息以及节点间联系的紧密程度,可以将网络中的节点划分到不同的组中,其中每个组内部的节点之间联系非常紧密,而不同组的节点之间联系相对较少,这样在同一个组内的节点关系更强,节点信任度较高,相互之间的影响也较大,组内的节点容易具有更多的相似性,因此可以根据组中已知节点的信息推断出组中未知节点的属性等信息。这是对社交网络进行社区发现的一个重要应用。另外,社交网络中节点之间的连接也具有一定的内在联系,根据网络中已有的节点连接信息或者节点的属性特点,可以推断出在将来一段时间里节点之间可能形成的新的连接信息。这是链路预测的研究内容,社交网站上的好友推荐这一应用就是根据社交网站上各个个体的连接信息,基于某人邻居的邻居也有可能是其邻居的三元闭包特性推荐用户可能认识的人,或者是根据个体其自身的

属性,如就读学校、所在单位、职业、兴趣爱好等信息,推荐其可能感兴趣的人。社交网络也是在动态变化的,对这一变化过程进行研究,利用统计数据分析过去的变化趋势,可以对未来可能出现的情况进行有效的预测,从而提前研究应该如何应对或是对其进行有效利用。网上商城中的商品个性化推荐等也是社交网络研究的典型应用,可以利用具有相似属性的个体浏览过的商品或是购买过一些相同商品的个体的购买信息进行商品推荐。

总的来说,对社交网络的研究主要集中在三个方面:社交网络的结构、社交网络的动态演化以及信息在社交网络上的传播。从本质上讲,对社交网络结构和动态演化的研究属于基础性研究,最终都是为研究信息在社交网络上的传播方式以及怎样可以使信息传播达到预期效果而服务的。

1.2 社交网络的结构特性分析与建模

1.2.1 社交网络的结构建模

社交网络是一种典型的复杂网络,其中存在着复杂的用户群体互动行为。与交通网络、通信网络和生物网络等其他复杂网络相比较而言,社交网络包含了更加海量和多元化的信息。在这些信息中,网络的结构信息十分重要:首先,它直观地反映了用户间通过相互关注建立的朋友关系,构成了网络中信息来往沟通的物理基础;其次,面向网络结构的研究不但可以印证在其他领域所发现的人类社会行为特征,如“富人俱乐部现象”等,还帮助人们发现新的社会行为规律,如社交群体中的社团聚集现象等;最后,由于网络结构是社交网络个体间信息传播的载体,所以与用户发表并传播文本内容等行为相关的用户话题建模、文本信息检索、影响力最大化问题等研究极大地依赖于对网络结构的深刻认识。可以说,社交网络的研究基础始于对其网络结构的研究。

理解社交网络最直观的方法是观察其拓扑结构。随着在线社交网络用户日益增多,通过可视化的方式来研究社交网络在很多情况下并不适用。近年来,大量工作关注于研究社交网络的统计特性。统计特性的发现不仅可以让我们更加有效地关注于社交网络的整体特点,而且它对社交网络中个体行为分析、信息传播、舆情发现及预警等关键的应用问题有着至关重要的推动作用。

社交网络结构建模是在社交网络结构特性分析的基础上,对其形成机理和演化规律认识的验证与深化。常见的模型可分为两大类:一类是网络构造模型,通过显式地设定网络中节点的加入和边的形成过程来构建网络,这类模型的优点在于过程直观并可形象地模拟人类社交行为,缺点在于参数求解比较困难;另一类模型是采用统计建模方法的随机生成模型,将复杂的网络结构生成过程简化为若干基本概率

步骤,并通过统计推断得到模型参数以还原这个生成过程,这类模型能够从宏观层面面对网络结构的形成机制进行解释,然而却不如构造模型直观和具体。

网络构造模型经历了一个由 ER (Erdos-Renyi) 随机图模型到小世界、无标度模型的演化过程。其中 ER 随机图模型由著名数学家 Erdos 和 Renyi 于 1960 年提出^[3],该模型通过假设网络中所有潜在边都具有独立同分布生成概率,成功地刻画了网络中节点间连边的无序性,并在接下来近 40 年的时间里深刻影响着人们对复杂网络现象的认知。

随着现代电子计算机技术和互联网的发展,人们得到了更加强大的数据分析工具,并接触到了更多类型的复杂网络数据,如以万维网和电子邮件网络所代表的科技网络、以科研合作和演员合作关系所代表的社交网络、以道路运输网络代表的交通网络等。在实验中,人们发现 ER 随机图并不能有效地解释复杂网络中的一些常见现象,如通过该模型生成的人工仿真网络并不具有平均路径长度较短、聚集系数较高和节点度分布服从重尾分布的特征等,而这些规律在真实世界的网络中是普遍存在的。由此,人们需要通过新的网络模型刻画这些网络结构。在世纪之交,两个具有里程碑意义的工作分别刊登于 *Nature* 和 *Science*, 其中 WS (Watts-Strogatz) 模型^[1]通过对规则栅格网络中的边进行概率重连接的方法,构造出了参数特征介于规则网络和 ER 随机网络之间的小世界网络,该网络具有较高的聚集系数与较短的平均路径长度,并在一定程度上解释了小世界现象的形成要素,而 BA (Barabasi-Albert) 模型^[2]通过引入“择优连接”机制,构造出了具有无标度特征的网络结构,解释了导致网络产生重尾效应的客观规律。类似地,人们围绕小世界现象和无标度特征还建立了不同的模型,其中最主要的一类是基于“择优连接”机制展开的,这些模型都通过将新的节点加入网络来扩增网络规模,并以择优连接机制建立网络中的连接。然而这类模型的共同问题在于其网络直径是缓慢增长的,这与实际网络中常见的“直径缩减”特性相悖。由于这些模型通常是显式地给出了网络的生成机制,所以难以估计模型的相关参数,或者进行这种估计的价值是不大的,这就解释了为什么很少能看到将其与观测数据进行拟合和进行参数估计的工作。

上述工作的共同问题在于,模型都只注重于刻画网络中的一个或少数几个静态特征,模型不存在显式解,造成了参数估计比较困难,而 Kronecker 图模型则有效地克服了这些问题。Leskovec 等^[4]发现可以用矩阵的 Kronecker 乘积操作来生成网络,将两个图之间的 Kronecker 积定义为它们的邻接矩阵的 Kronecker 积,就可以进行图的扩展操作,并且扩展生成的图具有自相似的特性。整个模型的生成过程是对一个初始图 K_1 , 进行 $i-1$ 次 Kronecker 积操作,最终形成 K_i 。由于该方法具有独特的递归构造形式,所以 Kronecker 图模型所生成的网络具有良好的可分析性,并且对于具有成百上千万节点规模的巨大网络,仍然能够较为容易地得到模型的参数估计值。实验中发现, Kronecker 图模型生成的网络可以很好地模拟静态网络的度分布、

特征值分布以及动态网络的直径、密度变化的幂律分布等特性。而这些优良特性是前述模型所不具备的。

上述模型的提出,大都是由相关研究人员根据网络结构的静态特征,基于经验假设并设计相应算法来实现的。另外一种建模的方法是:首先观察真实世界的网络演化过程来分析用户行为规律,进而在模型中设定相应的规则来模拟这些行为。这类模型建立了用户行为与网络结构演化之间的关联,能够帮助分析人类的社交行为对网络结构的影响。Leskovec 等^[5]观察到用户加入网络后连接到邻居的邻居这一行为规律,提出了森林火灾模型。在该模型的演化过程中,度高的节点更容易被新增的节点连接,并且递归调用过程使得新增节点可以形成很多边,导致了入度与出度的重尾分布。新加入的节点与代表节点的邻居形成许多有向边、网络中形成社区、新增节点在代表节点社区附近形成许多连接等原因造成了网络的密度以及有效直径的变化。这些都是与真实网络中观察到的规律相符的。区别于其他研究网络统计属性的模型,Kumar 等^[6]将网络中的用户分为 Passive、Linker 和 Inviter 三类,新加入的节点被随机地指定为三类节点中的一种。在添加边时,首先按照择优连接原则从网络中已有的 Inviter 和 Linker 节点中选择一个作为边的源端。如果源端是 Inviter 节点,则新增一个节点作为边的终端与其相连;如果源端是 Linker 节点,则从已有的 Inviter 和 Linker 节点中按照择优连接原则选出一个作为边的终端。按照该模型产生的网络与观察到的 Flickr 和 Yahoo!360° 的 3 部分构成相符,包括巨大分支、孤立社区和孤立节点。Leskovec 等^[7]对 4 个真实社交网络的演化过程进行了观察和研究,总结了用户的行为规律,并在模型中进行重现。他们提出的网络完全演化模型包含了节点到达、边初始化以及边端点选择过程。其中边初始化过程中使用了节点生存时间以及“睡眠”的概念:对于网络中的节点,按照其度的某个指数函数产生一个时间间隔 δ 并进入睡眠状态 δ 个时刻,当其从睡眠状态中苏醒时,如果生存时间还没有到期,则使用随机-随机三角闭合模型产生一个长度为 2 的边(即随机选择该节点的一个邻居,然后随机选择该邻居的一个邻居,然后将该节点与选择的邻居的邻居相连形成一个长度为 2 的边)。如果一个节点的生存时间到期了,则该节点就要停止增加边;否则继续执行睡眠-苏醒这一过程。该模型生成网络的聚集系数、度分布、最短路径长度等参数都与 Flickr 等真实网络相近,因而可用来生成任意规模的人工合成网络。

另外一类重要的结构模型是随机生成模型,这类模型不考虑网络中具体节点和边的生成过程,而是将形成网络观测数据的复杂机制简化为几个基本的概率步骤。先假定网络的观测数据由一个潜在的概率统计过程生成,然后通过统计推断得到模型参数来还原这个结构。

一个典型例子是 Handcock 等^[8]提出的隐含位置聚类模型(latent position cluster model),该模型考虑了网络连通性及用户节点在属性和聚集特性上所表现的“同质

性”等 3 个结构特征,并为每个用户分配一个在多维欧氏空间中的“社会定位”(social position)。模型通过假定相似的用户通常会聚集在同一个“社会定位”中,而相异的用户的“社会定位”相距较远,来模拟社交网络中用户交友关系的“同质性”现象。

社交网络结构的研究与分析,是社交网络研究中最早得到重视和探索的领域。本书针对社交网络结构特性和建模研究中的主要问题与研究进展进行了综述,给出了基于真实社交网络的经验数据之上统计属性的描述,包括小世界现象和无标度特性、网络弹性、核与核度、中心性、同配性、社区结构、网络特性与合作行为的关系。这些统计特征不仅包含复杂网络都具有的共性,如小世界现象,也有社交网络所具有的特性,如同配性及其演化。在结构建模部分,我们将现有研究工作分成网络构造模型和随机生成模型来阐述。

社交网络是一种特殊的复杂网络。其特殊性主要体现在它是一种智能化很强的网络,原因是它的基本要素是人。社交网络拓扑结构的形成受人的世界观、生活环境、所处位置、职业环境特别是人的成长环境的影响,导致社交网络的拓扑结构遵循“物以类聚、人以群分”的特点。

1.2.2 异质社交网络的结构建模

在如今的信息化时代,几乎所有人都会参与到各种社交网络之中,例如,在线社交网络、购物网络、学术合作网络等。在现实生活中,用户往往会参与到多个网络之中。网络用户的性质不同,以及网络本身的功能差异,使得这些网络呈现出了异质性的特点:网络本身的结构是异质的,网络的节点也是异质的(论文合作关系网络中,节点就包括作者和论文两类)。对于这些异质社交网络的结构进行分析和建模,有着重要的实际意义。

对这些网络的结构建模中,主要研究如下两个问题:①网络中节点的表示学习;②不同网络间节点的对齐。网络节点的表示学习是近年来快速发展的一个领域。在对网络数据的处理中,一个好的节点表达,对于最终任务效果的评价有着极其重要的影响。分布式表达学习在自然语言处理领域,取得了显著的效果,在多个评测任务上都取得了好成绩,也为网络表示学习领域带来了新的思路,催生了 Deepwalk^[9]、LINE^[10]等一系列优秀的网络表示学习的工作。Deepwalk 模型直接类比于文本表示学习中的 Word2vec 模型^[11],将在图上随机游走得到的序列作为“句子”,直接利用 Word2vec 模型来学习网络节点的表达。LINE 工作则在考虑节点间的直接连边形成的一阶关系外,还考虑了节点的共同邻居关系。引入这样的二阶关系后,节点的结构信息可以被刻画地更好,从而得到更好的节点表示。

在现实生活中,用户往往会参与到多个网络之中。如果我们能够准确地对齐同一个用户在不同网络中的账号,从而整合用户在不同网络场景中的所有信息,我们

就能够更好地建模一个用户。因此,不同网络间的节点对齐工作,有着重要的实际应用意义。早期的工作将节点对齐问题形式化为图同构问题,但是在社交网络中,节点的缺失使得传统的图同构算法不能很好地解决异质社交网络中的节点对齐问题。节点的表示学习,为这个问题带来了新的思考角度。通过学习节点的表达,可以更好地度量节点间的关联关系,进而完成节点的对齐工作。

1.3 在线社交服务的交友特性分析与发现

人们在社会中进行信息、思想、行为等的交互,信息、思想的交互变化过程引发了社会性的革新变化,研究发现人们在交互过程中展现出很多人类所特有的特性,本书从级联、博弈和偏好性三个角度分析了交友特性与规律。

1.3.1 级联交友特性分析

社会生活中的口口相传具有很多应用,如信仰宗教人群的扩大;新技术、新产品广泛应用;对政治大选中选举人的竞选观点的宣传产生积极影响。20世纪中期,社会学家通过观察口口相传现象,并深入分析了新思想的交互机制,称这种现象为“创新的扩散”。后来,社会学家马克·格兰诺维特、经济学家托马斯·司格灵在创新性扩散机制基础上进行了系统化研究,总结了信息、思想的交互过程的级联传播特性:初始几个节点小规模范围采纳最新的信息,然后影响其朋友、邻居等,使得越来越多的人群采纳新信息,信息通过邻居不断地向外动态传播,最终扩散到整个社会。社交网络中的级联特性对社会产生积极、有意义的影响,改变了人们的观念,提高了人们的生活质量。

社交网络中的内容共享是一种重要的机制,扩大了人们的视野,通过共享使得具有相同兴趣的人们更好地交互,也扩大了人们的交友空间。一个人分享内容给朋友,其中部分朋友再向自己的朋友二次共享,这个过程形成了共享级联,级联共享过程中信息不断地向外传播至广泛的人群中。研究人员分析了博客、电子邮件、产品推荐、社交网站中的级联特征,包括级联的结构属性、级联和内容的关系。针对级联的属性特征,主要从网络结构角度分析级联树、级联广度、级联规模等全局性属性。研究者观察级联规模发现很少发生大的级联,而且很难对级联范围进行精确预测。级联的复杂轨迹是可观察的。从内容管控角度来说,预测级联具有重要意义,可以分析相同内容被不同的人共享其级联规模的差异性规律。

级联预测问题是预测级联规模将要扩大还是保持,如果扩大,级联规模将扩大多少倍。随时间和内容变化,级联规模有很大差异。研究发现,级联具有独特的时间特征、形状特征、二次共享规律。初始级联通过大量的二次分发使得级联规模显著增长。从结构角度看,更容易预测级联的广度。而且级联增长特性和级联规

模具有显著关系。在独立级联模型^[12]基础上,通过偏序观察,加权的完美一致树(weighted perfect consistent tree, WPCT)算法引入一致树级联算法,推测级联结构。一致树关联了源节点,引入了似然性指标推测级联。WPCT 推测级联节点的精确度高于 70%。

1.3.2 博弈交友特性分析

人们接受新思想、新技术前经过了对比和抉择过程。在采用新技术会带来明显或潜在的收益前提下,人们才会采用新技术。人们在交互过程中进行着博弈,博弈后选取有利的举措。级联扩散模型说明了行为是通过直接交互向外传播的,并没有说明人们为何选择接受某思想。在级联扩散模型基础上,建立博弈分析模型分析人们思想的抉择过程,把人的主观思维决策过程和交互过程相结合,结合人的决策变化分析社会性交互机制。建立基于博弈的级联传播模型,对分析经济和政治前景具有重要意义。人们的思想碰撞、行为交互过程既有博弈又有合作,合作结果才能共赢。可以证明简单图的完全扩散界限为 $1/2$ 。如果人们选择了新技术或新思想,且在后续时间片中保持新的行为状态,称为积极改进,否则称为非积极改进。考虑积极改进和非积极改进过程下,任意图的完全传播界限最大为 $1/2$,证明了新技术的扩散范围。

1.3.3 偏好性分析

在线社交网络需要对群体或个体进行推荐,如推荐朋友、电影、商品。需要对用户的交友行为进行偏好性分析,通过挖掘用户兴趣,如职业、爱好、购物习惯,以提供有效的个性化服务。当网络规模变大时量化人们的偏好关系,需要系统化模型分析多维偏好性。

以往对偏好性的研究,主要从信息扩散角度分析。采用的模型主要有独立级联模型^[12]、线性界限模型^[12]、SIS(susceptible-infected-susceptible)模型^[13]、热扩散模型^[14]。人们的偏好是多维、立体、变化的,如总体选举,每个人有独立的选举权,人们对候选人的偏好受多方因素影响,各维偏好独立,选举了某候选人则是各维偏好性的综合评价。偏好性是一种超维属性,而不是单维属性。通过观察,群体的偏好性具有收敛特性,群体能够达成共识,产生共同偏好。早期的偏好性研究,把偏好性量化为一维实数值,假设偏好性是一个二元信号或实数值。基于向量的偏好传播模型中,采用向量表示多维偏好,量化了偏好性差异,并分析了偏好性扩散的收敛特性。从理论上分析了社会学中达成群体共识的必然性。

偏好性传播模型中,个人的偏好随朋友的偏好变化而变化。以选举为例说明偏好传播模型,在模型中包含了交互过程个体间的影响。偏好性具有收敛特性。偏好性传播过程中,无论经过几次传播,偏好性矩阵的总值不变。即最终达成一致偏好性,也就是说,同一社区内最后达成一致意见。