电子邮件网络信息传播与演化研究
——以Enron公司邮件为例

张乐君 著

哈尔滨工业大学出版社
HARBIN INSTITUTE OF TECHNOLOGY PRESS

电子邮件网络信息传播与演化研究

——以 Enron 公司邮件为例

张乐君　著

哈爾濱工業大學出版社

内 容 简 介

随着信息技术的飞速发展,社会网络的相关服务和网站越来越普遍,社会网络的相关研究也越来越多。社会网络是以人为中心,依靠人与人之间的关联关系进行信息传播的一种网络模型。个体是信息传播过程的主体,具有聚类性、主观能动性等特点。另一方面,由于网络的复杂性,使得在真实的网络上进行网络应用的研究、测试和模拟非常困难,一种有效的研究方法是建立贴近真实网络特性的信息传播模型和网络演化模型。本书以 Enron 公司的电子邮件网络为数据集,对电子邮件网络的信息传播模型和演化模型进行了讨论和研究,为其他社会网络中信息传播和网络演化分析提供理论基础和现实依据。

本书可作为数据挖掘等专业硕士研究生或博士研究生的教学参考用书,也可作为从事社会网络分析研究人员的参考用书。

图书在版编目(CIP)数据

电子邮件网络信息传播与演化研究:以 Enron 公司邮件为例/张乐君著. —哈尔滨:哈尔滨工业大学出版社,2014.9

ISBN 978-7-5603-4949-7

Ⅰ.①电… Ⅱ.①张… Ⅲ.①电子邮件-信息-传播-研究 Ⅳ.①TP393.098

中国版本图书馆 CIP 数据核字(2014)第 223290 号

责任编辑 杨秀华
封面设计 刘长友
出版发行 哈尔滨工业大学出版社
社 址 哈尔滨市南岗区复华四道街 10 号 邮编 150006
传 真 0451-86414749
网 址 http://hitpress.hit.edu.cn
印 刷 黑龙江省地质测绘印制中心印刷厂
开 本 787mm×960mm 1/16 印张 7.5 字数 139 千字
版 次 2014 年 9 月第 1 版 2014 年 9 月第 1 次印刷
书 号 ISBN 978-7-5603-4949-7
定 价 28.00 元

前　言

随着信息技术的飞速发展，社会网络的相关服务和网站越来越普遍，社会网络的相关研究也越来越多。但目前社会网络中个体行为特征的分析和识别技术研究还处于起步阶段，迫切需要针对社会网络中个体的行为特征、个体参与感兴趣信息行为特征以及在此基础上细化社会网络的信息传播和演化的研究。《电子邮件网络信息传播与演化研究——以 Enron 公司邮件为例》以 Enron 公司的电子邮件网络为数据集，对电子邮件网络的信息传播模型和演化模型进行了讨论和研究，为其他社会网络中信息传播和网络演化分析提供理论基础和现实依据。

本书介绍了社会网络中信息的传播与演化基础知识。全书共分 5 章。第 1 章主要介绍社会网络个体行为分析技术、信息传播模型和网络演化模型的研究现状，并介绍了 Enron 公司的邮件数据集的来源、数据整理以及其中的人员信息；第 2 章介绍基于社会网络分析（SNA）的 Enron 数据集分析的结果，并对 Enron用户职位的不同行为进行了分析；第 3 章在介绍邮件“用户—主题—关键词”三层网络模型的基础上，分析了不同角色用户参与主题的行为模型；第 4 章在比较了邮件传播网络和病毒传播网络异同的基础上，提出了一种基于邮件角色用户行为的网络信息传播模型；第 5 章提出了一种基于用户传播模式的电子邮件网络的演化模型，并与其他经典模型进行了比较分析。

本书得到了国家自然科学基金青年科学基金（61100008）、中国博士后科学基金（20100480980）以及中央高校基本科研业务费（HEUCF100602）的支持，在此表示感谢。

本书在写作过程中，哈尔滨工程大学计算机科学与技术学院的张健沛、杨静、国林等老师，对本书的研究结构和研究内容提出了许多宝贵的意见，在此深表谢意。同时，感谢哈尔滨工业大学出版社老师们的辛勤劳动，使得本书能够顺利出版。

社会网络分析的研究正越来越受到学术界的关注，本书虽尽力将作者所做的研究成果展现出来，但在研究过程中难免存在不足之处，我们将在后续的研究中进一步完善与升华。同时，也恳请读者批评指正，共同促进社会网络信息传播和演化技术研究的进一步发展。

作　者

2014 年 6 月

目　录

第1章 绪 论

1.1 研究意义

随着信息技术的飞速发展,社会网络的相关服务和网站越来越普遍,社会网络的相关研究也越来越多。社会网络是以人为中心,依靠人与人之间的关联关系进行信息传播的一种网络模型。个体是信息传播过程的主体,其具有聚类性、主观能动性等特点。另一方面,由于网络的复杂性,使得在真实的网络上进行网络应用的研究、测试和模拟非常困难,一种有效的研究方法是建立贴近真实网络特性的信息传播模型和网络演化模型。因此,研究个体行为驱动的社会网络信息传播及网络的演化规律是迫切需要解决的问题。然而,已有的研究成果不能很好地实现信息的传递和网络的演化分析,因为目前的研究大多将社会网络中的个体同样对待,并且仅注重网络演化的宏观规律,忽略了微观特性;而研究个体行为驱动的信息传播和网络演化规律的目标是要体现网络中个体的行为特征,正是注重了微观的特性。但是,社会网络中个体的多样性,使得单纯地在信息传播和网络演化的时候再考虑个体的行为特征无法达到令人满意的结果。

因此,研究社会网络信息传播和网络演化是十分必要的,但目前社会网络中个体行为特征的分析和识别技术研究还处于起步阶段,基于个体行为驱动的社会网络信息传播和网络演化还未见报道。迫切需要研究针对社会网络中个体的行为特征、个体参与感兴趣信息行为特征以及在此基础上细化社会网络的信息传播和演化的研究。本书以 Enron 公司的电子邮件网络为数据集,对电子邮件网络的信息传播模型和演化模型进行了讨论和研究,为其他社会网络中信息传播和网络演化分析提供理论基础和现实依据。

1.2 研究现状

20 世纪 60 年代,美国哈佛大学社会心理学家 Milgram 提出了“六度分割”理论[1],这被认为是社会网络理论的基础。随着计算设备信息处理能力的大幅提升以及不同学科的相互渗透,人们可以处理和分析规模巨大且类型不同的实

际网络数据，以达到揭示网络所共有结构特性的目的。其中，以 1998 年 Watts 和 Strogatz 建立小世界网络模型[2]，1999 年 Barabsi 和 Albert 建立无标度网络模型[3]为标志，人们对复杂网络的研究进入高速发展时期，也吸引了来自数学与系统科学、统计物理、生命科学、信息科学、社会科学、经济金融等多个领域的研究人员的关注。使用复杂网络理论对社会网络进行分析主要关注个体间相互关联和作用的拓扑结构，这也是理解社会网络性质和功能的基础。就计算机领域而言，关于社会网络研究的文章在近几年已经频繁出现于 KDD，VLDB，WWW 等国际会议上。社会网络的研究主要从两个方面考虑：社会网络的静态性质和动态特征。静态性质研究包括拓扑分析[4~5]、社区挖掘[6~7]、关键节点发现[8~9]等；但另一方面，随着时间的推移，社会网络中新节点会不断地加入，节点间的连接也会不断变化，从而使得社会网络具有动态特性。对于这方面的研究主要集中在社会网络形成和进化[10~11]等方面。另外，社会网络拓扑结构与其动力学的相互作用也被人们所关注，比如社会网络中的信息传播[12~13]等。下面，从社会网络个体行为分析技术、信息传播模型和网络演化模型三个方面介绍国内外的研究现状。

1.2.1 社会网络个体行为分析技术

社会网络信息服务的基本目的是为用户提供优良的信息平台，以便吸引并留住更多的用户个体，继而将用户转化为资源。因此社会网络的核心是用户，并能够为用户提供感兴趣的信息和工具，所以社会网络中用户个体的行为分析对社会网络的发展是非常重要的。

社会网络个体行为是指社会个体成员在社会网络中相互交往、相互作用的种种表现形式，而表现形式多种多样，从自我出发的表现形式有交往的方位、频度、紧密程度等；从整体出发的表现形式有体现参与者在网络中的地位的中心性指标等。从一般意义上讲，个体行为是相对于社会网络的群体行为而言的，是指在一定的思想认识、情感、意志、信念支配下，个体所采取的行动。在社会网络中，个体行为带有一些普遍的特征，这些特征表现为：

1. 行为的自发性

个体行为是具有其内在的动力自动发生的，外在环境因素可以影响个体行为的方向与强度，但却不能发动个体行为。

2. 行为的主动性

个体行为不是盲目的，任何行为的产生绝不是偶然出现的，任何行为都是受个体的意识支配。行为者可能并不自觉地意识到自己的行为的原因，但这绝不

表示其不受自己意识的控制。

3. 行为的持久性

由于行为是有目的性的，是个体主动发生的，通常，在个体没有达到自己的目标之前，这种行为不会停止下来。

4. 行为的可变性

个体在追求个人目标以及环境变化时，选择最有利的方式，达到个人的目标。

社会网络中群体行为决定着个体行为的方向，个体行为是群体行为的体现。群体是由个体构成的，因此，社会网络的群体行为离不开个体行为，但群体行为并不是个体行为的简单相加。其原因是：当某群体把个体凝聚在一起时，就具有该群体的意识和目的，并且具有其特定的社会性，该群体的活动效果反映着整个行为主体的状况，而不再以个体的意识、目的为转移。

美国 Northeastern University 复杂网络研究中心的科研人员通过对 5 万名随机挑选出的移动用户进行为期 3 个月的数据跟踪研究，构建并分析了移动用户关联网络，发现其中 93% 的人群行为可预测[14]，这进一步证明了利用复杂网络来研究人的心理与行为的巨大潜力。目前有一些尝试把社会网络分析应用于电信领域，例如：使用社会网络分析方法对用户通话记录进行分析，分析了客户在呼叫网络的信息传播过程中的重要地位，为客户细分提供依据[15]。社会网络用户行为分析的研究方法一般有两种：一种是通过大量的调查问卷，这种方法工作量大，而且需要许多人配合，数据准确度没有保障；另一种是通过软件工具采集社会网络的实时数据，该方法可以在短时间内收集大量数据，并且信息的实时性和准确性有保证。因此采用后者进行用户行为分析是未来的发展方向[16]。

目前，社会网络中由于个体行为所产生的动态特性受到了广泛关注。一些基于网络演化的分析方法被应用于各种社会网络当中[17]。在这些方法中，图模型被用来描述某一特定时间的网络快照。利用这些静态图序列就可以刻画网络的动态特性[18]。但是，这种方法忽略了社会网络中个体行为的随机性和突发性，而这些特性往往具有特殊的表现形式，并且会对最终演化结果产生一定的影响。简言之，基于这种“硬切分”的动态网络分析方法忽略了演化网络中两个重要的特性：噪声和事件。其中，噪声由具有社会化特征的个体行为的随机性和不确定性造成。这种个体行为往往持续时间短，并且不具有网络扩散特性。噪声如果得不到有效的处理，就有可能在演化分析过程中被放大，从而影响整个分析结果。而事件则是由个体或群体的异常行为所引起的，并且具有一定的持续时

间,往往也会具有扩散特性,从而造成局部或整个网络的异常性变化。

本研究所关注的是如何发现和利用个体行为模式在网络演化过程中所起的作用,发现信息传播和网络演化的内在驱动力。

1.2.2 社会网络信息传播模型

社会网络与传统的 Web 网络传播信息的模式存在着本质不同,传统的 Web 网络是以信息内容为主体进行传播;社会网络是以人(个体)为中心,依靠人与人之间的关联关系进行信息的传播。如今,人们在获取信息的时候更加关注信息的来源。这种获取信息的方式将关注的重心放在个体从谁那里获取信息,又会与谁分享信息。传统的传播行为,如计算机病毒在计算机网络上的蔓延[19]、传染病在人群中的流行[20]、谣言在社会中的扩散[21],都可以看作服从某种规律的传播行为。社会网络打破了传统传播方法,它利用人与人的关系改变人与信息的关系,反过来又用人与信息的关系影响人与人的关系。如何去描述社会网络中的传播行为,揭示它的特性,具有重要的理论和应用价值,是我们关注的焦点。为了进一步研究社会网络中节点的传播影响力和其中心性指标之间的关系,传统的方法通常采用传染病理论中的 SIR 模型,将用户节点分为传播节点、免疫节点、未感染节点,并定义传播规则:

(1)如果一个传播节点与一个未感染节点接触,则未感染节点会以概率 a 成为传播节点;

(2)如果一个传播节点与一个免疫节点接触,则传播节点会以概率 b 成为免疫节点;

(3)传播节点会以速度 v 变为免疫节点,无需与其他节点接触。

有些研究学者发现了社会网络上的传播行为与个体用户的行为特征存在着显著的联系,那些有影响力的用户个体能够极大地提升信息传播的速度和范围。如:Kwak 等人研究了 Twitter 的关注网络的拓扑特征[22],测量用户的关注数量、转发数量,以及被提到数量的分布特征,发现关注数量的分布极度不均匀,并且用户之间的交互大部分都是单向的,而且这三种衡量指标之间没有必然的联系。Weng 等人对 Twitter 中强影响力用户的发现有着类似的研究结果[23]。Bakshy 等人则研究了 Twitter 中转发的级联效应的分布,发现尽管拥有大量关注的用户更容易引发较大规模的转发,但是与转发规模却没有太大关系[24]。社会网络用户之间的连接强度也对信息传播有着影响,对 Facebook 数据的研究结果表明,尽管强连接本身具有更大的影响力,但是弱连接却传递了绝大多数的新信息,也就是说社会网络中的弱连接在信息传播的过程中起着更重要的作用。Wu 等人

以微博用户为关注对象，他们将 Twitter 用户分为普通用户和精英用户两类，将精英用户又划分为名人、博主、主要媒体和正式组织，并分析了不同类型用户之间的信息流动，发现尽管关注的分布集中在少数精英用户之上[25]，但是他们发出的信息并非直接传递给普通用户，而是要经过中间用户，也就是说要经历一个两级传播。

社会网络中信息传播研究的另一个角度关注信息的传播路径和方向。Sun 等人分析了信息传播链长度的影响[26]，认为与传统社会网络相比在线社会网络中信息传播链通常比较短，而且信息的传播范围与传播链的长度之间没有明显的关系。尽管在线社会网络中信息的传播路径相对容易获取，而完整的信息传播网络则很难获取。针对这个问题，Rodriguez 等人提出了一个追踪信息的传播路径，进而估计出传播网络的算法[27]，并用该算法研究了博客和新闻报道的传播网络。他们发现该网络具有一个核心——“外围结构”，信息主要从小部分核心媒体站点发出，这些站点之间通常被普通的站点连接成固定的环状。与他们的研究不同，以前的传播模型大多是基于规则网络的，且较少地考虑社会网络中个体的行为特征。但是社会网络上的传播行为确实与用户个体的行为存在着显著的联系。本研究针对该问题，利用复杂网络理论，对基于个体行为驱动的信息传播行为进行详细的理论建模和数值仿真研究。

1.2.3 社会网络演化模型

以前复杂网络的拓扑模型一般采用 ER 随机图理论来进行描述，ER 模型首先固定网络中的节点数，然后以某种概率在节点间建立连接，形成最终的仿真网络。尽管连接是随机设置的，但大部分节点的连接数目大致相同，节点度的分布服从钟形的泊松分布，随着连接数的增大，其概率呈指数式迅速递减，因此随机网络也称指数网络。由于缺乏大型的网络数据，ER 模型并没有真正应用到真实的网络拓扑模拟中。近年来对相关统计数据的分析表明，许多复杂网络并不是随机网络，具有不同于随机网络的统计特征。例如：WWW 中存在着很多高连通的节点，且度服从一定参数的幂律分布，Internet 的度分布也服从幂律分布[28]。这种度分布服从幂律分布的网络被称为无标度网络（Scale - Free Network）[29]。与随机网络相比，无标度网络具有一些重要的特性，如网络中可能存在度很大的节点，它们可以承受意外的故障，但面对协同式攻击却很脆弱等。幂律特性的发现对复杂网络的动力学、拓扑仿真、信息传播等方面的研究具有重要的意义。

真实系统通过自组织生成无标度的网络主要归功于两个因素：生长

(Growth)和优先附着(Preferential Attachment),理论及仿真证明二者缺一不可。根据这两个原则,给出了一个构造无标度网络的简单模型:简称 BA 模型,该模型初始时设立 m_0 个节点,然后在每一个时间步加入一个具有 m 个连接的新节点,该新节点按照某种概率分布选择网络中已有的 m 个节点,并与之建立连接。

1. 生长

开始于较少数量的节点(m_0),在每个时间步增加一个具有 $m(\leqslant m_0)$ 条边的新节点,连接这个新节点到 m 个已经存在于网络中的节点上。

2. 优先附着

在选择新节点的另一个端点时,新节点连接到节点 i 的概率 π 取决于节点 i 的度,即

$$\pi(k_i) = \frac{k_i}{\sum k_j}$$

经过 t 时间间隔后,该算法程序产生一个具有 $t + m_0$ 个节点,$m \times t$ 条边的网络,当 $t \to \infty$ 时,整个网络的平均连接度为 $2m$。且每个节点的连接度至少为 m。Barabási 和 Albert 利用平均场理论和仿真证明了网络中连接度为 k 的节点的概率服从幂指数为 -3 的幂律分布,该幂指数与 m 和 m_0 无关。BA 模型捕捉到了许多真实网络的幂律形成机制,但与真实网络相比,BA 模型有着明显的缺陷。真实网络在生长时,不会是每次增加固定的连接数,网络的平均连接度也不会是一个整数 $2m$。另一方面,许多实例表明在真实系统中节点的连接与增长率并不仅仅依赖节点进入网络的长短。例如,在社会系统中并不是每个人以同等的速率交朋友,而与其行为模式有紧密关系,有些颇有魅力的人会更容易交友。在万维网中,有些网站可以创建更好的内容吸引浏览者。通过“炒作”“广告”等行为可以使有些明星“一夜成名”,人气大涨,很容易超过那些比他们早进入娱乐圈的人。所有这些例子都说明了一个简单的道理,即网络的演化与节点的行为特征模式有着重要的联系。

2010 年有学者指出网络上的大量信息,如博客和论坛等,均是对现实社会的人及组织行为的映射,网络数据可用来分析个人和群体的行为模式,从而深化我们对生活、组织和社会的理解,并指出 3 个相互关联的问题:人群的交互方式、社会群体网络的形态及其演化规律[30]。这 3 个问题的研究可以帮助我们解答很多社会问题,如某个组织是达到了一个稳定的状态还是经常发生剧烈变化,具有创造力的团队应具有什么样的交互方式,目前社会的宏观网络结构是怎样的并将如何演变等。这种新兴理论对社会网络的研究和应用至关重要。

在复杂网络的研究中,各种各样的机理被提出来模拟真实网络的演化生长。从目前的研究现状来看,并不存在某种或某几种“放之四海皆准”的演化机理,与此相反,不同类型的网络可能有着完全不同的生长机理,即便它们表现出了非常相似的统计特性。举个例子来说,无标度网络结构的涌现就被公认为具有多种可能的原因。如:Kossinets 和 Watts 研究了在校大学生之间的熟人关系网络[31],发现两名原本没有社会关系的学生在将来是否会建立关系很大程度上受到他们当前共同熟人数目的影响。拥有越多共同熟人的学生在将来具有更大的可能性成为熟人。Liben-Nowell 和 Kleinberg 以及周涛等人研究了大量刻画节点接近性的定量指标,发现两个节点的共同邻居数目越多,它们之间存在直接连接的可能性就越大[32~33]。这些实证结果都暗示我们,两个节点建立连接行为更倾向于发生在具有共同邻居的节点中间。

因此研究社会网络中个体行为特征和真实演化过程的实证研究有助于我们了解网络生长可能的内在驱动力。寻找可能表征部分真实网络的演化规律并建立相应模型,一直是推动复杂网络演化模型研究的根本动力。

1.3 Enron 数据集

1.3.1 Enron 数据集介绍

Enron 数据集是第一个大规模真实的邮件通信数据,该数据最开始由联邦能源调查委员会(The Federal Energy Regulatory Commission,简称 FERC)开始收集,最初的数据集中的 92% 为 Enron 公司雇员的邮件,包括 158 个雇员的 619 449封电子邮件,FERC 版本的数据集存在着不完整性的问题。MIT 购买了这个数据集,并且 SRI 实验室的一组研究人员为了实施 CALO 项目[13]开始收集和完善这个数据集,SRI 研究人员修正了数据完整性问题,保证了数据的可用性。该数据集包括 Enorn 公司历时近 3 年的电子邮件数据,每封电子邮件按照 SMTP 协议格式的要求用文件的形式存储,文件中包括:发送者地址、接受者地址、发送时间、邮件主题、邮件正文等信息(不包括附件信息)。

本书使用 Enron 数据集是 2011 年 4 月份的版本,标记为“201104”。其中共包含517 425个文件,将文件中的群发邮件展开,共含有 3 520 792 个信息交互,平均每封邮件发送给 6.8 个用户。其中正常通信次数(利用 SMTP 协议中的“to”进行通信)2 982 658 个,抄送通信次数(利用 SMTP 协议中的“cc”进行通信)为 538 134 个。

1.3.2 Enron 数据集整理

Enron 数据量巨大,其中难免存在一些错误和容易引起混淆的数据,主要包括以下几种情况:

(1)时间错误:邮件发送/接受时间不符合实际情况,如发送时间为 1979 年;

(2)闭环数据:自己发送给自己的数据;

(3)与已知人员不相干的通信数据:发送双方都不是 Enron 公司中已经确认身份的人员;

(4)重复数据:有些邮件数据在不同的目录中重复出现,或者是一封邮件同时发送和抄送给同一个用户,本书利用每封邮件唯一的 Message ID 进行去重;

(5)用户 Micheal Swerzzbin(Vice President, mike. swerzbin@ enron. com)未在内部进行任何通信,因此将其从通信网络中去除。经过整理后从 158 个内部确认身份的员工中收集有效邮件通信共 49 005 个。其中正常通信次数 35 987,抄送次数 13 018。Enron 数据集中的邮件时间(SMTP 邮件服务器时间)是按照太平洋时间(PST,西八区时间)标注,安然公司总部位于美国德克萨斯州休斯顿市(西十区时间)。这里忽略用户在外地的情况,假设所有人员发送邮件时在德克萨斯州。因此在处理数据的时候将发送电子邮件的时间由西八区减两个小时,转换为西十区时间。

1.3.3 Enron 数据集人员信息整理

FERC 网站中提供了 Enron 中的 162 人的职位信息,通过在 Enron 数据集中 sent 和 sent_items 目录(通常为邮件客户端软件的发件箱,如:outlook, Foxmail 等)中查找发件人,并根据名字的相似性进行匹配,共确认 199 个邮箱地址(存在一个用户同时具有多个邮箱地址的情况),将邮箱地址和实际名字进行相似度计算,将满足“名”和“姓”最少一个与用户名匹配为匹配条件,共找到符合 Enron 公司 158 人邮件地址条件,并进行合并,如:用户 Vince Kaminski 可能的邮件地址包括 vince. kaminski@ enron. com, j. kaminski@ enron. com, kaminski@ enron. com, j. . kaminski@ enron. com,共 4 个。用户和邮件地址对应关系具体见附录一。Enron 数据集人员统计信息如表 1.1 所示。

表 1.1 Enron 数据集人员信息统计表

	Enron 20110402	整理后数据集	包含的邮件地址数
CEO	4	4	6
President	4	4	6
Vice President	28	25	31
Managing Director	5	5	9
Manger	13	13	18
Director	22	22	24
In House Lawyer	3	3	3
Trader	12	12	13
Employee	40	40	47
Director of Trading	1	1	1
N/A	29	29	41
总人数	161	158	199

本章参考文献

[1] MILGRAM S. The small world problem[J]. Psychology Today, 1967, 2(1): 60-67.

[2] WATTS D J, STROGATZ S H. Collective dynamics of "small world" networks [J]. Nature, 1998, 393(6684): 440-442.

[3] BARABSI A L, ALBERT R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512.

[4] TAN E, GUO L, CHEN S, et al. UNIK: unsupervised social network spam detection[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. ACM, 2013: 479-488.

[5] MOLAVI K A, KLIMAN-SILVER C, MISLOVE A. Iolaus: securing online content rating systems[C] //Proceedings of the 22nd International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 919-930.

[6] LIN W, KONG X, YU P S, et al. Community detection in incomplete

information networks[C]//Proceedings of the 21st International Conference on World Wide Web. ACM, 2012: 341-350.

[7] JUNG S, SEGEV A. Analyzing future communities in growing citation networks [C]//Proceedings of the 2013 International Workshop on Mining Unstructured Big Data Using Natural Language Processing. ACM, 2013: 15-22.

[8] KUO T T, YAN R, HUANGY Y, et al. Unsupervised link prediction using aggregative statistics on heterogeneous social networks[C]//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013: 775-783.

[9] BRONSTEIN J. Personal blogs as online presences on the internet: exploring self-presentation and self-disclosure in blogging [C]//Aslib Proceedings. Emerald Group Publishing Limited, 2013, 65(2): 4.

[10] DANESCU-NICULESCU-MIZIL C, WEST R, JURAFSKY D, et al. No country for old members: user lifecycle and linguistic change in online communities[C]//Proceedings of the 22nd International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013: 307-318.

[11] ROSSI R, GALLAGHER B, NEVILLE J, et al. Role-dynamics: fast mining of large dynamic networks[C]//Proceedings of the 21st International Conference Companion on World Wide Web. ACM, 2012: 997-1006.

[12] TANG J, WANG B, YANG Y, et al. Patentminer: topic-driven patent analysis and mining [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012: 1366-1374.

[13] LIU B, CONG G, XU D, et al. Time constrained influence maximization in social networks[C]//Data Mining (ICDM), 2012 IEEE 12th International Conference, 2012: 439-448.

[14] Lü L, MEDO M, YEUNGC H, et al. Recommender systems[J]. Physics Reports, 2012, 519(1): 1-49.

[15] TANG F Y, MAO C J, YU J H. The implementation of information service based on social network systems[C]// Proceedings of the 5th International Conference on New Trends in Information and Service Science. Macau, China, 2011: 46-49.

[16] VERBEKE W, DEJAEGER K, MARTENS D, et al. New insights into churn

prediction in the telecommunication sector: a profit driven data mining approach[J]. European Journal of Operational Research, 2012, 218(1): 211-229.

[17] GÖRKE R, MAILLARD P, SCHUMM A, et al. Dynamic graph clustering combining modularity and smoothness [J]. Journal of Experimental Algorithmics (JEA), 2013, 18(1): 1-5.

[18] WU L, YING X, WU X, et al. Examining spectral space of complex networks with positive and negative links[J]. International Journal of Social Network Mining, 2012, 1(1): 91-111.

[19] ZOU C C, TOWSLEY D, GONG W. E-mail virus propagation modeling and analysis[J]. Department of Electrical and Computer Engineering, Univ. Massachusetts, Amherst, Technical Report: TR-CSE-03-04, 2003.

[20] FAN W, YEUNG K H. Virus propagation modeling in Facebook[M]//The Influence of Technology on Social Network Analysis and Mining. Springer Vienna, 2013: 185-199.

[21] 高湘昀, 安海忠, 方伟. 基于复杂网络的时间序列双变量相关性波动研究[J]. 物理学报, 2012, 61(9): 098902.

[22] KWAK H, LEE C, PARK H, et al. What is twitter, a social network or a news media? [C]//Proceedings of the 19th International Conference on World Wide Web. ACM, 2010: 591-600.

[23] WENG J, LIM E P, JIANG J, et al. Twitter rank: finding topic-sensitive influential twitterers [C]//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. ACM, 2010: 261-270.

[24] BAKSHY E, ROSENN I, MARLOW C, et al. The role of social networks in information diffusion[C] //Proceedings of the 21st International Conference on World Wide Web. New York: ACM Press, 2012: 519-528.

[25] WU SHAO-MEI, HOFMAN J M, MASONW A, et al. Who says what to whom on twitter[C] //Proceedings of the 20th International Conference on World Wide Web. New York: ACM Press, 2011.

[26] SUN E, ROSENN I, MARLOW C A, et al. Gesundheit! modeling contagion through Facebook news feed[C] //Proceedings of the 3rd International AAAI Conference on Web Logs and Social Media, 2009.

[27] RODRIGUEZ M G, LESKOVEC J, KRAUSE A. Inferring networks of diffusion and influence [C] //Proceedings of the 16th ACMSIGKDD

International Conference on Knowledge Discovery and Data Mining. Washington D C: ACM Press, 2010.

[28] MISLOVE A, KOPPULA H S, GUMMADI K P, et al. An empirical validation of growth models for complex networks[M]//Dynamics on and of Complex Networks, Springer New York, 2013,2: 19-40.

[29] WANG L, LIU J. A scale-free based memetic algorithm for resource-constrained project scheduling problems[M]//Intelligent Data Engineering and Automated Learning-IDEAL 2013. Springer Berlin Heidelberg, 2013: 202-209.

[30] CIOFFI- REVILLA C. Computational social science [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(3): 259-271.

[31] KOSSINETS G, WATTS D J. Empirical analysis of an evolving social network [J]. Science, 2006, 311(5757): 88-90.

[32] LIBEN-NOWELL D, KLEINBERG J. The link-prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.

[33] ZHOU T, LU L Y, ZHANG Y C. Predicting missing links via local information [J]. The European Physical Journal B, 2009, 71(4): 623-630.