在线社交网络影响力分析

吴信东[1),2)] 李 毅[1),3)] 李 磊[1)]

[1)](合肥工业大学计算机与信息学院 合肥 230009)

[2)](佛蒙特大学计算机科学系 伯灵顿 VT 05405 美国)

[3)](北方民族大学计算机科学与工程学院 银川 750021)

摘 要 社交影响力分析是社交网络分析的关键问题之一.近十几年间,随着在线社交网络的蓬勃发展,研究人员才开始有机会在大量现实数据的基础上对社交影响力进行建模和分析,并取得了丰硕的研究成果和广泛的应用价值.文中回顾了近些年在线社交网络影响力分析的主要成果,阐述了社交影响力的相关概念和它们之间的关系,重点从网络拓扑、用户行为和交互信息等几个方面总结了影响力分析的建模和度量方法,介绍了与影响力传播密切相关的意见领袖发现和影响力最大传播问题的研究现状,并对在线社交网络影响力分析的前景进行了展望.

关键词 在线社交网络;社交影响力;方法和模型;影响力最大传播;意见领袖发现;社会计算

中图法分类号 TP399 **DOI 号** 10.3724/SP.J.1016.2014.00735

Influence Analysis of Online Social Networks

WU Xin-Dong[1),2)] LI Yi[1),3)] LI Lei[1)]

[1)](*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei* 230009)

[2)](*Department of Computer Science, University of Vermont, Burlington VT* 05405, *USA*)

[3)](*School of Computer Science and Technology, Beifang University of Nationalities, Yinchuan* 750021)

Abstract Social influence analysis is one of the major research focuses in Social Network Analysis. Over the past decade, benefiting from the rapid development of online social networks, researchers began to have the opportunity to model and analyze the social influence based on the vast amount of realistic data, and they have achieved fruitful research findings with great application values. This paper reviews important achievements on influence analysis of online social networks in recent years. First, the concept of social influence is elaborated with comparison to related definitions. Second, we discuss modeling and measuring for social influence analysis with respect to network topology, user behavior and interaction information. Third, two important problems, discovering opinion leaders and maximizing the spread of social influence are introduced in a close context to social influence spreading. Finally, we conclude the paper with an exploration of future research directions on influence analysis of online social networks.

Keywords online social networks; social influence; methods and models; maximal influence spreading; opinion leader discovery; social computing

---

收稿日期:2013-06-19;最终修改稿收到日期:2014-01-22.本课题得到国家"九七三"重点基础研究发展规划项目"社交网络分析与网络信息传播的基础研究"(2013CB329604)和教育部创新团队项目"多源海量动态信息处理"(IRT13059)资助.**吴信东**,男,1963 年生,博士,教授,博士生导师,IEEE Fellow,AAAS Fellow,主要研究领域为数据挖掘、大数据分析、知识库系统、万维网信息开采.E-mail:xwu@hfut.edu.cn.**李 毅**,男,1981 年生,博士研究生,讲师,主要研究方向为数据挖掘、社交网络分析、智能信息处理.**李 磊**,男,1981 年生,博士,副教授,主要研究方向为数据挖掘、社会计算、信任计算.

1 引 言

影响力分析是社交网络分析的重要内容[1]. 社交影响力可以通过用户之间的社交活动体现出来，表现为用户的行为和思想等受他人影响发生改变的现象. 自从 20 世纪 50 年代 Katz 和 Lazarsfeld[2] 发现社交影响力在社会生活和决策制定等方面发挥重要作用至今，影响力分析在多个领域得到广泛应用，比如推荐系统[3]、社交网络信息传播[4-8]、链路预测[9-11]、病毒式营销[12-18]、公共健康[19-20]、专家发现[21-22]、突发事件检测[23] 和广告投放[24] 等. Domingos 和 Richardson[15,17] 对用户影响力及其在营销网络中的传播规律首先展开探索，其后计算机学界投入大量工作到相关问题的研究之中. 人们发现具有广泛影响力的用户，他们在创新采用、社会舆论传播和导向、群体行为形成和发展等方面具有重要作用[2,25-26]，而且通过口口相传(Word of Mouth)的影响力传播方式，商业营销能以较低费用将新产品推广到整个社交网络，从而产生较大的社会影响和商业价值[18]. 此类问题一直受到研究人员的关注.

早期工作对影响力在社会活动中的表现和相关因素进行了探索和分析，对社交影响力的作用模式和产生机制进行了深入研究，发现了很多与影响力相关的社会现象及其深层原理. 但是当时的研究样本空间较小，能够获取的数据量有限，亟需大量客观数据的支持和验证[27].

随着 Web2.0 技术的迅速发展和在线社交网络的迅速普及[28]，研究人员首次有机会在海量交互信息和大规模社交网络中综合分析和研究用户之间的各种复杂关系. 影响力研究转而以在线社交网络产生的丰富数据为支持，建立各种影响力分析模型和广泛使用多种量化技术[29]，对用户本身体现出来的影响力、用户在线交互过程中表现的彼此影响、用户及其所在社团之间的相互作用以及影响力随时间的演化等诸多问题进行了研究和探讨，既验证和扩展了早期的很多假设和理论模型，同时也观察到更多有趣的现象和规律，而且在新的研究环境下发现了大量与社交影响力相关的科研问题和应用场景[30]. 如异构网络中的影响力分析[31-32]、资源受限的影响力传播[33-34]、隐性影响力度量[35-36]、群体影响力分析[37] 等问题都是当前的研究热点.

总体而言，在线社交影响力分析主要涉及 3 方面的内容：(1) 影响力自身的识别. 由于影响力概念本身具有较强的因果性，因此如何从繁杂的因素中鉴别影响力和相关要素的区别与联系，就成为首当其冲的问题；(2) 社交影响力的度量. 在对社交影响力定性识别的基础之上，面对复杂多变的社交关系，如何设计和选择既具有一定普适性，同时又能充分发掘社交网络特性的度量方法，是该领域的核心问题之一；(3) 社交影响力的动态传播. 人们之间形成的社会关系和社交影响力在频繁地产生变化，而在线社交网络又加剧了这种变化的速度和范围，因此研究社交影响力的动态特性，对分析社交网络演化、社会行为特征、信息传播模式等诸多问题都有重要价值.

在线社交网络的影响力分析涉及到大量复杂因素. Sun 和 Tang[29] 介绍了基于社交网络的边和节点的影响力度量模型，为了更全面地总结社交影响力度量的方法及近几年的研究成果，本文从社交影响力的基本概念出发，分析和对比了社交影响力的定义、度量及传播的重要方法和成果，并对相关领域的研究方向和问题进行了有益的探讨. 本文第 2 节介绍社交影响力的概念及相关因素；第 3 节从社交网络的结构、用户动作和交互信息三方面分析近些年社交影响力的模型和度量方法；第 4 节介绍与影响力传播相关的意见领袖发现和影响力最大传播问题的相关研究成果；第 5 节对本文内容进行总结并探讨未来的研究方向.

2 社交影响力的相关因素

2.1 影响力(Influence)

社交影响力只有通过人们之间的交互活动才能够体现出来，比如用户 A 在网上的发帖吸引了用户 B，使得后者成为 A 的粉丝，即 A 对 B 产生了影响力. 由于社交影响力的研究工作涉及众多学科和领域，对该术语的定义也有诸多版本[4-5,25-26,38-42]. 例如：

(1) Rashotte[38] 根据用户行为及其产生的效果，将社交影响力定义成人们由于和其他人或团体之间的交互而改变自身思想、感情、态度和行为的现象；

(2) Watts 和 Dodds[42] 在前人工作基础上，利用社交影响力的统计学特点，把影响力分布中度量值较大的 $q\%$ 的用户定义成有影响力的人；

(3) Cha 等人[5] 根据用户在社交网络上的追随行为、转发信息和谈论用户的行为，分别划分出 3 种

影响力；

(4) Bakshy 等人[4]依据交互信息在社交网络上的传播特点，利用信息的转发次序对用户影响力进行打分.

社交影响力的定义具有明显的因果性，而人们的思想、行为等产生变化的原因则是不胜枚举且因人而异，社交影响力只是其中之一. 这就给社交影响力的建模和度量带来了很大挑战，同时也是造成社交影响力模型众多的重要原因. 同样，在线社交网络中的影响力也与很多因素相关，目前大部分研究工作都是针对社交网络结构及其上的交互信息和用户行为特征进行量化和分析的，因此可以把能对信息传播过程或他人行为产生影响的个体视为具有社交影响力.

2.2　同质性(Homophily)

同质性指具有相似特征的个体选择彼此作为朋友的倾向[43-44]，即所谓的“物以类聚，人以群分”. 如果用户 B 发现用户 A 和自己有相同的兴趣和爱好，因而选择 A 作为自己的朋友，这种现象即是同质性的表现.

社会生活中有大量证据能够表明同质性的普遍存在[44]，比如具有相同宗教信仰的人之间更容易结成至交；相同性别的人遵循同样的交友模式；相近年龄的学生倾向于相互来往；相同地域、教育背景、职业或健康状态的人容易形成朋友.

仅从概念上就可以发现同质性和影响力具有较强的关联，而对这二者的鉴别向来是社交影响力分析和建模的关键问题之一[10,45-51]. 研究人员从 Facebook 中选取 130 多万用户作为分析对象，发现具有影响力的用户更倾向于在彼此之间形成社团，表明社交网络中存在明显的同质性现象[52]. 也有研究发现具有相同影视音乐爱好的学生倾向于结成好友，而结交后的学生却很少改变自己的特性[45]；虽然具有共同话题和兴趣的用户能够通过频繁交流产生较强的相互影响，但是，即便是交流不多的朋友之间，彼此的影响力依然很大[50]. 同质性和影响力在社交网络中的传播模式也非常相似，都表现为作用对象的相互联系趋于紧密，二者最大的区别体现在动态效应上，即影响力需要更长时间的复杂交互活动才能发挥显性效果. 为此 Aral 等人[46]设计了动态匹配的取样估计方法，在动态网络上对影响力和同质性进行了鉴别，发现由于同质性的存在，之前的研究可能高估了影响力在社交活动中的作用. 由于影响力和同质性彼此之间存在较强的反馈效应[10,51]，所以在没有很强假设条件的限制下，统计方法没有办法完全控制观察结果中的混杂因子，即同质性和影响力是天生交织在一起的[47]. 因此，在实际应用环境中将影响力和同质性结合起来使用，既可以提高彼此的预测精度[49]，也可以提高用户行为预测的准确性[10].

2.3　互惠性等其他因素

社交影响力除了受到同质性的影响，还有很多因素都会对影响力分析结果产生影响，比如互惠性[5,7,53-55](Reciprocity)、活跃度[56](Activity)、同时性[57](Simultaneity)、异质性[58](Heterogeneity)、环境因素和关联效应[59](Contextual and Correlated Effects)等.

上述因素中的互惠性指的是用户在社交过程中出于礼貌或习惯等原因对其他用户的行为给予相应回应的现象. 例如在社交网络中，用户 B 申请作为用户 A 的好友，A 在接受请求后出于礼貌又申请作为 B 的好友，这种投桃报李的现象就体现出互惠性. 显然，B 很可能受到 A 的影响而成为他的朋友，但 A 的行为明显不是社交影响力的结果. 研究表明 Flickr 和 Yahoo!、360 等网站中互惠性较明显[54-55]，而在 Twitter 中则要相对低很多[5,53]. 即使在同一社交网络中，互惠性的测算结果也会有较大差别[5,7]. 由于用户行为的结果相同而原因迥异，互惠性、影响力和前面所述的同质性之间的区分依然是影响力分析的热点问题. 而且，目前的工作主要聚焦在上述三者的差异性分析上，对它们的内在联系及其在演化过程中的作用规律所知甚少，这类问题尚有待相关研究者进一步的分析和探索.

为了获取高影响力，除了要发布有价值的信息和受到其他用户的关注等因素之外，用户还需要克服自身的消极性[6]. 不是被动接受社交网络中的各种信息然后再将其转发出去，而要主动参与信息的传播和加工，扩大自身的影响力.

社交网络的外部因素也会对影响力和信息传播产生作用. 例如，研究发现 YouTube 上的有些视频信息能以很快的速度传播开来，而这类现象可以用视频信息受到的外部影响进行解释[60]. Myers 等人[61]利用曝光曲线[62]对社交网络上的外部影响力进行了建模，认为随机出现在节点上，以“跳跃”形式分布的信息受外部影响力驱动，分析结果显示 Twitter 上有近三成的信息受外部因素影响. 文献[63]中则直接把外部影响力作为社交网络的外部连接引入模型中，称之为 ε-边，即与这类边相连的用户受到权重

为 ε 的外部影响力的作用.

在影响力分析时大家普遍接受的观点是用户的影响力促进了意见、行为、创新和产品的传播.虽然该观点比较流行,也得到了广泛研究和实验支持,但是 Watts 等人[42,56,64-65]对此假说提出了质疑.后续研究也有观点认为是敏感性(即个体受他人影响的趋势)而不是影响力在社会传播中起着关键作用[52],并且发现 Facebook 上的用户的影响力和敏感性呈此消彼长的关系,几乎没有同时兼具高影响力和高敏感性的用户,而这两种因素在用户行为传播时都会发挥作用.

社交影响力及其传播过程受诸多要素的制约,而且不少要素之间是紧密联系,相互依存的.如果要详细区分这些要素,势必会引入一系列参数或者协变量,而且在有些情况下根据有限的社交现象根本就无法对它们进行准确区分.比如在社交网络上的交友现象,两个人选择成为朋友,有可能是通过各种信息交流产生相互影响的结果,也有可能他们本身就具有很多共同点,比如有相同爱好、相似的习惯或者信仰等同质性特征,也有可能是其他更为复杂的外界因素使得两人成为朋友,还有可能是上述因素共同作用产生的结果.目前的影响力模型大都以上述要素的子集合为主进行构建,为模型设计和计算带来便利的同时,造成模型通用性不强,而全面考虑上述要素又会使得模型过于复杂或者根本无法实现.所以,在设计和完善现有影响力模型的过程中,既要全面考虑各种要素的作用和权重,使模型尽可能符合现实世界的真实情况,同时还有赖于对非独立要素相关性的分析和研究,找出它们之间的内在联系和变化规律,以减少模型中的参数数量,降低模型复杂度.

3 社交网络影响力的度量方法

社交网络中影响力度量的主要任务是分析和预测用户社交影响力的大小及演化规律,为基于社交影响力的研究和应用提供技术支持和理论依据.常用的影响力度量方法大致可以划分为基于网络拓扑结构、基于用户行为和基于交互信息的度量等类型.社交网络拓扑是用户在社交活动中残留下的"遗迹",从网络拓扑学角度体现了用户影响力的特征,而且社交网络结构的获取比较简单,基于其上的度量方法相对成熟,计算量较小.但是,网络拓扑无法刻画用户之间频繁的交互活动,而用户在社交活动中的行为变化能够更准确地反映用户社交影响力的产生和演变情况,用户的交互信息则能够进一步体现影响力产生及演化的细节.所以,在进行社交影响力分析时,既需要根据实际情况选择合适的度量手段,还可以综合使用上述方法,尽可能准确客观地刻画社交影响力的真实面貌.

3.1 基于网络拓扑结构的度量

人们在社交过程中形成的拓扑结构是分析影响力最直接的数据来源,同时相比社交信息更容易观察和获取,因此在影响力分析早期就得到社会学家和其他领域专家的充分研究[1].对社交网络进行影响力分析时,网络结构中的节点表示用户,而节点之间的连接则表示用户之间建立的关系,它们在分析影响力时都起着至关重要的作用,因此我们分别以网络节点和连接为对象介绍一些广泛使用的度量方法.该类方法和模型更全面的介绍可参阅文献[29].

在下文中,社交网络的拓扑结构用图 $G=(V,E)$ 表示,如无特别说明,一般指无向图.其中 $n=|V|$ 表示节点数;v_i 表示节点 i;e_{ij} 表示节点 i 和 j 之间的边,很多时候也被称之为连接;$\mathbf{A}_{n\times n}$ 表示图的邻接矩阵;$a_{i,j}$ 是其中的元素;如果是带权图,我们用 $w_{i,j}$ 表示节点 i 和 j 之间的权重.

3.1.1 节点的度量

节点中心度主要用于衡量网络中节点的重要程度,并且被广泛应用于以节点为对象的分析技术中[66-67].下面以中心度度量为重点介绍一些常用的度量方法及其特点,如表 1 所示.

节点的出度和入度可以衡量社交网络中与用户影响力相关的指标(如好友数、推荐数、跟帖数等),在一定程度上可表示节点的影响力大小,而它们的方向可以表示用户影响力或者信息的传播方向.度中心度[68](Degree Centrality)则可以用来衡量节点对其邻居的平均影响力.

基于社交网络上最短路径的方法有紧密中心度[69-70](Closeness Centrality)和介数中心度[68,71-72](Betweenness Centrality)等方法.紧密中心度可用来度量当前节点对其他节点的间接影响力,或者信息从该节点传播到其他节点的距离,也可间接度量该用户的社会关系强度.该值越大,表示当前用户和其他用户之间的距离越短,该用户影响其他用户的速度越快.介数中心度衡量节点在网络结构中所处位置的重要性.该度量值越大,表示网络中信息流动时经过该节点的信息量越大,即该节点在信息传播过程中的影响力越大.

表 1　节点影响力的度量方法

类型	度量方法	计算公式	度量方法描述
基于节点度	入度	$deg^{\text{in}}(v_i)=\sum_j a_{j,i}$	当前节点对邻居节点的影响力
	出度	$deg^{\text{out}}(v_i)=\sum_j a_{i,j}$	邻居节点对当前节点的影响力
	度中心度	$C^{\text{DEG}}(v_i)=\frac{deg(v_i)}{n-1}$	当前节点与邻居节点间的平均影响力
基于最短路径	紧密中心度	$C^{\text{CLO}}(v_i)=\frac{1}{\sum_{v_j\in V\setminus v_i} g'_{ij}}$	当前节点到网络中其他节点的距离. 其中 g'_{ij} 表示节点 i 到 j 的最短路径长度.
	介数中心度	$C^{\text{BET}}(v_i)=\frac{\sum_{s<t}\|\{g^i_{st}\}\|/\|\{g_{st}\}\|}{n(n-1)/2}$	当前节点在网络中所处位置的重要程度. 其中 $\|\{g_{st}\}\|$ 表示节点 s 和 t 之间的最短路径(也称为测地线)个数，$\|\{g^i_{st}\}\|$ 表示上述最短路径中经过节点 i 的个数.
基于随机游走	特征向量中心度	$\lambda x_i=\sum_{j=1}^{n} a_{i,j}x_j,\ i=1,2,\cdots,n$	当前节点的影响力取决于邻接节点影响力的线性和. 其中 λ 表示邻接矩阵的最大特征值.
	Katz 中心度	$C^{\text{Katz}}(v_i)=\sum_{k=1}^{\infty}\sum_{j=1}^{n}\alpha^k(A^k)_{ij}$	当前节点的影响力由从该节点出发的随机游走路径所决定. 其中 α 为惩罚因子.
	PageRank 度量	$Pr(v_i)=\frac{1-d}{n}+d\sum_{v_j\in L^{\text{in}}(v_i)}\frac{Pr(v_j)}{\|L^{\text{out}}(v_j)\|}$	当前节点的影响力排名. 其中 $L^{\text{in}}(v_i)$ 表示链入 v_i 的节点，$L^{\text{out}}(v_i)$ 表示链出 v_i 的节点，d 为阻尼因子.
	局部聚集系数	$C^{\text{CLU}}(v_i)=\frac{\|\{e_{jk}:v_j,v_k\in Ng_i,e_{jk}\in E\}\|}{\|Ng_i\|(\|Ng_i\|-1)/2}$	当前节点的邻居节点相互影响的程度大小. 其中 Ng_i 是节点 i 的邻居节点集合.

利用随机游走特征度量影响力的指标有特征向量中心度[73-74](Eigenvector Centrality)、Katz 中心度[75]和 PageRank 度量[7,76-78]等. 特征向量中心度根据节点的中心性计算它们的权重，然后将当前节点可达的其他节点的权重的线性和视为该节点的中心度数值. 该度量反映的思想是，节点的影响力是与和它相连的其他节点的影响力相关的，亦即与之相连的节点影响力越大，该节点在网络中的影响力随之增大，反之亦然. Katz 中心度利用两个节点间的游走路径来计算二者之间的影响力. 在游走路径上距离 v_i 越远的节点，通过惩罚因子 α 的作用，对节点 v_i 的 Katz 中心度贡献越小. 与 Katz 中心度类似的还有 Bonacich 中心度[79]和 Hubbell 中心度[80]. 特征向量中心度和 Katz 中心度之间存在密切的关系，研究证明当惩罚因子 α 趋近某个特定值时，邻接矩阵 $\boldsymbol{A}$ 的特征向量是 Katz 度量的极限[81]. PageRank 度量是在有向图中计算节点排名的指标，如果把用户影响力的传播过程看成随机游走，则该度量也可用来表示用户影响力的大小.

社交网络中的用户相互之间紧密联系，从而形成社团的趋势较强，聚集系数[82-83](Clustering Coefficient)可以用来度量这种趋势的大小. 局部聚集系数表示节点 v_i 的任意两个邻居 v_j 和 v_k 之间产生联系的可能性，比如用户 A 和用户 B、C 是朋友关系，该中心度可以测算 B 和 C 之间同样是朋友关系的概率.

众所周知，节点的度本质上就表示某节点和邻居节点之间的关联程度，而且基于节点的度的方法表达的意义直观，计算代价很小，所以被广泛应用于对社交网络中用户影响力的度量. 但是，基于度的方法只能反映用户及其邻居之间的联系，是对用户局部影响力的度量，无法很好地衡量用户在整个社交网络中的影响力. 如果用户影响力是通过节点之间的最短路径发挥作用的，则可利用基于最短路径的方法对用户影响力的作用距离进行度量，从而间接表示用户影响力的强弱. 相比基于度的方法，基于最短路径的方法能够从社交网络整体对用户影响力进行度量，但是它的计算复杂度比前者高，而且用户的影响力通过最短路径发挥作用是一种理想状态，在现实环境中很难实现. 基于随机游走的方法可以利用随机游走路径上的节点衡量当前节点的影响力，即用户 v_i 的影响力与游走路径上的其他节点的影响力相关，其他节点的影响力越大，则 v_i 的影响力越大，反之亦然.

3.1.2　连接的度量

对连接的影响力度量即是对两个用户相互之间影响程度的度量. 在早期的研究工作之中，很多分析信息传播和影响力扩散的模型出于简化模型或者计算所需，对表示用户间影响力大小的连接权重赋予经验值：有的是常数，有些则假设影响力服从某种分

布[63,84-86]. 显然,这类方法无法准确描述影响力的真实情况.

一般而言,两个节点的邻居重叠程度越高,这两个节点之间的关系越紧密,它们之间的影响力也越强烈[87]. 可以用 Jaccard 相似度度量这种关系:

$$J(i,j)=\frac{|Ng_i \cap Ng_j|}{|Ng_i \cup Ng_j|},$$

其中,Ng_i 表示节点 v_i 的邻居节点集合. Jaccard 相似度用于统计节点 i 和 j 的共同邻居在总邻居数中所占比例. 与 Jaccard 相似度相似,还可以用 Overlap 相似度和 Cosine 相似度等计算连接上的影响力[10].

与节点的介数中心度类似,边介数[88](Edge Betweenness)也可用于度量边在网络中的重要程度:

$$E^{\mathrm{BET}}(e_{ij})=\sum_{s<t}|g_{st}^{ij}|,$$

其中,$|g_{st}^{ij}|$ 表示节点 s 和 t 之间的最短路径同时经过节点 i 和 j(即通过边 e_{ij})的个数,边介数统计的是网络中经过边 e_{ij} 的最短路径的总数量.

文献[78]在分析博客空间的影响力传播问题时,用有向多重图表示节点间的影响力,弧的重数越多代表节点间的影响力越强,弧的方向表示影响力的作用方向. 随后作者提出一种影响力图用于刻画上述关系(如图 1 所示),该图是有向带权图,弧的方向表示影响力来源,权重代表影响力强度,用 $c_{u,v}$ 表示从节点 u 到节点 v 之间的平行边条数,其计算公式为

$$w_{u,v}=\frac{c_{u,v}}{deg^{\mathrm{in}}(v)}.$$

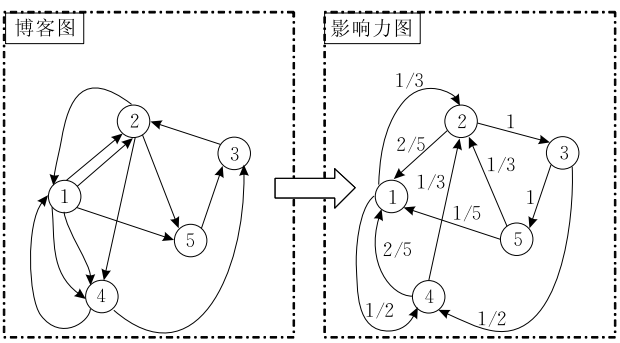

图 1 社交网络的影响力图[78]

综上所述,基于网络拓扑结构的度量是社交影响力最基本的度量方法. 这种方法抛开网络上海量的交互信息,抽取网络结构对用户影响力进行分析和建模,具有模型简单、易于应用和扩展的特点,在社交网络规模较大或者信息有限的情况下具有一定优势;同时还横跨了图论、概率统计、社会学和物理学等多个学科,具有坚实的理论基础.

但是,基于网络拓扑的影响力度量也存在一些先天缺陷:首先,研究人员获取的社交网络拓扑都是静态的,相当于原网络的一个快照,其上记录了从该社交网络诞生到被采集为止的所有显性社交关系,即十年前和一秒钟之前建立的连接被同时采集,只接收过一条公告的连接和两个朋友之间热烈交流的连接在计算模型中同等对待;其次,在这样的网络拓扑上,所有连接的权重都是相等或者同分布的,意味着但凡有连接的用户彼此之间具有相同的影响力,或者社交网络中用户之间的影响力满足简单的概率函数. 上述情形显然和实际情况不符,而造成这些缺陷的根本原因在于基于网络拓扑的方法对用户之间的行为和交互信息利用较少,从而导致该方法的度量结果和实际情况有偏差.

3.2 基于用户行为的度量

在线社交网络中的用户行为(也称动作)有发布信息、购买商品、话题评论、转发信息、建立好友关系等. 通过分析这些行为的分布规律和因果关系,就能够评估行为的发起者和传播者之间的影响力,还可以预测人们在社交网络上的行为,加深我们对人类社交行为的认识和理解[5,48,50,89-91].

一般情况下,在线社交网络都会记录人们通过交互活动产生的大量信息,其中就包括各种用户行为数据. 通过分析这些数据,可以衡量用户之间的影响力大小及其传播途径和范围,还可以据此建立用户之间的社交关系网络. 网络日志是一种重要的用户行为数据来源,Goyal 等人[48]利用日志信息分别计算了用户和动作自身的影响力:

$$infl(u)=\frac{|\{a|\exists v,\Delta t: prop(a,v,u,\Delta t)\wedge 0\le\Delta t\le\tau_{v,u}\}|}{A_u},$$

$$infl(a)=\frac{|\{u|\exists v,\Delta t: prop(a,v,u,\Delta t)\wedge 0\le\Delta t\le\tau_{v,u}\}|}{U(a)},$$

其中,u 和 v 表示不同用户;a 表示动作;Δt 表示动作之间的时间间隔;$\tau_{v,u}$ 是时间常量;$prop(a,v,u,\Delta t)$ 表示动作在用户之间的传播;A_u 表示用户 u 产生的动作数量;$U(a)$ 表示执行动作 a 的用户数量. 与基于网络拓扑的节点影响力度量方法不同,上述模型以动作的传播频率作为用户影响力的评估指标,并用动作的执行范围度量动作本身的影响力指标. 为了计算用户对其邻居的影响力,他们分别设计了静态概率模型、连续时间模型和离散时间模型,用机器

学习方法对相关参数进行了估计，并计算了 Flickr 上用户之间的影响力系数，而在此之前这些系数在大部分工作中都取经验值.

Saito 等人[92-93]在研究信息传播的独立级联模型[16,94-95]时探讨了类似问题，将用户影响力模型转化成一种最大似然问题，并且利用期望最大化[96](Expectation Maximization，EM)算法进行了求解. 因为该模型在 EM 算法的每次迭代过程都要对每条连接上的影响力系数进行计算，时间复杂度较高，所以并不适合于大规模社交网络中的影响力度量.

上述方法均需要利用社交网络结构才可以计算用户之间的影响力，而 Yang 和 Leskovec[8]则认为信息的传播过程受用户影响力的控制，与显性的网络拓扑以及用户间的相互连接没有必然联系. 他们建立了一种线性影响力模型 LIM(Linear Influence Model)，用于表示用户影响力与过去已经受到影响的其他用户之间的关系. 之后利用人们谈论信息的行为，在影响力函数和信息的谈论次数之间建立起联系，用于影响力的度量：

$$V(t+1)=\sum_{u\in A(t)} I_u(t-t_u),$$

其中，$V(t)$表示信息在 t 时刻被提到的次数；$I_u(l)$表示用户 u 的影响力函数，即用户 u 采用某信息后，其他人在 l 时刻提到该信息的次数；$A(t)$表示截止 t 时刻已经受到影响的用户集合. $I_u(l)$一般被设定成特定的参数方程，比如指数函数 $I_u(l)=c_u e^{-\lambda_u l}$ 或幂率函数 $I_u(l)=c_u l^{-\lambda_u}$. 为了更符合实际情况，他们还提出一种无参化方法，将影响力表示成随时间流逝递减的函数. 这些模型的度量效果与 $I_u(l)$函数关系密切，但是现有研究表明与社交影响力度量的相关因素很多，而且影响力的分布和变化规律在很大程度上仍旧是未解之谜，因此很难用特定函数对其进行统一表示.

除了发布、转发和评论信息等频繁发生的行为之外，其他行为也可作为用户影响力的评价指标. 比如用户登录社交网络的频率也可以判断他的影响力[90]，如果某用户登录频繁，而且跟他有联系的用户数量随之增长，则表明该用户有影响力. 文献[97]提出一种 NTT-FGM(Noise Tolerant Time-varying Factor Graph Model)模型，综合使用网络结构、用户属性和历史行为数据，利用朋友对用户行为的影响力预测当前时刻的用户行为.

上述基于用户行为的影响力度量方法的特点如表 2 所示. 其中连续模型能更细致地刻画影响力的传播过程，相比离散模型具有更强的预测能力，但其计算代价过高，而离散模型则可以比较高效地用各种启发式方法进行求解；无参模型对社交网络的约束较少，但描述真实环境的能力有限，适用于理论分析，而参数化模型刻画真实世界的能力较强，但其预测精度有赖于模型中参数的具体取值. 现有工作很少考虑用户行为自身的影响力，而是将它们平等对待，在模型中赋予相等的权重，这显然与实际情形不符. 比如在真实的社交环境中，用户申请作为别人的好友这一行为，相比转发一条此人信息的行为，其中所包含的社交影响力理应不同. 因此，对于社交网络中的各种社交行为在影响力计算中所起的作用，它们在影响力传播过程中是否遵循相同的传播模式，以及不同的社交行为对影响力预测精度的贡献和制约等问题，还亟需进一步分析和研究.

表 2　基于用户行为的影响力度量方法

模型	网络结构	离散/连续	参数化	动作权重
LIM	不需要	离散	否	无
Goyal 等	需要	离散/连续	是	有
Saito 等	需要	离散	是	无
NTT-FGM	需要	离散	是	有

此外，基于用户行为的影响力度量方法及其实验结果，明确反映出仅限于社交网络结构的度量方法的局限性. Kwak 等人[53]在分析有影响力的 Twitter 时，分别利用基于网络拓扑的 PageRank 算法和追随者数量计算了用户的影响力，随后又利用用户的转推(Retweet)行为重新计算，结果发现前两种方法的计算结果比较相近，而和基于转推动作计算出来的数值之间存在明显的差异，结果如图 2 所示. 文献[5]综合使用社交网络的拓扑和用户交互行为，分别使用入度、转推和谈论(Mentions)度量用户影响力的大小，同样得到了不同的度量结果，并且发现入度只能够衡量用户的流行程度，转推和谈论则能够代表用户发布信息的影响力价值.

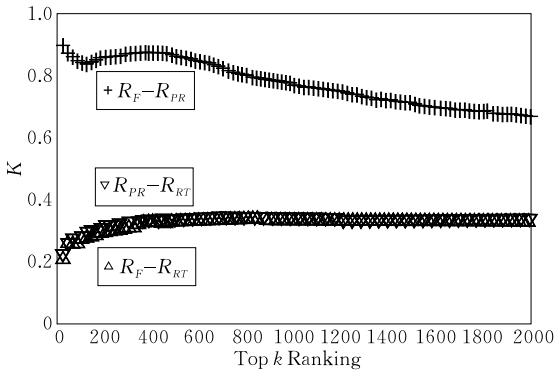

图 2　影响力排名结果示意图[53]

虽然基于用户行为的方法比单纯基于网络结构的方法功能更强,预测精度更高,但是由于隐私保护等原因往往不能获得完整的用户行为数据.因此,利用有限的用户行为数据提高基于网络结构的影响力模型的度量效果和精度也是今后值得研究的问题.所以,目前在进行社交网络影响力分析时,研究人员一般都会综合考虑使用包括网络拓扑、用户行为和用户交互信息在内的多种数据进行建模,以期尽可能提高影响力度量和预测的精度及准确性.

3.3 基于用户交互信息的度量

与现实世界中通过相同地域、共同活动、亲属关系等因素构成的社交网络不同,在线社交网络上的用户主要通过信息发布、共享、评论和转发等方式进行交流从而产生联系.因此,在线社交网络中的信息承载着用户网上活动的所有记录,在影响力分析时起着不可或缺的重要作用.但是在线社交网络上产生的数据量极大,要迅速分析这些数据并获取用户的实时影响力不太现实,所以研究人员一般都是采集部分数据样本进行分析,而且在分析过程中针对的都是交互信息中的重要数据,比如话题信息和关键词等.

3.3.1 基于交互信息内容的度量

信息内容作为承载和传播影响力的载体,研究它们在社交网络中的传播方式和特性有助于深入理解社交影响力的作用模式.由于用户的社交影响力能够促进信息的传播,所以分析在线社交网络中信息内容的传播范围和时间,能够比较准确地反映用户的影响力.

在社交网络中,很多传播范围广泛的流行消息都是由影响力较大,拥有大量粉丝的用户发起的,因此流行消息的传播可以作为用户社交影响力的判断依据.Bakshy 等人[4]使用消息扩散产生的树结构计算和预测用户的影响力,发现 Twitter 上的很多流行话题都是由影响力较大的用户传播开的,但是要基于交互信息准确预测用户的社交影响力却非常困难.

除了信息的传播范围,用户发布的信息在社交网络中流传的时间长短可以反映用户影响力的深远程度,也是衡量社交影响力的重要指标.Romero 等人[62]同时分析了 Twitter 上的流行标签在传播范围和时间上的特点,发现不同标签的传播存在明显的区别.他们把标签曝光(即用户看到标签)次数与用户采用该标签的概率之间的关系称为标签的粘着性,把用户反复看到该标签后,持续对用户产生影响的程度称为标签的持久性,还定义了一种影响力曲线(也称"曝光曲线")描述标签的上述性质,如图 3 所示.给定某个标签,$P(k)$表示用户在经受 k 次曝光后才开始使用该标签的概率,图中曲线表示实验数据集中最流行的 500 个标签的 $P(k)$值的平均分布.明显可以看出大多数标签的 $P(k)$值在 k 为 2～4 时达到峰值,之后随着 k 值的增大呈下降趋势,该结果和 Kempe 等人观察到的影响力递减现象一致[98],也间接证明了时间因素在影响力度量中的重要作用.其他研究成果表明,用户自身的属性,比如活跃性和专注度[5],也能对信息传播过程和影响力的计算结果产生影响.

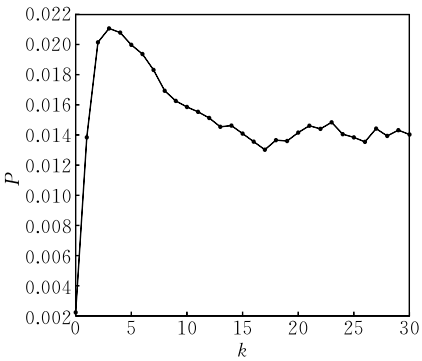

图 3 影响力曲线示意图[62]

上述工作主要从用户交互信息的传播范围和时间等因素出发,从广度和深度两方面对用户影响力进行了定性分析,但是基于交互信息定量计算社交影响力的工作则相对较少.同时,基于交互信息的度量方法还引发出另外一个重要问题:由于信息本身具有内在的传播属性,那么交互信息在社交网络中的传播,其自身所起作用和用户影响力的作用各占几何,如何进行区分和度量?研究和解决该问题有助于人们更深入地理解社交影响力的本质及其在社交活动中扮演的角色.

3.3.2 基于话题的度量

在人们的社交活动中,很多时候信息是以话题的形式产生和传播的.研究工作发现不同的话题具有不同的影响力,即便是相同的话题在不同人群中产生的影响力也不尽相同[87,94,99].所以,使用话题作为社交影响力度量的基本对象,能够从多个角度对用户的影响力进行细致刻画.在建立社交影响力模型时,可以直接从话题内容和用户对话题的参与度构建用户和话题之间的联系,无需用户之间通过好友申请或被关注等行为建立的社交网络拓扑结构作为模型输入,我们将前者分析出的影响力称为隐性

影响力，而将基于后者的计算结果称为显性影响力.同时，在线社交网络中的实体包括用户、文本、多媒体信息等，它们内部以及相互之间形成比同质网络更为复杂的异质网络结构.

Tang 等人[22]研究了用户间基于话题的影响力问题，定义了一种话题因子图 TFG(Topical Factor Graph)，用该模型统一包含与话题影响力分析相关的信息，包括社交网络结构、用户间的话题相似性和话题信息的分布情况等，同时支持同质/异质网络环境下的影响力分析.如图 4 所示，其中 $g(\cdot)$，$f(\cdot)$ 和 $h(\cdot)$是分别定义在用户、用户之间的连接以及全体用户之上的特征函数，$\boldsymbol{y}_i$ 是定义在用户 v_i 上的隐向量，表示其他用户和 v_i 在话题级别上的影响力.

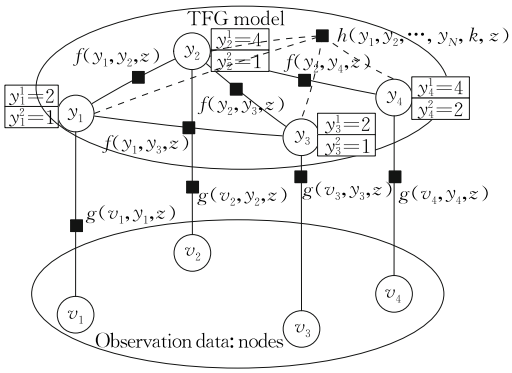

图 4　话题因子图[22]

基于上述工作，Liu 等人[31]将话题分布和影响力分析结合到一起考虑，在用户和各种文本信息构成的异质网络中，对基于话题的社交影响力进行了分析和建模，并利用文本内容的相似性挖掘用户之间的隐性影响和预测用户的行为.

进一步地，Cui 等人[100]研究了更小的信息粒度—信息条目(Item)—与社交影响力的关系，发现与话题信息的影响力因话题内容和因人而异相似，不同的信息条目也具有不同的影响力，即便是同一用户在不同信息条目上的影响力也在发生变化.他们根据社交网络中的历史交互信息分析信息条目、用户以及用户和信息之间的关系，度量和预测隐性社交影响力强度，并据此设计了预测影响力的方法 HF-NMF(Hybrid Factor Non-Negative Matrix Factorization)，最后用投影梯度矩阵因子分解方法予以求解.

在基于话题的影响力计算上，Weng 等人[7]使用了一种两阶段策略：首先用文本分析的方法提取用户感兴趣的话题，从而建立起话题之间的关系；接着使用 TwitterRank 算法分析了由话题相似度和网络结构两部分信息构成的用户影响力.该方法在基于网络结构的分析方法上引入话题信息，实验表明这种策略能够改善话题敏感类算法的功能和预测精度.

上述基于话题的影响力分析工作的特点如表 3 所示.这类方法在社交网络结构上融合了以话题为基本单位的用户交互信息，通过分析话题信息和用户之间的关系，能够更准确地度量用户影响力的产生和变化过程.尽管对该类方法的研究已取得不少成果，我们认为至少还有以下问题值得深入探索：首先，目前的在线社交网络除了文本数据，还包含大量的图像声音等多媒体信息，在这种异质网络中进行建模是基于话题的影响力分析需要应对的新挑战；其次，隐性影响力能够揭示更为隐秘的社会关系，那么它和来自同一社交网络的显性影响力之间是互补关系抑或替代关系，如何量化它们之间的联系；再次，话题的改变会导致用户影响力在计算时产生变化[5,93,101]，不同话题的传播过程也可能遵循不同的传播模型，但现实环境中用户的影响力具有相对稳定性，因此话题信息对在线社交网络中用户影响力的影响程度和方式，还有待更深入的研究和探索.

表 3　基于话题的影响力度量方法

模型	网络结构	显性/隐性	异质/同质	数据源
TFG	需要	显性	异质/同质	Wikipedia，ArnetMiner
Liu 等	需要	显性/隐性	异质	Twitter，Digg，Cora
HF-NMF	不需要	隐性	同质	人人网
TwitterRank	需要	显性	同质	Twitter

3.4　时间因素、转移熵等其他度量

在线社交网络是非常复杂的系统，有很多因素都会对网络用户影响力的度量产生约束，综合考虑这些因素既能提高模型的预测效果和精度，同时也为社交影响力的研究提供了新视角.

时间因素与社交网络中用户影响力的产生和传播紧密相关[25]，可以结合网络拓扑结构和信息传播的时间序列对用户影响力进行分析.文献[102]的工作发现信息传播大多数发生在时间序列的早期，所以将首次接触到信息的用户视为有效读者(Effective Reader)，而且利用上述特点提出的方法能够发现之前未被发掘出来的具有影响力的用户.但是也有相关研究观察到在信息传播早期就受到影响的用户，本身不一定具有较高影响力[103].文献[104]在交互信息缺失的情况下，利用个体的状态信息对影响力进行建模，并分析了影响力随时间变化的情况.

可以通过分析网络结构上不互联、在交互内容上却相互影响的社会关系获取用户之间的隐性影响力.转移熵可以表示这种随机过程中不确定性的变化情况,能利用社交网络中交互信息的演化过程计算用户间的隐性影响力[105-106],但是这类方法对实验数据量的要求较高.

社交影响力通过口口相传的方式在用户之间发挥作用,能够提高商品的营销效果,增加创新的采用范围.Huang 等人[107]通过统计分析发现,影响力不止在购买商品之前,在购买行为之后也能够显著提升用户对推荐商品的评价指标.

4 社交网络中影响力的传播

社交影响力本质上具有动态属性,从参与社交活动开始,每个人在社会群体中的影响力都在随着他的言行和社会属性发生变化,也随着社交活动在社交网络中进行传播.因此分析和研究社交影响力的动态传播过程,对认识影响力的本质特性,理解社交网络的形成和演化,以及发现社交网络中信息的传播规律和人们的行为模式等诸多问题具有重要意义.在 Katz 和 Lazarsfeld 提出的经典传播模型中[2],信息或者创新的传播首先从具有较强社会影响力的群体开始,再经由他们把信息和创新传播到更大范围的人群中去.

由于具有较强影响力的用户在传播过程中起着非常重要的作用,同时又能对数量众多的用户产生直接或者间接的影响,他们也被称为意见领袖.意见领袖发掘是社交影响力分析的热点问题之一.人们从商业营销中也发现,少数客户可以对大多数顾客的购买行为产生影响,如果能够找到这部分有影响力的客户,设法使他们采用最新的产品,就能通过口口相传的病毒式传播方式,最终使得整个社交网络上的大部分用户都采用该产品.这种影响力作用和传播过程被归结为影响力最大传播问题,以其重要的理论意义和应用价值吸引了大量研究人员投入到该问题的探索之中.

4.1 意见领袖问题

意见领袖就是能在相应环境下对其他人产生影响的个体[2],很多时候也被视为有影响力的人[108-109].根据分析过程中数据对象的不同,意见领袖发掘方法大致可以分为基于网络结构的方法和基于以交互信息为主的方法.

社交网络结构能在很大程度上反映某个节点在网络中所处位置的重要程度.如果用户在社交网络中的位置能够如实体现他在社交活动中的领袖地位,利用基于 PageRank 算法的排名方法就能对其进行度量[110-113].文献[113]把 PageRank 算法打分最高的1%的用户看作意见领袖,结果发现意见领袖比普通用户具有更高的社会地位.文献[110]分析和对比了多种衡量用户影响力的排名算法,强调了网络结构在算法运行和设计中的重要性,并提出一种类似 PageRank 的方法计算用户的专家排名.Song 等人[111]提出 InfluenceRank 算法,可以对用户的影响力和发布信息的新颖性进行度量,以此确定用户的排名.文献[114]则根据用户在社交网络上的中心度推断某人是否为意见领袖,并发现中心度高的用户可能是局部意见领袖,而紧密中心度和介数中心度可作为全局意见领袖的判断指标.

基于网络结构的方法过分强调了意见领袖的"领袖"特点,即在社交网络中地位显赫,而忽视了他发布的"意见"的重要性.虽然社交网络上的意见领袖都有较大的节点度,但是具有这种特点的用户不一定都是意见领袖.因此分析用户所发信息的影响力及其传播特性,能够更客观准确地发现网络中的意见领袖.博客中的意见领袖可以根据他发布的有影响力的帖子数量进行识别,Agarwal 等人[115]综合利用博文的引用数量、评论数量、新颖程度和内容长度评价帖子的影响力,从而发掘出意见领袖.通过观察和分析网络用户的行为特征及其传播规律,也能对用户的影响力大小和范围进行度量.文献[116]认为社团中的领袖发起的行为能在特定时间段内对一定数量的用户产生影响,并据此使用频繁模式挖掘方法识别社交网络中的领袖.樊兴华等人[117]分析了影响力扩散模型 IDM 的缺陷,对其进行改进后用于在线网络中的意见领袖发现.也有部分工作对用户在线问答关系中体现的权威性[118]和 BBS 中的意见领袖[119]进行了建模和分析.

上述两类方法各有优势和不足.基于网络结构的方法模型简单,计算效率较高,能够处理大规模的社交网络,但是准确率相对较低,在真实网络环境中存在误判的可能性;基于交互信息的分析方法得到的结果客观准确,但是由于涉及到大量信息的预处理和内容相关性的计算,难以适应规模较大的社交网络.因此可以考虑结合这两种方法的优点,采用两阶段的选择策略提高算法的效率和准确度:先利用基于网络结构的方法筛选意见领袖的备选集合,然后再用基于交互信息的方法选取真正的意见领袖.

4.2　影响力最大传播问题

影响力最大传播问题首先是由 Domingos 和 Richardson 提出的[15,17]，简言之就是在社交网路中寻找最有影响力的成员集合. Kempe 和 Kleinberg 等人[95]形式化表示了该问题，并总结出独立级联模型(Independent Cascade Model)、线性阈值模型(Linear Threshold Model)和这两种模型的泛化模型，随后采用离散优化方法对该问题进行了求解. 相对而言，独立级联模型在很多工作中都得到了应用，只有考虑到邻居的影响力对用户产生的累积效应时，人们才会优先使用线性阈值模型.

在上述模型中，个体之间的影响力权值默认是已知的常量或者提前赋予的数值，但是在现实环境中这种假设是不成立的. 因此有不少工作都对个体间的影响力因素及其相互关系进行了分析和量化[10,22,56]. 后来 Goyal 等人[48]利用用户的动作日志对影响力系数进行测算取得了较好的结果，也有工作尝试在日志数据缺失的情况下，利用网络拓扑和历史交互数据量化用户间的影响力[120]. 由于影响力最大传播问题是 NP 难的[95]，所以该问题的启发式求解方法也得到了深入研究.

4.2.1　影响力的传播模型

社交影响力是通过人们日常的交互活动发挥作用的. 在网络环境中，最主要的交互活动就是信息的发布、共享和扩散，所以影响力在社交网络中的作用过程和信息的扩散过程有内在紧密的联系和十分相似的机制，因此信息传播模型在影响力传播问题的研究过程中发挥着非常重要的作用，文献[121]有更为详尽的介绍.

(1) 独立级联模型[16,85,94-95](Independent Cascade Model)

IC 模型和描述传染病传播过程的 SIR(Susceptible-Infective-Recovered)模型原理相似[72]，可以描述为：在社交网络 $G=(V,E)$ 中，部分用户 $V_0\in V$ 在初始时刻就处于激活状态，用户 v_i 与其邻居 v_j 间的影响力用 $p_{i,j}$ 表示，$p_{i,j}$ 的取值是独立的，在传播过程中不受 v_i 和其他邻居节点间关系的影响. 如果用户 v_i 在某一时刻 t 由非激活状态变成激活状态，则 v_i 对处于非激活状态的每一个邻居节点，仅在 t 时刻有一次机会尝试激活他. 例如 v_i 的邻居 v_j 在 t 时刻还没有激活，则 v_i 以概率 $p_{i,j}$ 激活 v_j，如果成功，v_j 从 $t+1$ 时刻起处于激活状态，但是不管成功与否，v_i 再也不能试图去激活 v_j. 如果 v_j 在 t 时刻同时有多个邻居都变成了激活状态，则他们尝试激活 v_j 的顺序是任意的. 系统从初始态开始传播过程，直到没有新的用户可以被激活为止.

(2) 线性阈值模型[95,122-123](Linear Threshold Model)

线性阈值模型在新产品采用等问题中应用广泛，是诸多阈值模型的核心[95]. 该模型可以描述为：在社交网络 $G=(V,E)$ 中，用户 v_i 与其邻居节点 v_j 间的影响力权重为 $w_{i,j}$，且 v_i 所有邻居的影响力权重之和最大为 1，即

$$\sum_{v_j\in Ng_i} w_{i,j}\leq 1.$$

任意一个用户 v_i 都会随机选定属于自己的阈值 $\theta_i\in[0,1]$，表示只有当 v_i 的邻居节点对 v_i 的影响力超过该阈值，v_i 才会被激活. 与独立级联模型类似，初始时刻处于激活状态的用户集合为 V_0，如果截止 t 时刻用户 v_i 被激活的邻居集合为 Ng_i^{act}，而 v_i 尚未被激活，则在满足以下条件时：

$$\sum_{v_j\in Ng_i^{\text{act}}} w_{i,j}\geq\theta_i,$$

v_i 从 $t+1$ 时刻起变为激活状态，且保持该状态不变. 系统从初始状态开始演变，直到没有新用户能够被激活为止.

(3) 扩展及其它模型

Kempe 和 Kleinberg 等人[95]对 ICM 和 LTM 模型进行了扩展，舍弃了模型中的独立性条件，为级联模型定义了增量函数 $p_{v_j}(v_i,S_j)$ 表示用户 v_i 激活 v_j 的概率，S_j 是尝试激活 v_j 但最终失败的邻居节点集合，又为阈值模型定义了阈值函数 $f_{v_i}(Y_i)$，Y_i 是前一时刻 v_i 的邻居中被激活的节点集合，此时节点 v_i 被激活的条件变为

$$f_{v_i}(Y_i)\geq\theta_i.$$

上述两种扩展模型是相互等价的，可以很容易地实现相互转换，因此可以归结到同一个框架内.

Kempe 和 Kleinberg 等人[98]还发现，级联模型中用户 v_i 被邻居激活的概率随着时间的推移在递减，也就是说 v_i 被越多邻居尝试去激活而没有成功，则新激活的邻居对 v_i 的影响力就越弱，即

$$p_{v_i}(v_j,S_i)\geq p_{v_i}(v_j,S_i'),\ S_i\subseteq S_i',$$

并据此设计了递减级联模型(Decreasing Cascade Model)用于目标选择问题的建模.

Bass[12]提出的商品扩散模型也可以用来描述社交影响力的传播过程，在该模型中影响力的传播呈 S-形分布，即早期在用户之间传播较慢，后来突然成指数级增长，最后趋于平缓，但是该模型更多地

被用于定性分析而不是定量计算.

应用广泛的 ICM 可以很好地描述流行病的传播、信息在网络上的扩散等问题,但是如果从社交影响力角度出发对其进行解读,就会发现不少有悖事实的地方:ICM 中的用户只会被某个邻居激活,亦即社交网络中的用户最有可能和某个影响力强的邻居产生共鸣,其他邻居的影响力最终会被忽略;用户影响力在传播过程中只对其邻居产生一次作用,而且这种作用会同时发生.不论是 ICM 还是 LTM,用户受某些影响力的作用产生状态上的改变后,这种改变是不可逆的,也显然与现实情形不符.因此,设计更符合社交网络中影响力特性的传播模型仍将是一项挑战.

4.2.2 最大传播问题的定义和分析

该问题定义为,给定网络 $G=(V,E)$ 和常数 $k\leq|V|$,如果初始节点集合为 I,传播过程结束后预期激活节点集合为 $\sigma(I)$,找出节点集合 $S\subseteq V$ 且 $|S|=k$,使得 $\sigma(S)$ 最大.

函数 $\sigma(\cdot)$ 满足单调性,即对任意元素 v,给定集合 S,则有

$$\sigma(S\bigcup\{v\})\geq\sigma(S).$$

在实际应用中的含义就是初始用户集合越大,他们能影响到的用户数应该越大,至少不会减少.

$\sigma(\cdot)$ 还满足次模函数性质,给定集合 T,若集合 $S\subseteq T$,则

$$\sigma(S\bigcup\{v\})-\sigma(S)\geq\sigma(T\bigcup\{v\})-\sigma(T),$$

该性质的含义和上文所述的递减级联模型的含义类似.

可以证明影响力最大化问题在独立级联模型和线性阈值传播模型中都是 NP 问题[95,123-126],使用贪心算法求解后,算法的解以 $1-1/e$ 近似逼近最优解,即如果 S^* 是原问题的最优解,则有[95]

$$\sigma(S)\geq(1-1/e)\times\sigma(S^*).$$

由于影响力最大化问题的时间复杂度高,而且在线社交网络的规模日益庞大,所以设计启发式方法以期获得最优解和提高算法的执行效率一直是重要的研究方向.Kempe 等人[95]求解该问题的方法,首先需要多次调用蒙特卡洛算法以使模型获得足够的精度,而后贪心策略又需要调用上述过程 $O(nk)$ 次,其中 n 是网络中的用户数,k 为初始集合大小,因此时间耗费较高.在此基础之上,国内有工作利用线性阈值模型的累积效应提出了新的求解框架[127].Leskovec 等人[128]利用模型中次模函数的性质,在选择初始节点时提出一种"Lazy-Forward"优化机制生成初始用户集合,这种方法在取得近似最优解的同时效率比贪心算法提高了将近 700 倍.即便如此,该算法求解大规模问题仍有不足.Chen 等人[129]进一步优化了贪心算法的效率,在独立级联模型上提出一种度减小(Degree Discount)优化策略,使得实验结果与贪心算法相近,而运行效率有了很大提升.影响力最大传播在一般线性阈值模型上属于 NP 难问题,但在有向无环图上却能在线性时间予以求解,Chen 等人[123]为网络上的节点构造了局部有向无环图 LDAG(Local Directed Acyclic Graph),然后再用贪心策略求解原问题,实验表明该方法既具有较高的效率,而且能处理百万用户级别的社交网络.

除了在算法性能上的不断改进,随着影响力最大传播问题的深入研究和广泛应用,不断有新方法和新技术运用到该问题的建模和分析之中.投票模型可以模拟用户意见在社交网络上的传播情况[130],也可用于最优初始用户集合的选取[131].Kimura 和 Saito[132]认为影响力大都通过用户间的最短路径进行传播,由此入手对影响力最大传播问题进行了建模和求解,在此基础上又出现了基于用户间最大影响路径的方法[124].但是认为影响力通过最短路径传播的假设限制性太强,Wang 等人[133]发现影响力的传播大多发生在社团之间,由此提出一种贪心策略结合动态规划的算法用于初始用户的选取,较大提升了算法的执行效率.

目前的影响力最大传播模型只考虑到最小初始用户集合的选取,并没有将激活用户的代价和时间计算在内.在实际应用中,使用最小的费用或者耗费最短的时间实现影响力的最大传播也是常见情形,而此类问题还需进一步研究.现有模型认为用户之间的影响力只会对其传播有促进作用,但是现实营销环境总是面临各种竞争因素,因此竞争环境下的影响力最大传播问题也逐渐受到研究人员的关注[134-137].

5 总结与展望

随着在线社交网络的蓬勃发展和线上用户的急剧增长,以交友、信息共享等为目的的社交网络迅速成长为人们传播信息、推销商品、表述观点、产生影响力的理想平台.在线社交网络中的影响力分析和建模是社交网络分析的重要内容,通过分析人们相互之间的影响模式和影响力传播方式,既能够从社

会学角度加深理解人们的社会行为，为公共决策和舆情导向等提供理论依据，同时还能促进政治、经济和文化活动等多个领域的交流和传播，具有重要的社会意义和应用价值.

本文主要介绍了在线社交网络兴起以来社交影响力分析的主要成果，首先阐述了社交影响力的基本概念和与之相关的其他因素，介绍了区分影响力和这些因素的研究工作，接着重点总结了影响力分析的建模和度量方法，以及与影响力传播相关的意见领袖发现和影响力最大传播问题的研究现状. 虽然社交影响力分析已经取得丰硕的理论和应用成果，我们认为至少还有以下问题有待深入研究和探索：

(1) 社交网络用户数量众多，用户之间形成的关系也非常复杂，在这样的环境下对社交影响力的定性分析也受到很多因素的影响和干扰. 尽管有不少工作试图客观准确地厘清影响力和其他因素之间的关系，使用了包括随机化方法[51-52,138]在内的很多技术手段，但是最终都无法很好地解决该问题. 这种局面既与社交影响力复杂的产生和传播机制相关，也与影响力自身的定义有关系. 现阶段的影响力概念只是影响力效果的一种描述，实质上并未解释“影响力是什么”这一问题，造成在研究时社交影响力模型众多，缺少基准对比模型和方法. 或许我们无法精确定义社交影响力这一概念，但是就在线社交网络这一具体环境而言，有必要深入研究社交影响力的评判指标，为新模型的设计提供方向性指导，使其能够更准确地描述在线社交网络中的复杂现象.

(2) 目前，社交影响力的建模方法大致可分为两类，即经验方法和推断方法. 经验方法通过实验数据的观察和分析，总结出符合样本的数学模型，再通过和实际数据的拟合获取模型中的参数值. 推断方法则根据相关理论直接推导出影响力模型，同样再采用学习和拟合等方法确定模型中的参数值. 这两种方法都有各自的优势和成功的应用，但是还没有什么方法能够普遍准确地刻画社交网络中的影响力. 这种局面的改变，需要在影响力的定义、社交信息和影响力的关系等理论工作上取得突破，同时也需要建模方法的改进. 我们可以综合利用社交网络上的拓扑结构、用户交互数据和动作记录等信息，从多角度全面分析社交网络中用户影响力的产生及传播过程，而在实时性要求较高的场合，还可以考虑使用增量模型以减少计算量.

(3) 大量关于社交影响力分析的工作都聚焦在用户自身的影响力度量和演化，以及用户及其邻居和所在社团之间的相互影响上，但是以用户群体为基本目标进行影响力分析的研究还不是很多[33]，而且也有工作表明，当前的传播模型更适合于对大量用户的平均影响力进行评估[4]. 所以对具有特定属性的用户形成的群体以及群体之间的相互影响进行分析和度量，可以从更为宏观的层面上理解社交影响力的作用原理和传播机制，促进该领域知识的进一步发展.

(4) 在线社交网络的数据采集和共享也是影响力分析亟待解决的问题. 海量交互数据是影响力分析不可或缺的宝贵资源，也是当前影响力分析有别于早期研究的主要特点之一. 但是从网络上在线获取这类信息需要花费大量资源，而且由于涉及到商业利益和用户隐私保护等问题，许多著名社交网站都对数据的合法下载和使用设置了严苛的条件. 因此该问题的解决需要学术界和相关企业通力合作，以期在不损害企业利益和用户隐私的前提下，获取足够的分析数据，促进学术和商业的共同发展.

致　谢　感谢 Haewoon Kwak 博士为本文提供文献中的原始图表，同时感谢评审专家提供的宝贵意见和建议！

参　考　文　献

[1] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge, U. K.: Cambridge University Press, 1994

[2] Katz E, Lazarsfeld P F. Personal Influence: The Part Played by People in the Flow of Mass Communications. Piscataway: Transaction Publishers, 2006

[3] Rashid A M, Karypis G, Riedl J. Influence in ratings-based recommender systems: An algorithm-independent approach// Proceedings of the SIAM International Conference on Data Mining. Newport Beach, California, USA, 2005: 556-560

[4] Bakshy E, Hofman J M, Mason W A, Watts D J. Everyone's an influencer: Quantifying influence on twitter//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 65-74

[5] Cha M, Haddadi H, Benevenuto F, Gummadi K P. Measuring user influence in twitter: The million follower fallacy// Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, DC, USA, 2010: 10-17

[6] Romero D M, Galuba W, Asur S, Huberman B A. Influence and passivity in social media//Proceedings of the European Conference on Machine Learning and Principles and Practices

of Knowledge Discovery in Databases. Athens, Greece, 2011: 18-33

[7] Weng J, Lim E-P, Jiang J, He Q. Twitterrank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York City, USA, 2010: 261-270

[8] Yang J, Leskovec J. Modeling information diffusion in implicit networks//Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia, 2010: 599-608

[9] Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group formation in large social networks: Membership, growth, and evolution//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 44-54

[10] Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 160-168

[11] Aiello L M, Barrat A, Schifanella R, et al. Friendship prediction and homophily in social media. ACM Transactions on the Web, 2012, 6(2): 1-33

[12] Bass F M. A new product growth model for consumer durables. Management Science, 1969, 15(5): 215-227

[13] Brown J J, Reingen P H. Social ties and word-of-mouth referral behavior. Journal of Consumer Research, 1987, 14(3): 350-362

[14] Mahajan V, Muller E, Bass F M. New product diffusion models in marketing: A review and directions for research. The Journal of Marketing, 1990, 54(1): 1-26

[15] Domingos P, Richardson M. Mining the network value of customers//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 57-66

[16] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001, 12(3): 211-223

[17] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 61-70

[18] Leskovec J, Adamic L A, Huberman B A. The dynamics of viral marketing//Proceedings of the 7th ACM Conference on Electronic Commerce. Ann Arbor, USA, 2006: 228-237

[19] Christakis N A, Fowler J H. The spread of obesity in a large social network over 32 years. New England Journal of Medicine, 2007, 357(4): 370-379

[20] Fowler J H, Christakis N A. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. British Medical Journal, 2008, 337(a2338): 1-9

[21] Dong W, Pentland A. Modeling influence between experts//Proceedings of the ICMI 2006 and IJCAI 2007 International Conference on Artificial Intelligence for Human Computing. Banff, Canada, 2007: 170-189

[22] Tang J, Sun J, Wang C, Yang Z. Social influence analysis in large-scale networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 807-816

[23] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes twitter users: Real-time event detection by social sensors//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 851-860

[24] Bakshy E, Eckles D, Yan R, Rosenn I. Social influence in social advertising: Evidence from field experiments//Proceedings of the 13th ACM Conference on Electronic Commerce. Valencia, Spain, 2012: 146-161

[25] Rogers E M. Diffusion of Innovations. 5th Edition. New York: Free Press, 2003

[26] Keller E, Berry J. The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy. New York: Free Press, 2003

[27] Manski C F. Identification of endogenous social effects: The reflection problem. The Review of Economic Studies, 1993, 60(3): 531-542

[28] Ellison N B. Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 2007, 13(1): 210-230

[29] Sun J, Tang J. A survey of models and algorithms for social influence analysis//Aggarwal C C ed. Social Network Data Analytics. New York: Springer, 2011: 177-214

[30] Goyal A. Social influence and its applications: An algorithmic and data mining study[Ph. D. dissertation]. The University of British Columbia, Vancouver, USA, 2013

[31] Liu L, Tang J, Han J, et al. Mining topic-level influence in heterogeneous networks//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 199-208

[32] Liu L, Tang J, Han J, Yang S. Learning influence from heterogeneous social networks. Data Mining and Knowledge Discovery, 2012, 25(3): 511-544

[33] Goyal A, Bonchi F, Lakshmanan L V S, Venkatasubramanian S. On minimizing budget and time in influence propagation over social networks. Social Network Analysis and Mining, 2013, 3(2): 179-192

[34] Wang Y, Huang W, Zong L, et al. Influence maximization with limit cost in social network. Science China Information Sciences, 2013, 56(7): 1-14

[35] Iwata T, Shah A, Ghahramani Z. Discovering latent influence in online social activities via shared cascade Poisson processes//Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 266-274

[36] Shuai X, Ding Y, Busemeyer J, et al. Modeling indirect influence on twitter. International Journal on Semantic Web and Information Systems (IJSWIS), 2012, 8(4): 20-36

[37] Mehmood Y, Barbieri N, Bonchi F, Ukkonen A. CSI: Community-level social influence analysis//Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases. Prague, Czech Republic, 2013: 48-63

[38] Rashotte L. Social Influence: The Blackwell Encyclopedia of Social Psychology. Vol IX. Malden: Blackwell Publishing, 2007

[39] Gladwell M. The Tipping Point: How Little Things Can Make a Big Difference. Boston: Back Bay Books, 2000

[40] Goldenberg J, Han S, Lehmann D, Hong J. The role of hubs in the adoption processes. Journal of Marketing, 2009, 73(2): 1-13

[41] Weimann G. The Influentials: People Who Influence People. Albany, New York: State University of New York Press, 1994

[42] Watts D J, Dodds P S. Influentials, networks, and public opinion formation. Journal of Consumer Research, 2007, 34(4): 441-458

[43] Lazarsfeld P F, Merton R K. Friendship as a social process: A substantive and methodological analysis. Freedom and Control in Modern Society, 1954, 18(1): 18-66

[44] McPherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in social networks. Annual Review of Sociology, 2001, 27: 415-444

[45] Lewis K, Gonzalez M, Kaufman J. Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences, 2012, 109(1): 68-72

[46] Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences, 2009, 106(51): 21544-21549

[47] Shalizi C R, Thomas A C. Homophily and contagion are generically confounded in observational social network studies. Sociological Methods & Research, 2011, 40(2): 211-239

[48] Goyal A, Bonchi F, Lakshmanan L V. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 241-250

[49] Wang G, Hu Q, Yu P S. Influence and similarity on heterogeneous networks//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, Hawaii, USA, 2012: 1462-1466

[50] Singla P, Richardson M. Yes, there is a correlation: -from social networks to personal behavior on the web//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 655-664

[51] Fond T L, Neville J. Randomization tests for distinguishing social influence and homophily effects//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 601-610

[52] Aral S, Walker D. Identifying influential and susceptible members of social networks. Science, 2012, 337(6092): 337-341

[53] Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media?//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 591-600

[54] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 611-617

[55] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the flicker social network//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, 2009: 721-730

[56] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 7-15

[57] Godes D, Mayzlin D. Using online conversations to study word-of-mouth communication. Marketing Science, 2004, 23(4): 545-560

[58] Van den Bulte C, Lilien G L. Medical innovation revisited: Social contagion versus marketing effort. American Journal of Sociology, 2001, 106(5): 1409-1435

[59] Manski C F. Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press, 1995

[60] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653

[61] Myers S A, Zhu C, Leskovec J. Information diffusion and external influence in networks//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 33-41

[62] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter//Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, 2011: 695-704

[63] Rodriguez M G, Leskovec J, Krause A. Inferring networks of diffusion and influence//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 1019-1028

[64] Watts D. Challenging the influentials hypothesis. Word of Mouth Marketing Association Measuring Word of Mouth, 2007, 3(4): 201-211

[65] Watts D J, Peretti J, Frumin M. Viral Marketing for the Real World. Boston: Harvard Business School Pub, 2007

[66] Borgatti S P, Everett M G. A graph-theoretic perspective on centrality. Social Networks, 2006, 28(4): 466-484

[67] Ghosh R, Lerman K. Predicting influential users in online social networks//Proceedings of the 4th KDD Workshop on Social Network Analysis. Washington, DC, USA, 2010

[68] Freeman L C. Centrality in social networks conceptual clarification. Social Networks, 1979, 1(3): 215-239

[69] Sabidussi G. The centrality index of a graph. Psychometrika, 1966, 31(4): 581-603

[70] Newman M E. A measure of betweenness centrality based on random walks. Social Networks, 2005, 27(1): 39-54

[71] Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977, 40(1): 35-41

[72] Newman M E. The structure and function of complex networks. Society for Industrial and Applied Mathematics Review, 2003, 45(2): 167-256

[73] Bonacich P. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology, 1972, 2(1): 113-120

[74] Bonacich P. Some unique properties of eigenvector centrality. Social Networks, 2007, 29(4): 555-564

[75] Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18(1): 39-43

[76] Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998

[77] Haveliwala T H. Topic-sensitive pagerank//Proceedings of the 11th International Conference on World Wide Web. Honolulu, USA, 2002: 517-526

[78] Java A, Kolari P, Finin T, Oates T. Modeling the spread of influence on the blogosphere//Proceedings of the 15th International Conference on World Wide Web. Edinburgh, UK, 2006: 22-26

[79] Bonacich P. Power and centrality: A family of measures. American Journal of Sociology, 1987, 92(5): 1170-1182

[80] Hubbell C H. An input-output approach to clique identification. Sociometry, 1965, 28(4): 377-399

[81] Bonacich P. Simultaneous group and individual centralities. Social Networks, 1991, 13(2): 155-168

[82] Holland P W, Leinhardt S. Transitivity in structural models of small groups. Comparative Group Studies, 1971, 2(2): 107-124

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

[84] Barabasi A-L. The origin of bursts and heavy tails in human dynamics. Nature, 2005, 435(7039): 207-211

[85] Leskovec J, McGlohon M, Faloutsos C, et al. Cascading behavior in large blog graphs//Proceedings of the Society for Industrial and Applied Mathematics International Conference on Data Mining. Minneapolis, USA, 2007: 551-556

[86] Malmgren R D, Stouffer D B, Motter A E, Amaral L A N. A Poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences, 2008, 105(47): 18153-18158

[87] Granovetter M S. The strength of weak ties. American Journal of Sociology, 1973, 78(6): 1360-1380

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

[89] Xiang R, Neville J, Rogati M. Modeling relationship strength in online social networks//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 981-990

[90] Trusov M, Bodapati A V, Bucklin R E. Determining influential users in internet social networks. Journal of Marketing Research, 2010, 47(4): 643-658

[91] Leenders R T A. Modeling social influence through network autocorrelation: Constructing the weight matrix. Social Networks, 2002, 24(1): 21-47

[92] Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model//Proceedings of the 12th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. Zagreb, Croatia, 2008: 67-75

[93] Saito K, Kimura M, Ohara K, Motoda H. Selecting information diffusion models over social networks for behavioral analysis//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Barcelona, Spain, 2010: 180-195

[94] Gruhl D, Guha R, Liben-Nowell D, Tomkins A. Information diffusion through blogspace//Proceedings of the 13th International Conference on World Wide Web. New York, USA, 2004: 491-501

[95] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2003: 137-146

[96] McLachlan G, Krishnan T. The EM Algorithm and Extensions: Whiley Series in Probability and Statistics. New York: John Wiley & Sons, 1997

[97] Tan C, Tang J, Sun J, et al. Social action tracking via noise tolerant time-varying factor graphs//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 1049-1058

[98] Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks//Proceedings of the 32nd International Colloquium on Automata, Languages and Programming. Lisbon, Portugal, 2005: 1127-1138

[99] Krackhardt D. The strength of strong ties: The importance of philos in organizations//Nohria N, Eccles R G eds. Networks and Organizations. Cambridge, USA: Harvard Business School Press, 1992: 216-239

[100] Cui P, Wang F, Liu S, et al. Who should share what?: Item-level social influence prediction for users and posts

ranking//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 185-194

[101] Macskassy S A, Michelson M. Why do people retweet? Anti-homophily wins the day//Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. Barcelona, Spain, 2011: 209-216

[102] Lee C, Kwak H, Park H, Moon S. Finding influentials based on the temporal order of information adoption in twitter//Proceedings of the 19th International Conference on World Wide Web. Raleigh, USA, 2010: 1137-1138

[103] Bakshy E, Karrer B, Adamic L A. Social influence and the diffusion of user-created content//Proceedings of the 10th ACM Conference on Electronic Commerce. Stanford, California, USA, 2009: 325-334

[104] Pan W, Dong W, Cebrian M, et al. Modeling dynamical influence in human interaction: Using data to make better inferences about influence within social systems. IEEE Signal Processing Magazine, 2012, 29(2): 77-86

[105] Steeg G V, Galstyan A. Information transfer in social media//Proceedings of the 21st International Conference on World Wide Web. Lyon, France, 2012: 509-518

[106] Steeg G V, Galstyan A. Information-theoretic measures of influence based on content dynamics//Proceedings of the 6th ACM International Conference on Web Search and Data Mining. Rome, Italy, 2013: 3-12

[107] Huang J, Cheng X-Q, Shen H-W, et al. Exploring social influence via posterior effect of word-of-mouth recommendations//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, Washington, USA, 2012: 573-582

[108] Merton R K. Social Theory and Social Structure. New York: Free Press, 1968: 441-474

[109] Grewal R, Mehta R, Kardes F R. The role of the social-identity function of attitudes in consumer innovativeness and opinion leadership. Journal of Economic Psychology, 2000, 21(3): 233-252

[110] Zhang J, Ackerman M S, Adamic L. Expertise networks in online communities: Structure and algorithms//Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada, 2007: 221-230

[111] Song X, Chi Y, Hino K, Tseng B. Identifying opinion leaders in the blogosphere//Proceedings of the 16th ACM International Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 971-974

[112] Hajian B, White T. Modelling influence in a social network: Metrics and evaluation//Proceedings of the 3rd IEEE International Conference on Social Computing. Boston, USA, 2011: 497-500

[113] Tang J, Lou T, Kleinberg J. Inferring social ties across heterogenous networks//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 743-752

[114] Bodendorf F, Kaiser C. Detecting opinion leaders and trends in online social networks//Proceedings of the 2nd ACM Workshop on Social Web Search and Mining. Hong Kong, China, 2009: 65-68

[115] Agarwal N, Liu H, Tang L, Yu P S. Identifying the influential bloggers in a community//Proceedings of the 1st International Conference on Web Search and Data Mining. Palo Alto, USA, 2008: 207-218

[116] Goyal A, Bonchi F, Lakshmanan L V S. Discovering leaders from community actions//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, USA, 2008: 499-508

[117] Fan Xing-Hua, Zhao Jing, Fang Bin-Xing, Li Yu-Xiao. Influence diffusion probability model and utilizing it to identify network opinion leader. Chinese Journal of Computers, 2013, 36(2): 360-367(in Chinese)
(樊兴华，赵静，方滨兴，李欲晓. 影响力扩散概率模型及其用于意见领袖发现研究. 计算机学报，2013，36(2)：360-367)

[118] Jurczyk P, Agichtein E. Discovering authorities in question answer communities by using link analysis//Proceedings of the 16th ACM International Conference on Information and Knowledge Management. Lisbon, Portugal, 2007: 919-922

[119] Zhai Z, Xu H, Jia P. Identifying opinion leaders in BBS//Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and International Conference on Intelligent Agent Technology. Sydney, Australia, 2008: 398-401

[120] Wang C, Tang J, Sun J, Han J. Dynamic social influence analysis through time-dependent factor graphs//Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining. Kaohsiung City, Taiwan, China, 2011: 239-246

[121] Guille A, Hacid H, Favre C, Zighed D A. Information diffusion in online social networks: A survey. SIGMOD Record, 2013, 42(2): 17

[122] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978, 83(6): 1420-1443

[123] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model//Proceedings of the 2010 IEEE International Conference on Data Mining. Sydney, Australia, 2010: 88-97

[124] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 1029-1038

[125] Chen N. On the approximability of influence in social networks. SIAM Journal on Discrete Mathematics, 2009, 23(3): 1400-1415

[126] Ben-Zwi O, Hermelin D, Lokshtanov D, Newman I. An exact almost optimal algorithm for target set selection in social networks//Proceedings of the 10th ACM Conference

on Electronic Commerce. Stanford, California, USA, 2009: 355-362

[127] Tian Jia-Tang, Wang Yi-Tong, Feng Xiao-Jun. A new hybrid algorithm for influence maximization in social networks. Chinese Journal of Computers, 2011, 34(10): 1956-1965(in Chinese)
(田家堂, 王轶彤, 冯小军. 一种新型的社会网络影响最大化算法. 计算机学报, 2011, 34(10): 1956-1965)

[128] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 420-429

[129] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-208

[130] Clifford P, Sudbury A. A model for spatial conflict. Biometrika, 1973, 60(3): 581-588

[131] Even-Dar E, Shapira A. A note on maximizing the spread of influence in social networks//Proceedings of the 3rd International Workshop on Internet and Network Economics. San Diego, USA, 2007: 281-286

[132] Kimura M, Saito K. Approximate solutions for the influence maximization problem in a social network//Proceedings of the 10th International Conference in Knowledge Based and Intelligent Information and Engineering Systems. Bournemouth, UK, 2006: 937-944

[133] Wang Y, Cong G, Song G, Xie K. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA, 2010: 1039-1048

[134] Bharathi S, Kempe D, Salek M. Competitive influence maximization in social networks//Proceedings of the 3rd International Workshop on Internet and Network Economics. San Diego, USA, 2007: 306-311

[135] Carnes T, Nagarajan C, Wild S M, van Zuylen A. Maximizing influence in a competitive social network: A follower's perspective//Proceedings of the 9th International Conference on Electronic Commerce. Minneapolis, USA, 2007: 351-360

[136] Goyal S, Kearns M. Competitive contagion in networks//Proceedings of the 44th Symposium on Theory of Computing. New York, USA, 2012: 759-774

[137] Borodin A, Filmus Y, Oren J. Threshold models for competitive influence in social networks//Proceedings of the 6th International Workshop on Internet and Network Economics. Stanford, California, USA, 2010: 539-550

[138] Aral S, Walker D. Identifying social influence in networks using randomized experiments. IEEE Intelligent Systems, 2011, 26(5): 91-96

WU Xin-Dong, born in 1963, Ph.D., professor, Ph.D. supervisor, IEEE Fellow, AAAS Fellow. His research interests include data mining, big data analytics, knowledge-based systems and web information exploration.

LI Yi, born in 1981, Ph. D. candidate, lecturer. His research interests include data mining, social network analysis and intelligent information processing.

LI Lei, born in 1981, Ph. D., associate professor. His research interests include data mining, social computing and trust computing.

Background

Social influence can be observed when people exchange their opinions, emotions, attitudes or behaviors with others. It plays an important role in the domains of economics, politics, culture, etc. Researchers have obtained lots of valuable findings through persistent explorations in more than half a century, including its generation, evolution, spreading and modeling.

However, online social networks rose abruptly in the past decade and have provided massive data produced by social behaviors when people interact with each other. This has been the first time that researchers could validate past findings on social influence with massive realistic data. Unfortunately, many of the findings need to be improved in new application environments. Furthermore, new problems are continually proposed, and new interesting findings emerge simultaneously.

This paper reviewed important achievements made by computer scientists on social influence analysis in recent years. The authors elaborated the concepts concerning social influence such as homophily and reciprocity, etc. Then important models and measuring methods were introduced along with two well-known propagating problems, discovering opinion leaders and maximizing the spread of social influence. Finally, we discussed future research directions of online social network analysis.

This work is supported by the National Basic Research Program (973 Program) of China under Grant No. 2013CB329604 and the Program for Changjiang Scholars and Innovative Research Team in University of the Ministry of Education, China, under Grant No. IRT13059. Our study focuses on group influence and its interactive effects in social networks.