# 社交网络影响力传播研究

**陈 卫**
微软亚洲研究院 北京 100080

**摘要**
随着互联网和大数据的研究应用日益广泛，对社交网络影响力传播的研究成为数据挖掘和社交网络分析中的热点。从影响力传播模型、影响力传播学习和影响力传播优化3个方面总结了近些年计算机科学领域对影响力传播研究的主要成果，展示了影响力传播研究中对随机模型、数据挖掘、算法优化和博弈论等技术的综合运用。最后，简要讨论了影响力传播研究和应用中存在的问题、挑战及今后的研究方向。

**关键词**
社交网络；社会影响力；影响力传播模型；影响力最大化；社会影响力学习；病毒营销

doi: 10.11959/j.issn.2096-0271.2015031

# *Research on Influence Diffusion in Social Network*

**Chen Wei**
Microsoft Research Asia, Beijing 100080, China

***Abstract***
With the wide spread of internet and big data research and applications, influence diffusion research in social network becomes one of the hot topics in data mining and social network analysis in recent years. The main results on social influence diffusion research from the field of computer science in the last decade, which covers the three main areas -- influence diffusion modeling, influence diffusion learning, and influence diffusion optimization, were summarized. Different techniques, such as stochastic modeling, data mining, algorithmic optimization, and game theory, were demonstrated in their application to influence diffusion research. Finally, some discussions on the current issues, challenges and future directions in influence diffusion research and applications were provided.

***Key words***
social network, social influence, influence diffusion model, influence maximization, social influence learning, viral marketing

## 1 引言

任何社会性动物在个体与个体、群体与个体之间都存在着相互影响的关系，例如个体依从群体的行为会有利于猎食或减少被猎食的可能。而人类作为具有复杂交流手段的高级社会性动物，社会影响力在社会生活中更是无处不在。小到听一首歌曲、选一个餐馆，大到确定政治观点或买一处房产等，人们的各种选择和决定常常受家人、同事、朋友以及更广泛的大众倾向的影响。深入认识影响力的产生和传播模式有助于理解人类群体和个体的行为，从而能够预期人们的行为，为政府、机构、企业等各部门的决策提供可靠的依据和建议。比如企业在进行新产品推广时，可以利用对用户影响力及其传播的了解，选择有影响力的用户和传播渠道帮助产品推广，而政府可以选择合适的影响力群体和渠道来扩大其政策的影响或抵御谣言的传播。

社会影响力的研究在社会科学和市场学领域已有较长的历史[1,2]，为影响力传播的途径和范围带来了新的认识。比如Christakis和Fowler利用美国一个城市上万人跨32年的医疗记录数据验证了肥胖症和吸烟行为会在社交网络中相互影响和传播[3,4]。而伴随着互联网、在线社交网络和大数据的兴起以及日益广泛的应用，在更大规模下更深入地研究影响力的传播也成为可能。比如近期基于著名的社交网站脸谱（Facebook）平台的两项研究，都通过在线随机试验方式分别验证了影响力在选举意愿和应用选择中的存在性及其决定性因素[5,6]。

在计算机科学领域，基于互联网和大数据的影响力传播研究也从21世纪开始兴起。本文集中介绍这十几年来计算机科学领域在社交网络影响力传播方面的研究成果，并对面临的挑战和今后的方向加以简要讨论。概括来讲，影响力传播研究有三大支柱（如图1所示）。第一是影响力传播模型，主要描述影响力在社交网络中如何传播、有何特点和性质。第二是影响力传播的学习，即如何利用网络大数据挖掘学习影响力传播模式和具体传播模型的参数。第三是影响力传播优化，着重考虑在不同的传播模型下，如何通过施加外部作用（比如选取有影响力的初始传播用户和改变传播途径等）来扩大希望传播的影响力或者控制和减弱不希望传播的影响力，也包括有效地监控影响力的传播等。下文就分别对影响力传播的这三大支柱进行一一讲述。

影响力的研究和应用也是一个涵盖很广的课题，也有其他的综述型文章对其加以介绍[7,8]。与其他综述不同的是本文重点介绍对影响力动态传播特性的研究，而其他方面（如从静态图特性估计节点影响力等）请参见其他综述文章的相关介绍[7,8]。Chen、Lakshmanan和Castillo在近期发表了信息和影响力传播方面的专著[9]，对影响力传播的研究有较详尽的介绍。本文对该专著覆盖的内容进行了提炼、概括，并包含了对该专著出版后的最新研究成果的介绍和一些新观点的讨论。

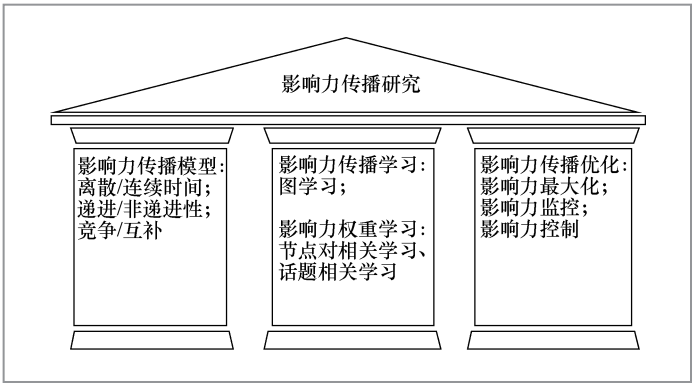



图1 社会影响力传播研究的三大支柱

## 2 影响力传播模型

信息和影响力在社交网络中的传播复杂多样，但排除一些干扰因素后仍然有章可循。在下文中统一用影响力传播来概括在社交网络中信息、概念、想法、创新、产品、文化基因（meme）等的传播。

首先把一个社交网络描述成一个有向图$G=(V, E)$，其中$V$是节点的集合，$E\subseteq V\times V$是有向边的集合。每一个节点$v\in V$代表一个社交网络中的人，每一条边$(u,v)\in E$代表节点$u$到节点$v$的影响力关系。边是有向的，表明影响力是有方向的，节点$u$对节点$v$有影响力，但节点$v$对节点$u$可能没有影响力。在后面具体建模中还通常会对边加上权重以表示影响力的强度。对于一条有向边$(u,v)\in E$，它叫做节点$u$的出边，节点$v$的入边，节点$v$是节点$u$的一个出邻居，而节点$u$是节点$v$的一个入邻居。一个节点$v$的所有出邻居的集合用$N^+(v)$表示，所有入邻居的集合用$N^-(v)$表示。

通常情况下，针对某一具体传播的实体（信息、想法、产品等），将图中的每个点描述为两种可能状态：不活跃（inactive）和活跃（active）。不活跃状态表示该个体还没有接受对应实体（信息、想法或产品），而活跃状态表示该个体已经接受对应的实体。节点从不活跃状态变为活跃状态表示该节点接受了对应实体，也称之为被激活。

影响力传播模型用来刻画影响力在社交网络中的传播模式，也即社交网络中节点的状态如何影响其相邻节点的状态，并造成某一状态（通常指活跃状态）在网络中扩散传播。传播模型分很多种类，其中大多数以随机模型（stochastic models）来描述，也有用博弈论模型（game-theoretic models）来描述的。本文着重描述随机模型，因为它更直接地反映了影响力传播中的不确定性，也是当前研究的主流。

随机模型又可分为离散时间和连续时间模型、递进性（progressive）和非递进性（non-progressive）模型等。离散时间模型将影响力传播和节点的状态转换规定在离散的时间点发生，以便于计算和分析，而连续时间模型允许传播和节点状态转换在连续时间轴上发生。递进性模型假设任意节点一旦从不活跃变为活跃就会一直保持在活跃状态，不会再回到不活跃状态。这类模型多用于信息、产品等的传播，因为它们一旦拥有，就一般不会再失去，或者只关注传播过程中所有曾经接受该信息或产品的人群。非递进模型则允许节点在两个（或多个）不同状态之间来回切换。这类模型多用于描述观点、看法的传播，因为人的观点和看法经常会随着时间和周围人群的观点而改变。在这种情况下也许对状态的描述用支持、反对等词语比用不活跃和活跃更合适。在众多模型中，离散时间递进性模型是研究最多的。本文以介绍经典的离散时间递进性模型和其上的应用问题为主线，附带简略介绍其他模型。

### 2.1 经典离散时间递进性传播模型

影响力传播模型的研究在社会和管理科学中由来已久[1,2]，但在计算机科学中基于计算和大数据的社交网络影响力传播模型的研究还是21世纪之后的事情。首先是Domingos和 Richardson 于2001年提出了基于马尔科夫随机场（Markov random field）的社交网络影响力模型[10]。严格地说，这个模型是关于图中节点被激活的相关性模型，而不直接表达影响力传播的因果关系。2003年，Kempe、Kleinberg

和 Tardos提出了独立级联(independent cascade)和线性阈值(linear threshold)等离散时间递进性传播模型和它们的若干拓展模型[11]。这些模型总结了前人在社会心理学、市场学及统计物理方面的模型，简单直观，基本符合人们对影响力传播的直觉理解，同时模型具有较好的性质，便于进一步分析和计算。这些模型如今已成为研究影响力传播的经典模型，被广泛应用到影响力最大化、影响力学习和影响力传播模型拓展等各个研究方面。下面对独立级联和线性阈值模型加以介绍，并在以后各部分中以独立级联模型为主要实例，介绍模型在各方面研究的应用。

### 2.1.1 独立级联模型

如图2所示，在独立级联模型中，每一条图中的有向边$(u,v)\in E$都有一个对应的概率值$p(u,v)\in[0,1]$。直观上说，$p(u,v)$表示当节点$u$被激活后，节点$u$通过边$(u,v)$独立激活节点$v$的概率。独立级联模型下的动态传播过程在离散时间点以如下形式完成：在$t$=0时刻，一个预先选好的初始集合$S_0$首先被激活，而其他节点都处于不活跃状态。这个初始节点集合被称作种子节点集合(seed set)。对任何时刻$t\geq1$，用$S_t$表示到这个时刻为止所有活跃点的集合。在任何时刻点$t\geq1$，对任何一个在上一时刻刚被激活的节点$u\in S_{t-1}\backslash S_{t-2}$(设$S_{-1}=\phi$)，节点$u$会对它的每个尚未被激活的出邻居节点$v\in N^+(v)\backslash S_{t-1}$尝试激活一次，而这次尝试成功的概率为$p(u,v)$，且这次激活尝试与所有其他的激活尝试事件相互独立。如果尝试成功，则节点$v$在时刻$t$被激活，即$v\in S_t\backslash S_{t-1}$；如果尝试不成功，且节点$v$的其他入邻居也未在时刻$t$成功激活节点$v$，则节点$v$在时刻$t$仍为不活跃状态，即$v\in V\backslash S_t$。当在某一时刻不再有新的节点被激活时，传播过程结束。

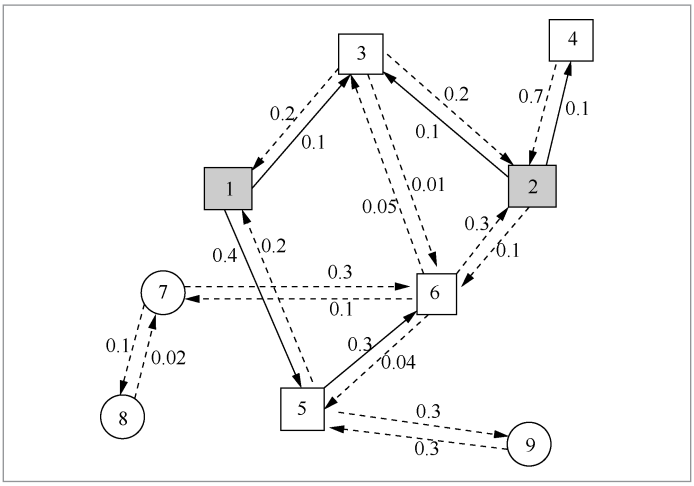



图2 独立级联模型示意

图2给出了独立级联模型一次传播结果的示意。实心方框表示种子节点，空心方框表示传播结束时被激活的节点；圆圈表示未被激活的节点；实线边表示影响力在该边上成功传播，虚线边表示影响力未在其上传播；边上的数字是该边上影响力传播的概率。在$t$=0时刻，种子节点1和2被激活；在$t$=1时刻，节点1、2分别激活节点5、4，并且同时激活了节点3；在$t$=2时刻，节点5成功激活节点6但没有成功激活节点9；在$t$=3时刻，节点6没有成功激活节点7；传播至此结束，节点7、8和9没有在这次传播中被激活。

用$S_\infty$表示在传播过程结束时所有活跃节点的集合。如果总节点数为$n$，而每一步至少激活一个新节点，则在这个模型下传播最多在$n-1$步后结束，即$S_{n-1}=S_\infty$。由于传播过程是随机过程，因此$S_\infty$是随机集合。在影响力传播中经常关心的是传播结束后被激活节点个数的期望值，即$\mathbb{E}[|S_\infty|]$，用$\sigma(S_0)$表示，并称之为(最终)影响力延展度(influence spread)。

注意到在独立级联模型中，任何一个节点$u$对它的任何一个出邻居$v$只有一次尝试激活机会，且发生在节点$u$刚被激活的下一时刻。这看起来似乎是模型的一个局

限。但如果只关心最终的影响力延展度，一个节点$u$在何时尝试激活另一节点$v$或者是否多次尝试激活节点$v$并不重要，只要用$p(u,v)$表示节点$u$多次尝试激活节点$v$的总成功概率，影响力延展度和引入多次激活尝试的扩展模型下的延展度是一样的[9]。如果要考虑中间某时刻的影响力延展度，也可将独立级联模型进行适当扩展，以使其更适合实际情况[12]。

独立级联模型抽象概括了社交网络中人与人之间独立交互影响的行为。它通过边上的概率来描述人与人之间发生影响的可能性和强度。很多简单实体（如新消息在在线网络的传播或新病毒在人际间的传播）很符合独立传播的特性[13]。独立级联模型也在基于实际数据的影响力学习中被初步验证是有效的。所以独立级联模型是目前研究最广泛、最深入的模型。

### 2.1.2 线性阈值模型

在线性阈值模型中，每条有向边$(u,v)\in E$上都有一个权重$w(u,v)\in[0,1]$。直观上说，$w(u,v)$反映了节点$u$在节点$v$的所有入邻居中影响力的重要性占比。要求$\sum_{u\in N^-(v)}w(u,v)\leqslant 1$。每个节点$v$还有一个被影响阈值$\theta_v\in[0,1]$，这个阈值在0到1的范围内均匀、随机地选取，一旦确定在传播中就不再改变。与独立级联模型一样，在$t=0$时刻有且仅有种子集合$S_0$中的节点被激活。在之后每个时刻$t\geqslant 1$，每个不活跃节点$v\in V\setminus S_{t-1}$都需要依据它所有已激活的入邻居到它的线性加权和是否已达到它的被影响值来判断是否被激活，即是否满足$\sum_{u\in N^-(v)\cap S_{t-1}}w(u,v)\leqslant\theta_v$；若是，则节点$v$在时刻$t$被激活（$v\in S_t$）；否则，节点$v$仍然保持不活跃状态。当某一时刻不再有新的节点被激活时，传播过程结束。

线性阈值模型中节点$v$的阈值$\theta_v$表达了节点对一个新实体的接受倾向：阈值越高，节点$v$越不容易被影响；反之，阈值越低越容易被影响。节点$v$的入邻居对节点$v$的影响是联合发生的，可能任何一个入邻居都不能单独激活节点$v$，但几个入邻居联合起来就可能使对节点$v$的影响力权重超过节点$v$的阈值，从而激活节点$v$。这对应了人类行为中在面对一个相对复杂选择时（如购买新型手机、选择移民、参与暴乱等）经常出现的从众行为[2,13]，也是与独立级联模型相比最主要的不同点。

线性阈值模型的随机性完全由节点被影响阈值的随机性所决定，一旦随机阈值被确定，后面的传播过程完全是确定性的。在线性阈值模型中阈值在0和1之间随机选取，这反映了对节点阈值的不了解。然而，在实际中人的被影响阈值虽然有随机性，但应该在更窄的范围内波动。另一方面，如果用更窄范围的随机阈值（如固定阈值）会使模型的分析和计算难度显著加大[9,11]。所以，线性阈值模型在阈值选取上面临两难选择，这也是这一模型不如独立级联模型应用广泛的一个原因。

### 2.1.3 独立级联和线性阈值模型的推广

Kempe等在独立级联和线性阈值模型的基础上又对其进行了推广[11]，引入了诸如触发模型（triggering model）、通用级联模型（general cascade model）、通用阈值模型（general threshold model）等。总体来讲，是让独立级联模型中的独立概率或线性阈值模型中的线性权重变得更灵活、覆盖更广的传播形式。由于篇幅关系，在这里不再展开介绍。感兴趣的读者请看原文或相关综述[9,11]。

## 2.2 其他传播模型

除了上文介绍的离散时间递进性经典模型，根据不同实际需要还有很多其他模

型，用来刻画社交网络中信息和影响力的传播。在这里只做简要介绍。

#### 2.2.1 连续时间模型

连续时间模型（continuous-time model）将网络中两个相连节点的传播时延用一个连续时间的密度函数表示，这样节点的激活可以在任何连续时间内发生[14]。这个模型避免了对实际数据离散化分段，在数据分析时经常是一种有效模型。现在的研究大多是对独立级联模型的连续化，对线性阈值模型的连续化还有待进一步研究。

#### 2.2.2 传染病模型

顾名思义，传染病模型（epidemic model）集中研究传染病或病毒在人群中的传播[15]，现在也被延伸用来研究信息和影响力传播。经典传染病模型将人的状态分为几类，比如易感S（susceptible）、感染I（infected）、治愈R（recovered）等。然后，根据可行的状态转换定义出不同的模型，如SI 模型描述人从易感变为感染；SIS模型允许人从感染回到易感状态然后再被感染；SIR 模型刻画人从易感变为感染然后再痊愈并永久免疫的情况。传染病模型有考虑人群整体行为的，也有基于人际之间接触网络的。前面介绍的独立级联模型与SIR模型在网络中的传播基本具有相同的性质。

#### 2.2.3 选举模型

选举模型（voter model）原是统计物理里一个常用的模型，现在也被用到社交网络影响力传播的研究中[16,17]。在最基本的选举模型中每个节点有两个状态，每个节点$u$在每个离散时刻从它的邻居节点中随机挑选一个节点$v$，将节点$v$在上一时刻的状态作为自己的当前状态。这一过程类似于社交网络中人们通过和朋友交流而采纳朋友意见的过程，所以选举模型和它的变种常用来刻画人们的看法、意见等在社交网络中的演变。因为节点的状态可在多个状态中反复变化，所以选举模型属于非递进性模型，一般用于分析在某一时间点或稳态下的状态分布和相关性质。

#### 2.2.4 博弈论模型

博弈论模型（game-theoretic model）将每一个节点描述为利益最大化的自私节点，其状态就是它的博弈策略。用于刻画传播的网络博弈论模型经常将每条边描述为其两个顶点的一个协调博弈（coordination game），当两个顶点选取同一策略时各自的收益都最大[18,19]。这种模型反映了人际之间的趋同效应和某些产品的网络外部效应（network externality），比如双方都用Skype作为网络通信工具对双方都有益处。这种模型在某节点的一次状态转换过程类似于阈值模型，而状态的反复交替又与选举模型有类似性质。

#### 2.2.5 多实体传播模型

在网络中很可能有多个实体同时传播它们的影响力，它们之间有可能是相互竞争的关系（比如小米手机和iPhone，或者关于某热点事件的官方消息和谣言等），也有可能是互补合作关系（比如iPhone和Apple Watch、微软视窗操作系统和联想笔记本电脑等）。多实体的传播会造成更复杂的传播现象和结果。近几年，已有不少工作着眼于将单实体传播模型（如独立级联和线性阈值模型）扩展为多实体传播模型（multi-item diffusion model）[9,20~24]。绝大多数扩展模型只考虑竞争性实体的并发传播。这些扩展在网络传播上基本继承单实体的传播模型，但在节点上设置先来

先用、后来放弃的规则，并辅以同时到达时的打破平局（tie-breaking）规则。Lu等人最近将多实体的竞争性模型又进一步扩展为既可以描述竞争也可以描述互补合作的比较影响力传播模型（comparative influence diffusion model）[24]。该模型利用节点自动机和少数几个参数刻画了节点在接受一个实体前后会接受另一个实体的不同概率。参数的不同取值范围可以囊括从完全竞争到部分竞争、相互独立、部分互补和完全互补的各种可能情况。总的来说，由于多实体传播模型引入了更复杂的交互和传播机制，模型的性质分析和其上的优化问题等也变得更为复杂。

## 3 影响力最大化问题

影响力传播建模的一个主要目的是控制和优化影响力的传播，这其中被广泛研究的一个核心问题就是影响力最大化（influence maximization）问题。本节以独立级联和线性阈值模型为基础，介绍影响力最大化的研究技术和主要成果，并附带介绍其他影响力传播中的优化问题。

### 3.1 影响力最大化问题的定义

影响力最大化是在给定社交网络结构$G=(V, E)$、影响力传播模型及其参数（如独立级联模型和边上的概率）的情况下，选择$k$个节点作为种子节点集合$S^*$，使得以$S^*$为种子节点产生的影响力延展度$\sigma(S^*)$最大，即$S^*=\text{argmax}_{S\subseteq V,|S|=k}\sigma(S)$。

影响力最大化问题是对病毒营销（viral marketing）的一个直接数学刻画。比如一个厂家要推广产品，希望用病毒式营销手段，先选择网络中少数人送以免费试用产品，希望选中的人试用以后喜欢新产品并主动在其朋友圈推广，使得更多的人接受和购买该产品，而这些新用户又会在他们的朋友圈中进一步推广该产品。厂家的期望是，基于对网络中影响力传播的了解（参见第4节影响力传播学习），能够找出接受试用产品的最佳用户（种子节点），使得最终接受产品的人最多（影响力延展度最大）。这个问题正是影响力最大化的优化目标。

### 3.2 子模函数（submodular function）和影响力最大化的贪心算法技术

上述影响力最大化问题属于组合优化问题，更具体地说，影响力最大化在经典的独立级联和线性阈值模型下都属于图上覆盖问题的一种扩展，因而与图覆盖问题一样，在这些模型下影响力最大化是NP难的问题。解决NP难优化问题的一个重要方法是利用有效的近似算法，比如即使找不到使影响力延展度达到最大的种子集合，但可能找到一个较好的集合，使得该集合的影响力延展度接近最优值，而两者之间的比例就是近似算法的近似比。影响力最大化的近似算法设计核心依赖于影响力延展度函数的子模性质和其带来的贪心算法技术。

对于一个将有限集合$V$的任意子集映射到实数值的函数$f: 2^V \to \mathbb{R}$，称$f$满足子模性，对于任意一个子集$S\subseteq V$和它的任意一个超集$T$（$S\subseteq T\subseteq V$）以及$T$外的任意一个元素$u\in V\setminus T$，$f$满足$f(S\cup\{u\})-f(S)\geqslant f(T\cup\{u\})-f(T)$。子模性反映了元素$u$在集合$S$基础上的增量效应随着$S$的增大而递减，这就是在经济学中经常用到的边界效用递减现象。很多图覆盖问题都具有子模性，因为覆盖的重叠现象会造成边界效用递减。重要的是，影响力延展度作为种子

集合的函数$\sigma(S)$已被证明在独立级联和线性阈值模型以及它们的很多扩展模型下都满足子模性[11]。

和子模性经常在一起使用(但非绝对必要)的还有集合函数的单调性,称集合函数$f$满足单调性,对于任意一个子集$S\subseteq V$和它的任意一个超集$T(S\subseteq T\subseteq V)$,$f$满足$f(S)\leqslant f(T)$。影响力延展度函数$\sigma(S)$同样具有单调性。

一个单调子模函数的重要性质是可以用如下的贪心算法得到函数最大值的近似解。

算法:单调子模函数的贪心算法。

输入:单调子模函数$f$,预算$k$。

输出:大小为$k$的子集$S$。

初始化:$S=\phi$

for $i$ =1 to $k$ do

$v=\arg\max_{u\in V\setminus S}(f(S\cup\{u\})-f(S))$

$S=S\cup\{v\}$

end for

返回$S$

贪心算法分$k$轮,每一轮都要找到一个元素,使得它对已找到的元素来说边界增量最大。如果$f$是单调子模的,且$f(\phi)=0$,则贪心算法找到的贪心解保证至少是最优解的(1-1/e),即大约63%[25]。所以贪心算法是单调子模函数最大化的(1-1/e)的近似算法。

由于影响力延展度函数$\sigma(S)$在独立级联和线性阈值模型下都具有单调性和子模性,且显然$\sigma(\phi)=0$,所以可以用贪心算法来解决影响力最大化问题,以达到的(1-1/e)近似比。

### 3.3 可扩展的影响力最大化算法

然而,第3.2节给出的单调子模函数的贪心算法并未完全解决影响力最大化问题,因为其中的关键一步需要计算一个种子集合$S$的延展度$\sigma(S)$,而计算$\sigma(S)$的精确值本身在独立级联和线性阈值模型下都是很难的问题(技术上称为NP难问题[26,27])。在Kempe等人的论文中[11],简单地提出用随机模拟的方法(通称蒙特卡洛方法)来模拟影响力传播,从而估算$\sigma(S)$的近似解。在这种近似解情况下,贪心算法的解能达到的$(1-1/e-\varepsilon)$近似解,其中$\varepsilon$是一个大于零的数,对应$\sigma(S)$估算的精确性。

但是简单地在影响力最大化中用蒙特卡洛方法有一个严重的问题,就是时间效率很低。在一个不算大的上万个节点的图中,如果对每一次延展度估计都用并不算多的2 000次蒙特卡洛模拟,找出50个种子节点的贪心算法要运行好几天[9]。为了解决这个效率问题,诸多研究提出了各种可扩展的影响力最大化(scalable influence maximization)算法。这些算法基本可分为两大类,一类是利用模型具体特点的启发式算法[26~30],另一类是改进蒙特卡洛方法的贪心近似算法[31~35]。

在启发式算法中,PMIA是一个有代表性的针对独立级联模型的算法[26]。PMIA的主要思想是将在一般图上针对某一节点影响力的传播转化为在该节点附近区域的一棵有代表性的最大影响力传播子树上的传播。这样做的好处是:独立级联模型的影响力延展度计算在树结构上可在线性时间内完成;构造以某一节点为中心的最大影响力子树(maximum influence arborescence)可以用Dijkstra最短路径算法在近线性时间完成;只考虑节点附近的子树会大大减少计算量,同时又不会损失太多计算精度,因为影响力在几步传播后已变弱到可以忽略不计。PMIA和当时已做过优化的蒙特卡洛贪心算法相比,速度提高了1 000倍,而选出种子的影响力在很多实际网络的模拟实验中都很接近

贪心算法。之后，又有不少工作对算法做了进一步改进和提高，比如IRIE算法利用图上整体迭代方法提高了算法速度，同时节省了内存使用[30]。针对线性阈值模型也有LDAG算法[27]和SIMPATH算法[29]。这些算法的优点是速度很快，通常效果也很好，但它们缺乏理论保证，所以究竟它在哪些实际网络中适用，还有待进一步论证。

在对蒙特卡洛贪心算法的改进方面，一种改进是依据延展度函数的子模性利用偷懒估值方法（lazy evaluation）减少对函数估值的次数，如CELF算法[32]。单纯用这种方法虽然对最原始的蒙特卡洛方法有上百倍的提高（运行时间从几天降低到几个小时），但与高效的启发式算法相比还有上千倍的差别（几小时和几秒钟的差别）。最近，由Borgs等人率先提出的反向蒙特卡洛算法改变了这种局面[31]。反向蒙特卡洛算法的核心思想是不从种子节点去模拟估算种子节点的影响力，而是随机选取图上节点，从该节点出发以所有边的相反方向进行蒙特卡洛模拟，得到的集合实际上是最可能影响该节点的集合。这样的集合被称作反向可达集合（reserve reachable set），简称RR集合。而如果一个节点经常在RR集合中出现，那么该节点就是一个影响力大的节点。基于这种思想，Borgs等人理论上证明了他们的算法可达到近乎最优的近线性时间，同时仍有（1-1/e-$\varepsilon$）的近似比保证。之后Tang等人对他们的算法加以改进，提出了TIM/TIM+和IMM算法[33,34]，并进行了模拟实验验证，最新的IMM算法在实验中已超越了启发式算法（如IRIE、SIMPATH）的速度，同时仍有一定的理论保证（$\varepsilon$=0.5，所以理论保证较弱，但对任意图适用）。同时他们指出这种方法适用于独立级联、线性阈值和更广的触发模型。但基于RR集合的这些算法有一个问题是，当选出$k$个种子节点时，算法并不保证也同时找到所有小于$k$个种子集合的近似解。Cohen等人提出的SKIM算法避免了这个问题[35]：SKIM算法通过刻画节点在随机意义下的可达性草图（reachability sketches）来高效计算节点影响力和选择种子节点。理论上，SKIM算法不保证近线性时间但有近似比保证；实验上，它与TIM/TIM+相当。

### 3.4 其他基于影响力的优化问题

基于影响力传播还可以提出很多的优化问题或对模型的拓展。这仍然是现在学术界十分活跃的领域。下面简要介绍一下这方面的几个问题和相关研究。

(1)种子集合最小化

种子集合最小化是影响力最大化的对偶问题。它要求影响力延展度达到一定数值情况下选取的种子集合尽量小。这个问题的解法也是基于单调子模函数的贪心算法，但由于优化目标变为最小化种子集合的大小，近似比变为了$O(\ln\eta)$，其中$\eta$是影响力延展度要求达到的阈值[36,37]。

(2)利润最大化

利润最大化考虑到选取种子有成本，而被影响的非种子节点才会产生收益。所以，利润最大化的目标是选取合适的种子节点（不再受硬性的个数限制），使得最终的期望收益减去种子成本最大。与影响力最大化相比，利润最大化的一个重要区别是它的目标函数（即给定种子集合下的期望利润）不再是单调的。因为当种子集合达到一定程度时，再加一个节点作为种子带来的额外期望收益可能已经不能抵消加入这个种子的费用，但是利润函数仍具有子模性，在这种情况下，利润最大化要利用非单调子模函数的优化技术[38]。

(3)影响力传播监控

影响力传播可能达到网络的各个角落，如何布置有效的监控节点对各种影响力传播提供及时、准确的报告，也是一个重要课题。在技术层面，选择有效的网络监控节点和选择有效的种子节点有相似性，在适当的模型和问题描述下都具有单调性和子模性，所以都可以用贪心算法来解决[32]。

(4)多实体传播模型下的影响力最大化

多实体的传播会给影响力优化带来很多变种。比如在已知一个竞争实体分布的种子节点情况下，如何选取我方的种子节点从而最大化我方的影响力[9]或者尽量减少对方的影响力，也称为影响力阻断最大化(influence blocking maximization)[20,22]。影响力阻断最大化可以应用在抵御谣言的传播。也有学者研究社交网络平台在有多个竞争实体下如何公平分配种子资源的问题[23]。Lu等人在他们最新的研究中还考虑了在互补性实体间的影响力最大化问题[24]，比如在已知一个互补实体的种子节点情况下，如何选取本方实体的种子节点以最大化本方的影响力(即自我影响力最大化(self influence maximization))或者最大化互补的对方的影响力(即互补影响力最大化(complementary influence maximization))。可以看出，多实体传播下的影响力最大化种类繁多，具体应用要具体分析。绝大多数问题仍然基于子模函数的最大化，但是多实体模型在不少情况下不再具备子模性，所以需要寻找新的解决途径。

(5)网络拓扑的优化

影响力传播研究中，也有研究如何有效地改变网络拓扑结构来优化影响力的。比如如何有效删除图中的边或节点使得种子节点的影响力尽量小，这对应了防止传染病传播中的隔离和免疫措施。也可以考虑如何增加点或边以最大化影响力，这在一定程度上对应了社交网络平台朋友推荐的情形。Khalil等人针对一种拓扑变化下定义的目标函数，论证了它的子模性或对称的超模性(supermodularity)，从而用子模或超模函数的优化技术进行处理[39]。值得一提的是，他们定义的一个集合的影响力并不是集合整体的影响力延展度，而是集合中每个个体的影响力延展度的算术平均。这个定义使得他们能够得到对应的子模或超模性结论，但这样的模型只适用于单一种子从种子集合中随机选取的情形。

(6)非子模性的影响力优化问题

当对影响力传播模型进行一定扩展或对优化目标进行一定改变后，新的模型或问题经常就不再具有子模性(或超模性)。在最近的研究中对非子模性的影响力优化问题也提出了一些解决方法，比如利用整数规划[40]，将其转化为相近的子模问题[41]，假设图的一部分对应的带权重的邻接矩阵有常数秩[42]，将非子模函数夹于两个子模函数之间的三明治方法[24]或者利用基于传播模型的启发式算法[43]。这些方法对某些具体问题有较好的效果，但非子模性的影响力优化问题的系统性研究还有待完善。

影响力传播中还有很多其他相关问题和相关算法，受篇幅限制，本文不能面面俱到。

## 4 社会影响力传播学习

前面介绍了影响力传播模型和其上的影响力优化问题。要使影响力传播研究在实际中发挥更大的作用，基于实际数据的影响力学习(influence learning)也是必不可少的一个方面。基于实际数据的网络影响力分析在国内外社交媒体网站也都有出现，比如国外的Klout.com、国内的新浪

微博影响力排名等。这些影响力分析侧重对名人的排名，分析方法大多利用网络拓扑结构（如粉丝数、PageRank）、用户活跃度等。而基于影响力传播的学习是希望从数据中挖掘用户行为的传播方式和对应的参数，从而为影响力传播建模和优化服务。

### 4.1 影响力传播学习的基本思想

在影响力传播学习方面也有不少工作。这些工作基于的数据基本上是两类：一类是社交网络结构的数据，比如微博中用户$B$关注了用户$A$，那么就有一条有向边从用户$A$到用户$B$，边的方向在这里表示信息从用户$A$传向用户$B$，与影响力的方向一致。当收集了大量用户的关注数据后，就可以建立一个关于这些用户的有向图。当然有些网络（如Facebook）对应的是无向图，每条无向边表示的是朋友关系。第二类数据是用户的某一类行为的时间序列，比如一条记录是微博用户$A$在时刻$t_1$发布了一条带有某个链接$L_1$的微博，用$(A,L_1,t_1)$表示。一般来讲，用户的行为序列是由$(u,a,t)$组成的序列，其中，$u$表示一个用户（对应图上一个节点），$a$表示一个动作，$t$表示用户$u$执行动作$a$的时间。

目前来讲，影响力传播学习的基本思想是如果相连的两个用户在相近时间先后执行同样的动作，那么认为这是先执行动作的用户对后执行动作的用户的一次成功影响。比如在上文的微博例子中，如果在记录$(A,L_1,t_1)$后面有一条记录$(B,L_1,t_2)$，而时间$t_2$大于$t_1$但又不大很多，说明在用户$A$发布了包含链接$L_1$的微博不久，关注用户$A$的用户$B$也发布了同样链接的微博，这可被理解为用户$B$看到用户$A$的微博而转发的行为，所以在发布链接这个行为上可以认为用户$B$受到一次用户$A$的影响。如果数据中发现用户$B$经常在用户$A$之后发布与用户$A$相同的链接，那么可以推测在发布链接这类行为上用户$A$对用户$B$的影响力较大。

上述的思想比较直观，但严格地说所发现的是用户行为的相关性，并不能直接反映影响力的因果关系。比如上述微博例子中也有可能是用户$B$并未看到用户$A$的微博，或者即使看到，用户$B$发同样微博是因为用户$B$和用户$A$都对同一类链接内容感兴趣，而并不是因为用户$B$受到用户$A$的影响，这称为社会关系中的同质性（homophily）。在一组收集数据中要区分相关性行为的来源是同质性还是影响力并不是一件容易的事情。为此，Anagnostopoulos等人提出了洗牌测试（shuffle test）的方法[44]，将实际发生事件的时间顺序像洗牌一样随机打乱后，再观察关于这个序列的某些特征值是否改变。如果发生改变，说明实际的时间顺序是重要的，这是支持影响力的因果关系造成实际事件顺序的证据；而如果不发生改变，说明时间顺序并不重要，这是支持由同质性造成的相关性事件序列的证据。洗牌测试对判定影响力的存在性有一定作用，但在区分影响力和同质性方面仍有不少需要进一步完善的工作要做。

在影响力传播中下一个要解决的问题是在一个节点执行一个动作之前，有多个该节点的邻居节点都执行了同样动作，在这种情况下如何判定是哪一个或哪几个邻居节点真正影响了该节点？现有的方法基本分两种：一种是用最大似然估计（maximum likelihood estimate），一种是基于信用分配（credit distribution）的频度分析（frequency analysis）。

### 4.2 最大似然估计

最大似然估计是基于一个随机传播模

型（如独立级联模型）得到一次传播结果的似然度，然后求得参数使得实际出现的传播结果似然度最大[45,46]。直观上说，虽然一个节点有可能被多个邻居节点影响，但如果实际数据中一个节点的动作经常跟随它的某一个邻居节点的动作，这说明这个特定节点对它的影响力可能较大。最大似然估计就是将这一想法严格数学化的方法。

直接应用最大似然估计很可能在图中很难计算，通常会用中间变量和期望最大化迭代的EM算法[46]。但这种算法在大图中效率不高，且不一定保证能收敛到全局最优解。Netrapalli和Sanghavi对最大似然估计做了改进，将其计算变为一个凸规化（convex program）问题，从而能有效求解且保证全局最优[45]。

### 4.3 信用分配和频度分析

最大似然估计的形式化和计算仍然比较复杂，对此Goyal、Bonchi和Lakshmanan提出了基于信用分布的频度分析方法[47]。它的基本思想是当需要决定在一次传播中究竟是哪个已被激活的邻居节点激活了一个节点时，将部分信用积分（partial credit）平摊到所有参与的邻居节点中（每次的总信用为1）。这种信用积分的分配可以是完全平均，也可以不平均，比如激活时间上离被激活节点时间最近的信用积分最高。这种简单的分配方式虽然是启发式的，但避免了复杂的最大似然分析。当部分信用积分分配对所有的传播实例都完成后，一个节点对它的邻居节点的影响力就由直接的频度分析得到，也即从得到的信用积分总和除以在数据中总共被激活的次数，这个比值表示了当被激活后被激活的频度，而这个频度考虑了对的部分信用积分。这种计算方法效率很高，适合于大规模图的学习。

影响力传播学习并不一定需要知道社交网络的图结构。在缺乏图结构时，认为任何在激活时间上相接近的两个节点都有可能存在边而发生传播。这相当于把图看成是全连通图。在学习结束后可以把权重很低的边删掉，从而一定程度上恢复原图。如果已知原图，则学习的效率和准确度都会大大提高。但从另一方面讲，社交网络中的图结构并不能准确表达所有的传播路径，不基于图结构的影响力传播学习可能会挖掘出隐含的影响力关系，也有它的好处。另外，影响力的传播在不同领域和不同话题下经常是不一样的，为此Barbieri等提出了与话题相关的影响力传播模型和在其上的学习方法[48]。

## 5 影响力传播研究和应用的问题、挑战和方向

影响力传播研究经过本世纪十几年的发展，已经取得长足的进步，使大家对影响力传播的模式和其上的优化问题都有了较深的认识。但是进一步发展其研究和应用，还要解决很多问题。

其中一个主要问题是影响力传播学习方面的准确、有效问题，这仍然是当前一个很大的挑战。与很多大数据分析不同，影响力传播的大数据分析要求分析的是任意两个关联用户之间的影响力强度，这比只分析一个用户的特征或一个群体的特征难度要大很多。不仅如此，影响力传播涉及对人的行为分析，而且是较为复杂的如产品购买、接受新思想等行为，这种行为数据在社交媒体数据中并不容易挖掘，因为大多数社交媒体数据都是无意义的噪声，而诸如转发等的行为传播又过于简单，与真正针对产品、思想等的行为传播可能

很不同。而且如前文所述，从数据中区分影响力和同质性也是一个较难的问题。所以，在影响力传播的研究中影响力传播的有效分析是目前的一大瓶颈。简单地说，就是在这方面大数据还远不够大，在真正理解和分析用户行为的大规模传播方面还有很多路要走。

在影响力建模方面，已发展出很多模型，其中以独立级联模型为代表的一些模型在实际数据中也得到一定程度的印证。但是目前为止，对于更适于描述复杂传播行为的阈值模型还缺乏实际数据的有效验证。线性阈值模型对阈值的随机性要求有局限性，而如果用更一般的阈值模型很可能会使模型不具备子模性等性质，从而无法设计有效的算法。所以对于阈值模型，从数据分析到建模和优化还都有不少问题要解决。

另外，绝大多数影响力传播研究都是在静态网络中进行，而实际网络都是动态变化的。如何将传播的动态性和网络的动态性合理结合，以达到有效的分析、建模和优化，也是一个需要更多关注的课题。

在影响力优化方面，其应用有效性还需实际检验。这是因为影响力优化需要因果关系的验证，而这通常需要在实际系统中进行随机可控试验（randomized controlled experiment）才能真正验证。绝大多数研究者还不具备大规模的社交网络平台和影响力传播数据用以实施这样的试验。所以如何加强合作，构建这样的共享平台和共享大数据，是让影响力传播和最大化研究走出实验室得以广泛应用的关键课题。

尽管存在很多问题和挑战，影响力传播的研究仍然蓬勃发展，甚至展示了它在一些意料之外方面的应用。比如Shakarian等人将影响力最大化应用到芝加哥警察局挑选暴力团伙成员参加学习劝导班，使其影响其他团伙成员远离暴力犯罪[49]，而Wang等人将影响力传播模型和最大化借用到文本概括（text summarization）领域，通过建立单词之间的一个影响网络来帮助文本概括[50]。随着大数据技术的发展和影响力传播研究的深入，影响力传播研究会有更广泛的应用前景。

## 6 结束语

本文将影响力传播研究分为三大方面：影响力传播模型、影响力传播学习和影响力传播优化，并对3个方面的主要成果和近期进展进行了介绍。简而言之，影响力传播研究通过建立人们行为的传播模型，从实际数据中学习传播模型及其参数和基于传播模型的各种影响力优化和控制技术，使大家对影响力的传播机理和模式有了深入的了解，并将这种认识和理解转化为对传播行为的预测、优化和控制。本文也讨论了当前影响力传播研究和应用方面的问题和挑战，比如如何利用更大规模的数据来支持影响力传播的研究、如何结合网络的动态性、如何在实际中检验优化结果等。随着大数据研究和应用的不断深入和发展，影响力传播的研究也会取得更加丰硕的成果，并在产业界和实际生活中得到广泛的应用。

## 参考文献

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

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

[3] 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

[4] Christakis N A, Fowler J H. The collective dynamics of smoking in a large social network. New England Journal of Medicine, 2008, 358(21): 2249~2258

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

[6] Bond R M, Fariss C J, Jones J J, *et al*. A 61-million-person experiment in social influence and political mobilization. Nature, 2012(489): 295~298

[7] Charu C, Aggarwal. Social Network Data Analysis. New York: Springer, 2011: 177~214

[8] 吴信东, 李毅, 李磊. 在线社交网络影响力分析. 中国计算机学报, 2014, 37(4): 735~752
Wu X D, Li Y, Li L. Influence analysis of online social networks. Chinese Journal of Computers, 2014, 37(4): 735~752

[9] Chen W, Lakshmanan L V S, Castillo C. Information and Influence Propagation in Social Networks. California: Morgan & Claypool Publishers, 2013

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

[11] Kempe D, Kleinberg J M, 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 (KDD), Washington DC, USA, 2003: 137~146

[12] Chen W, Lu W, Zhang N. Time-critical influence maximization in social networks with time-delayed diffusion process. Proceedings of the 26th National Conference on Artificial Intelligence (AAAI), Toronto, Canada, 2012

[13] Centola D, Macy M. Complex contagion and the weakness of long ties. American Journal of Sociology, 2007, 113(3): 702~734

[14] Gomez-Rodriguez M, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks. Proceedings of the 28th International Conference on Machine Learning (ICML), Bellevue, Washington, USA,2011: 561~568

[15] Newman M E J. Networks: an Introduction. Oxford: Oxford University Press, 2010

[16] Even-Dar E,Shapira A. A note on maximizing the spread of influence in social networks. Proceedings of the 3rd Workshop on Internet and Network Economic (WINE), San Diego, USA, 2007: 281~286

[17] Li Y, Chen W, Wang Y, *et al*. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy, 2013: 657~666

[18] Immorlica N, Kleinberg J M, Mahdian M, *et al*. The role of compatibility in the diffusion of technologies through social networks. Proceedings of the 8th ACM Conference on Electronic Commerce (EC), San Diego, USA, 2007: 75~83

[19] Montanari A,Saberi A. Convergence to equilibrium in local interaction games. Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS), Atlanta, USA, 2009: 303~312

[20] Budak C, Agrawal D, Abbadi A E. Limiting the spread of misinformation in social networks. Proceedings of the 20th International Conference on World Wide

Web (WWW), Hyderabad, India, 2011: 665~674

[21] Chen W, Collins A, Cummings R, *et al*. Influence maximization in social networks when negative opinions may emerge and propagate. Proceedings of SIAM International Conference on Data Mining, Mesa, USA, 2011: 379~390

[22] He X, Song G, Chen W, *et al*. Influence blocking maximization in social networks under the competitive linear threshold Model. Proceedings of SIAM International Conference on Data Mining, Anaheim, USA, 2012: 463~474

[23] Lu W, Bonchi F, Goyal A, *et al*. The bang for the buck: fair competitive viral marketing from the host perspective. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Chicago, USA, 2013: 928~936

[24] Lu W, Chen W, Lakshmanan L V S. From competition to complementarity: comparative influence diffusion and maximization. Proceedings of the 42nd International Conference on Very Large Data Bases (VLDB), New Delhi, India, 2016 Accepted

[25] Nemhauser G, Wolsey L, Fisher M. An analysis of the approximations for maximizing submodular set functions. Mathematical Programming, 1978(14): 265~294

[26] Wang C, Chen W, Wang Y. Scalable influence maximization for independent cascade model in large-scale social networks. Data Mining and Knowledge Discovery, 2012, 25(3): 545~576

[27] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold Model. Proceedings of the 10th IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 2010: 88~97

[28] 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 (KDD), Paris, France, 2009: 199~208

[29] Goyal A, Lu W, Lakshmanan L V S. SIMPATH: an efficient algorithm for influence maximization under the linear threshold model. Proceedings of the 11st IEEE International Conference on Data Mining (ICDM), Vancouver, Canada, 2011: 211~220

[30] Jung K, Heo W, Chen W. IRIE: scalable and robust influence maximization in social networks. Proceedings of the 12nd IEEE International Conference on Data Mining (ICDM), Brussels, Belgium, 2012: 918~923

[31] Borgs C, Brautbar M, Chayes J, *et al*. Maximizing social influence in nearly optimal time. Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA), Portland, USA, 2014: 946~957

[32] Leskovec J, Krause A, Guestin C, *et al*. Cost-effective outbreak detection in networks. Proceedings of the 13rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Jose, USA, 2007: 420~429

[33] Tang Y, Shi Y, Xiao X. Influence maximization in near-linear time: a martingale approach. Proceedings of ACM SIGMOD Conference (SIGMOD), Melbourne, Australia, 2015: 1539~1554

[34] Tang Y, Xiao X, Shi Y. Influence maximization: near-optimal time complexity meets practical efficiency. Proceedings of ACM SIGMOD Conference (SIGMOD), Snowbird, USA, 2014: 75~86

[35] Cohen E, Delling D, Pajor T, *et al*. Sketch-based influence maximization and computation: scaling up with guarantees. Proceedings of the 23rd ACM International

Conference on Information and Knowledge Management (CIKM), Shanghai, China, 2014: 629~638

[36] Goyal A, Bonchi F, Lakshmanan L V S, *et al*. On minimizing budget and time in influence propagation over social networks. Social Network Analysis and Mining, 2012, 2(1)

[37] Long C, Wong R CW. Minimizing seed set for viral marketing. Proceedings of the 11st IEEE International Conference on Data Mining (ICDM), Vancouver, Canada, 2011: 427~436

[38] Lu W,Lakshmanan L V S. Profit maximization over social networks. Proceedings of the 12nd IEEE International Conference on Data Mining (ICDM),Brussels, Belgium, 2012: 479~488

[39] Khalil E, Dilkina B, Song L. Scalable diffusion-aware optimization of network topology. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), New York, USA, 2014: 1226~1235

[40] Goldberg S, Liu Z. The diffusion of networking technologies. Proceedings of the 24th ACM-SIAM Symposium on Discrete Algorithms (SODA), New Orleans, USA, 2013: 1577~1594

[41] Zhang P, Chen W, Sun X, *et al*. Minimizing seed set selection with probabilistic coverage guarantee in a social network. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), New York, USA, 2014: 1306~1315

[42] Chen W, Li F, Lin T,*et al*. Combining traditional marketing and viral marketing with amphibious influence maximization. Proceedings of the 16th ACM Conference on Economics and Computation (EC), Portland, USA, 2015: 779~796

[43] Yang DN, Hung HJ, Lee WC, *et al*. Maximizing acceptance probability for active friending in online social networks. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Chicago, USA, 2013: 713~721

[44] 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 (KDD), Las Vegas, USA, 2008: 7~15

[45] Netrapalli P, Sanghavi S. Learning the graph of epidemic cascades. Proceedings of ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS), London, UK, 2012: 211~222

[46] Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model. Proceedings of the 12nd International Conference on Knowledge-based Intelligent Information and Engineering Systems (KES), Zagreb, Croatia, 2008: 67~75

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

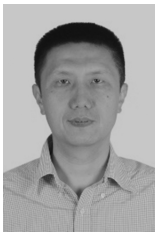
[48] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowledge Information Systems, 2013, 37(3): 555~584

[49] Shakarian P, Salmento J, Pulleyblank W, *et al*. Reducing gang violence through network influence based targeting of social programs. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), New York, USA, 2014: 1829~1836

[50] Wang C, Yu X, Li Y, *et al*. Content

coverage maximization on word networks for hierarchical topic summarization. Proceedings of the 22nd ACM International Conference on Information and Knowledge Management(CIKM), San Francisco, USA, 2013: 249~258

## 作者简介



**陈卫**，男，微软亚洲研究院高级研究员，清华大学客座教授，中国科学院计算所客座研究员，多个国际顶级数据挖掘和数据管理会议（KDD、WSDM、SIGMOD、ICDE、WWW等）的程序委员会成员，中国计算机学会大数据专家委员会首批成员，《大数据》期刊编委。近期主要研究方向包括社交与信息网络算法和数据挖掘、网络博弈论和经济学、在线学习等。近几年在社会影响力最大化方面的一系列开创性研究成果，在KDD、ICDM、SDM、WSDM、ICWSM、AAAI、VLDB等顶级数据挖掘、人工智能和数据库学术会议上发表后得到良好反响，并引发这一方向众多的后续工作。最早发表的KDD' 2009论文被引用次数排同会议所有论文第二位，而第二篇KDD' 2010论文被引用次数排同会议所有论文第一位。2013年与另外两位合作者合写了一部关于影响力传播和最大化的专著（Information and Influence Propagation in Social Networks, Morgan & Claypool, 2013），系统总结了这方面的研究成果和最新发展。另外，在与社会和信息网络相关的方向，如社区检测、网络中心化度量排序、网络博弈、网络定价、网络激励机制等方面也都做出开创性的工作，其中将博弈论引入网络社区检测的论文获得了2010年欧洲机器学习及数据挖掘会议最佳学生论文奖。

**收稿日期：** 2015-08-26
**基金项目：** 国家自然科学基金重点项目（No.61433014）
**Foundation Item:** The National Natural Science Foundation of China (No.61433014)
**论文引用格式：** 陈卫. 社交网络影响力传播研究. 大数据, 2015031
Chen W. Research on influence diffusion in social network. Big Data Research, 2015031