基于最小点覆盖和反馈点集的社交网络影响最大化算法

许宇光[①] 潘惊治[①] 谢惠扬*[②]
[①](北京大学信息科学技术学院 北京 100871)
[②](北京林业大学理学院 北京 100083)

摘 要：社交网络中的影响最大化问题是指在特定的传播模型下，如何寻找 k 个最具影响力的节点使得在该模型下社交网络中被影响的节点最多，信息传播的范围最广。该问题是一个优化问题，并且已经被证明是 NP-难的。考虑到图的最小点覆盖和反馈点集中的顶点对图的连通性影响较大，该文提出一种基于最小点覆盖和反馈点集的社交网络影响最大化算法(Minimum Vertex Covering and Feedback Vertex Set, MVCFVS)，并给出了具体的仿真实验和分析。实验结果表明，与最新的算法比较，该算法得到的节点集在多种模型下都具有优异的传播效果，例如在独立级联模型和加权级联模型中超过当前最好的算法，并且还具有更快的收敛速度。

关键词：社交网络；影响最大化；传播模型；最小点覆盖；反馈点集
中图分类号：TP393 **文献标识码**：A **文章编号**：1009-5896(2016)04-0795-08
DOI: 10.11999/JEIT160019

Minimum Vertex Covering and Feedback Vertex Set-based Algorithm for Influence Maximization in Social Network

XU Yuguang[①] PAN Jingzhi[①] XIE Huiyang[②]
[①](*School of Electronics Engineering and Computer Science*, *Peking University*, *Beijing* 100871, *China*)
[②](*College of Science*, *Beijing Forestry University*, *Beijing* 100083, *China*)

Abstract: Influence maximization is an optimization issue of finding a subset of nodes under a given diffusion model, which can maximize the spread of influence. This optimization issue has been proved to be NP-hard. Leveraging the fact that vertices in minimum vertex covering and feedback vertex set are of great importance for the connectivity of a graph, a heuristic algorithm for influence maximization based on Minimum Vertex Covering and Feedback Vertex Set (MVCFVS). Extensive experiments on various diffusion models against state of the art algorithms are carried out. Specifically, the proposed algorithm performs excellent on Independent Cascade Model (ICM) and Weighted Cascade Model (WCM), which exhibits its great advantages in terms of influence range and convergent speed.

Key words: Social network; Influence maximization; Diffusion models; Minimum vertex covering; Feedback vertex set

1 引言

社交网络是指由个体及个体之间的关系所组成的一个复杂网络，它与交通网络，通讯网络和生物网络等其他复杂网络相比，包含了更加海量和多元化的信息。自从社交网络出现以来[1,2]，它便在社会个体的信息传播、思想引导和相互影响中发挥着重大作用。近年来，随着大规模在线社交网络(如人人，Facebook, Twitter和微博等)的迅速发展，从个人到个人和从个人到群体的相互作用中探索社会影响引起了人们的广泛兴趣[3−6]，这是因为社会影响可以作为一种微妙的力量控制社交网络的动态性。为此，关于在大规模社会网络中挖掘对信息和思想的传播有影响的个体集的研究受到了大量学者的青睐。其中，一个关键问题是影响最大化问题，即如何选择 k 个初始节点，使得它们在社交网络中的影响最大化。

关于影响力最大化算法的研究，目前主要有基于贪心思想的方法和启发式方法。其中，基于贪心思想的方法选出的节点传播效果较好，但是选择节点时算法效率较低，因此这类方法目前主要的研究方向是如何降低算法的运行时间，提高算法效率。文献[7,8]首次将影响最大化问题引入到社交网络

收稿日期：2016-01-15；改回日期：2016-02-26；网络出版：2016-03-09
*通信作者：谢惠扬 xhyang@bjfu.edu.cn
基金项目：国家自然科学基金(61370193)
Foundation Item: The National Natural Science Foundation of China (61370193)

中，他们考虑了个体之间的社会关系并提出了一种概率信息传播模型。随后，文献[9,10]首次将影响最大化问题描述成离散优化问题，并在两个不同的模型下，即线性值模型和独立级联模型，研究了此问题。他们证明了影响最大化问题在上述两个模型下是 NP-难的。同时，他们提出了一种贪心算法，并证明了所提算法在这两种模型下的性能比为 $1-1/\mathrm{e}$。考虑到贪心算法效率不高的问题，文献[11]提出了“Lazy-forward”的优化策略来选择初始节点。2014 年，文献[12]在研究基于线性阈值模型下的影响最大化问题时，为线性阈值的传播方程推导出了理论上界[13]。2014 年，文献[14]提出贪心算法本质上是一种自一致性排序，即自身的排序和影响范围的增益相一致。

通常启发式方法选出的节点传播效果不如基于贪心思想的方法，但启发式方法算法效率较高，因此这类方法目前主要的研究方向是如何在保持其效率较高优势的前提下，改善选出节点的传播效果。文献[15]基于度提出了“Degree Discount”方法。文献[16]根据模拟退火法启发式求解影响最大化问题。文献[16–18]首次将潜伏限制加到线型阙值模型下的影响最大化问题中，并称其为快速信息传播问题(fast information propagation problem)。他们证明了该问题是NP-难的，并给出了两个启发式的算法来求解该问题。近年来，文献[19]提出了概括影响力公式的问题。文献[20]研究了社区结构和影响最大化问题的关系。文献[21]提出一种基于k-核的社会网络影响最大化算法。文献[22]提出一种在独立级联模型下估计节点级联影响力的方法。

基于上述考虑以及图最小点覆盖和反馈点集中顶点的重要性，本文提出了一种基于最小覆盖集和反馈点集的近似求解影响最大化的算法(Minimum Vertex Covering and Feedback Vertex Set, MVCFVS)。该算法同时考虑了最小点覆盖和反馈点集中顶点的影响力，具有很好的效果。实验结果表明，所提算法可以较快地找到具有较好影响范围的节点集。本文第 2 节介绍 3 种常用的传播模型；第 3 节介绍与本算法相关的概念，详细描述本文的算法，并对本算法的时间复杂度进行分析；第 4 节介绍实验设计及实验结果分析；第 5 节对本文成果进行了概括并探讨未来的工作。

2 传播模型

在研究社交网络时，社交网络通常被抽象成一个有向(无向)图，图中的节点代表参与社会活动的人，边代表人与人之间的联系。对于给定的社会网络，在网络中寻找影响力节点集，需要借助于相应的传播模型。线性阈值模型、独立级联模型和加权级联模型是3个常用的传播模型。在这些模型中，节点有活跃和不活跃两种状态可以选择。其中个体处于活跃状态时，表示该个体接受了这个信息；处于不活跃状态时，表示该个体没有接受这个信息。随着不活跃节点的活跃邻居数目的增加，节点也越倾向于变为活跃状态。

线性阈值模型(Linear Threshold Model, LTM) 在线性阈值模型中，一个节点是否受到影响从不活跃状态变为活跃状态是由其邻居的共同影响力决定的。对于节点 w 的邻居节点 v，将 v 对 w 的影响记为 $b_{w,v}$，则有 $\sum_v b_{w,v} \le 1$。在该模型中，如果节点 w 受活跃邻居的影响总和超过某个阈值 θ_w，则 w 由不活跃状态变为活跃状态。即满足公式 $\sum_{w\in AN(v)} b_{w,v} \ge \theta_w$ 时，节点 w 被激活(即由不活跃状态变为活跃状态)，其中$AN(v)$表示v的活跃邻居节点集。可以看出，θ_w 越大，则节点 w 越不容易被激活。因此 θ_w 可以反映出节点 w 被激活的倾向性[23]。

独立级联模型(Independent Cascade Model, ICM) 在独立级联模型中，节点只有在刚被激活时才可以尝试去激活其邻居。假设 v 是一个在时刻 t 刚被激活的节点，则对于 v 的每个不活跃邻居 w，v 可以以概率 $p_{v,w}$ 激活 w。若激活成功，则节点 w 在时刻 $t+1$ 变为活跃节点；若不成功，w 仍然保持不活跃状态。无论 w 是否成功激活其邻居，在 $t+1$ 以后的时刻 v 都不能再尝试激活其他节点。如果在时刻 t 不活跃节点 w 有多个邻居都刚被激活，则其邻居节点对其的激活顺序对于最后结果没有影响。概率 $p_{v,w}$ 不依赖于之前所有对 w 的激活尝试。传播过程以这种方式不断进行直到没有刚被激活的节点时停止[10]。

加权级联模型(Weighted Cascade Model, WCM) 加权级联模型可以看作是独立级联模型的一个特例。在该模型中，节点 v 激活其不活跃邻居节点 w 的概率 $p_{v,w}$ 与节点 w 的度 $d(w)$ 有关，$p_{v,w} = 1/d(w)$。故当一个节点有很多邻居时，每个邻居对其的影响就会被平均到一个非常小的值，这在某种程度上可以反映出真实世界中的人际关系。例如，如果一个人只有一个朋友，那么这个朋友对他的建议就非常具有影响力，而如果他有很多朋友，那么其中一个朋友的建议对他作何决定的影响并不大[10]。

3 MVCFVS 算法

由于社交网络通常被抽象成图来研究，所以本文的算法在理论上把图作为研究对象，最后在真实

的社交网络数据集上进行实验来验证。

本文所言之图皆指无向有限连通简单图(无环无重边)。对于任意图 G，用 $V(G)$和 $E(G)$分别表示图 G的顶点集和边集。图 G的一个点覆盖是指 $V(G)$的一个顶点子集 K，使得 $E(G)$中的每条边至少有一个端点在 K 中。如果 G 不存在任何覆盖 K'满足 $|K'|<|K|$，那么称 K 是 G 的最小点覆盖。求一个图的最小点覆盖并非易事，该问题是 NP-完全的[24]。

对于一个图 G，令 v 是 G中的一个顶点。我们用 $N_G(v)$表示 G中与 v相邻的所有顶点构成的集合。顶点 v 的度，记作 $d_G(v)$，定义成 $N_G(v)$含有顶点的个数，即 $d_G(v)=|N_G(v)|$。若 $d_G(v)=k$，那么称 v 是 G 的一个 k-点。$\Delta(G)$ 和 $\delta(G)$ 分别表示图 G 的最大度和最小度。若 G中任意两个顶点之间都存在一条路，那么 G称为是连通的；否则，G称为不连通。如果 G是不连通的，那么 G至少含有两个连通分支，用 $\omega(G)$表示 G 的连通分支的个数。不含圈的图称为无圈图，连通的无圈图称为树。

树最小点覆盖算法如表 1 所示。

表 1 树最小点覆盖算法

算法 1　树最小点覆盖算法[25]
输入：树 T
　　$C=\varnothing$
输出：T的最小点覆盖 C
步骤 1　找出 T中的所有 1-点构成的集合，记作 V_1；
步骤 2　令 $N(V_1)$表示与 V_1中的顶点相邻的顶点集，并且令 $C=C\cup N(V_1)$；
步骤 3　在 T中删去 V_1和 $N(V_1)$中的顶点及它们的关联边，得到的图记为 T_1；
步骤 4　如果$|V(T_1)|=0$，结束；否则转向步骤 1。

定理 1　对于任意树 T，算法 1 都得到 T 的最小覆盖集。

证明　首先，我们证明 T存在一个最小点覆盖不含 1-点。否则若 T 的某个最小点覆盖 C 含有 1-点，那么将 1-点从 C中删去，然后将与其相邻的顶点加入 C中，从而得到一个新的最小点覆盖。令C'是 T的一个不含 1-点的最小点覆盖，那么C'即是按算法 1 得到的。因为，每个 1-点关联的边如要被覆盖，那么与该 1-点相邻的顶点必定在最小覆盖中。故C'包含算法 1 所选出的所有的顶点。另一方面，容易验证算法 1 选出的顶点集是 T的一个覆盖，所以，结论成立。　　证毕

本文为给出 MVCFVS 算法，需要先找到图的反馈点集，于是本文提出了一个简单的图的反馈点集算法：

令 G 是一个简单无向图。$F\subset V(G)$称为是 G的一个反馈点集如果$G-F$是一个无圈图。可见，G中的每个圈至少有一个顶点在 F中。把含有顶点数最少的反馈点集称为 G的最小反馈点集。

图反馈点集算法如表 2 所示。

表 2 图反馈点集算法

算法 2　图的反馈点集算法 findfeedbackset(G)
输入：图 G
　　$F=\varnothing$
输出：G的反馈点集 F
步骤 1　反复的删去图 G中的 1-点，直到所得之图 G'为空图(即不含边)，或 $\delta(G')\geq 2$；
步骤 2　如果 G'是空图，算法结束，返回 F；如果 $\delta(G')\geq 2$，那么对于 G'的每个分支 G_i，删去 G_i中的最大度顶点 v'，并令 $F=F\cup\{v'\}$。然后返回步骤 1。

定理 2　对于任意图 G，算法 2 都得到 G的一个反馈点集。

证明　首先，对于任意连通图，经过步骤 1 后所得之图都是连通图，因为每次删去的都是 1-点。其次，若一个图含有圈，那么经过步骤 1 后一定不是空图。所以，当算法结束时，即对应的G'是空图，所以G'对应的 G的每个分支都是树，从而，结论成立。　　证毕

下面将应用上述两个算法给出本文求解影响最大化的算法 MVCFVS。我们的思想是：对于一个图 G，首先利用算法 2 求出 G的一个反馈点集 F；其次，对于$G-F$的每个树分支 T_i，利用算法 1 求出树分支的最小点覆盖集 C_i。第三，令$G-F$含有 m个树分支，对于$V'=F\cup\bigcup_{i=1}^{m}C_i$ 中的顶点 x，我们将按下述定义的影响力函数 $t(x)$，从大到小选出最有影响力的 k 个顶点。

$$t(x)=d_G(x)+(N_G(x)\cap(S-M_x))\frac{|V(G)-V'|}{|V'|}$$

式中，M_x 表示在顶点 x之前已经被选出的顶点集。

MVCFVS 算法如表 3 所示。

表 3　MVCFVS 算法

算法 3　算法 MVCFVS
输入：图 G，数 k
输出：大小为 k的节点集 S
步骤 1　求 G的一个反馈点集 F (算法 2)；
步骤 2　求 $G-F$ 的每个树分支的最小点覆盖(算法 1)；
步骤 3　从 $V'=F\cup\bigcup_{i=1}^{m}C_i$ 中，依次选出 $t(x)$ 值最大的 k 个顶点加入 S。

从时间复杂度来看，本文算法的时间复杂度为$O(n^2)$。其中，寻找反馈点集的时间复杂度为$O(n^2)$，因为我们在删除节点时并不需要真的删除，可以巧妙地通过标记处理，而寻找需要删除的目标节点的时间复杂度为$O(n)$，所以寻找反馈点集的时间复杂度可形式化的表示为$T(n)=T(n-1)+O(n)$，即时间复杂度为$O(n^2)$；同理，寻找最小点覆盖的时间复杂度也为$O(n^2)$。本文算法是在基于这两个算法选出的节点的影响力函数值基础上选出来的，最后的时间复杂度为$O(n^2)$。

因为反馈点集中的每个顶点都属于一个圈中，故有理由认为他们的全局影响力比较大。另外，最小点覆盖中的顶点的局部影响力比较大，从而可以认为由此算法筛选出来的 k 个顶点的影响力比较大。下一节将给出具体的实验。

4 实验

为了验证算法的有效性，我们在真实的社交网络数据上进行了实验，用基于最小点覆盖和反馈点集的影响最大化算法(MVCFVS)在这些真实社交网络数据上选出种子节点，并通过不同的传播模型模拟他们的实际影响传播效果，然后和其他几种节点选择方法的影响传播效果比较，评价各自的优缺点，最后分析有这样的表现的原因，总结本算法的使用条件。

本节主要分为两个部分：第1部分简要介绍实验中用到的数据集和用于作对比的其他节点选择算法；第2部分从不同的方面来分析实验结果，得出结论。

4.1 数据集和算法介绍

为了显示算法的实验效果，实验中用到的来自真实社交网络中的数据集有如下3个：

CA-HepTh[26] 该数据来自于arXiv，涵盖了提交到该网站的高能物理(High Energy Physics-Theory)分类下的作者之间的科学合作。如果作者i和作者j合作了一篇文章，那么节点i和节点j之间存在一条无向边；如果一篇文章是由k个作者共同合作完成的，那么这k个节点之间构成了一个k个节点的完全图。该数据涵盖了从1993年1月至2003年4月期间(124个月)的论文，共包含9877个节点，25998条边。

Email-Enron[27] 该数据是美国联邦能源监管委员会在调查安然公司破产案的过程中发布到网上的安然公司的邮件通信网络。其中，节点代表电子邮件的地址，边代表邮件地址之间的通信，如果两个地址之间至少发过一封邮件，那么这两个地址之间存在一条边。该网络是一个无向简单网络，覆盖约五十万封电子邮件数据集内的所有电子邮件通信，共包含36692个节点，183831条边。

Facebook-Combined(FC)[28] 该数据来自Facebook的“社交圈”(或“朋友列表”)，是一个ego网络(所谓ego网络，指的是网络的节点是由唯一的一个中心节点(ego)，以及这个节点的邻居(alters)组成的，它的边只包括了ego和alter之间，以及alter与alter之间的边)。共包含4039个节点，88234条边。

为了和其他节点选择算法作对比，选择了如下几个节点选择方法：

随机选择(Random) 一种简单的节点选择方法，这种方法完全随机地选出k个初始节点。

按照度选择(Degree) 这种方法选择网络中度最大的k个节点作为初始活跃节点集合。

局部中心度算法(Local Centrality, LC) 节点v的邻居节点集为out1(v), out1(v)中所有节点的邻居节点集的总集合为out2(v), out2(v)中所有节点的邻居节点集的总集合为out3(v)；设$\deg 1(v)$，$\deg 2(v)$，$\deg 3(v)$分别为v的相应层次邻居节点集out1(v), out2(v), out3(v)的影响度。对于无符号网络，影响度定义为邻居节点集的元素个数。节点v的潜在影响力值PI定义为：$\mathrm{PI}(v)=\deg(v)+\omega_0(1-\mathrm{e}^{-\inf(v)})$，其中$\inf(v)$定义为$v$对所有未激活邻居节点的影响力之和：

$$\inf(v)=\sum_{u\in \mathrm{out1}(v),\mathrm{active}(u)=0} b_{vu}$$

式中$\deg(v)$为v的局部中心度，ω_0为累积影响权重，ω_0可以视具体网络选取和修正。

混合度分解算法(MDD) 在计算k-核的过程中考虑了剩余度(residual degree)和排出度(exhausted degree)，该算法中的可调参数λ在实验中被设置为0.7。

基于度的启发式算法(DegreeDiscountIC, DDIC) 该算法是在传统基于度的启发式算法基础上改进后适用于独立级联模型的算法。在Degree Discount 中，如果某个节点已经被选入种子集合中，则该节点的邻居节点(不在种子集合中的节点)的度相应地减1。而DegreeDiscountIC算法使用了不同的折扣方法，当节点v的邻居节点u已经被选入

种子节点中，那么 v 将以概率 p 被 u 影响，这样的话就没必要将 v 选入种子节点中。当 p 比较小的时候，忽略 v 的多跳邻居节点对其的间接影响，只关注直接影响。$dd_v = d_v - 2t_v - (d_v - t_v)t_v p$，其中，$d_v$ 表示 v 节点的度，t_v 初始为0，对于 v 的邻居节点中未加入种子节点集合的节点每加1，t_v 也加1。

基于最小点覆盖和反馈点集的影响最大化算法(MVCFVS)　使用第 3 节中介绍的算法 3 选择初始活跃节点集合，该算法先找到网络 G 的反馈点集 F，再在 G - F 的每个分支上找最小点覆盖集合 C，最后在候选集 $R = F \cup C$ 根据影响力函数 $t(x)$ 选出 k 个节点。

4.2 实验结果

将以上6种节点选择方法选出的节点作为初始的活跃节点，分别按照线性阈值模型、独立级联模型、加权级联模型的传播方式进行影响力传播，对最终的传播效果进行比较。由于基于度的启发式算法(DegreeDiscountIC)是适用于独立级联模型的节点选择算法，所以我们只在独立级联模型下加入了这种算法(DDIC)作比较。为了保证算法的有效性，在初始活跃节点集上进行10000次模拟，取这10000次结果的平均值作为传播模型的最终结果。其结果如下：

图1比较了在CA-HepTh网络上4.1节中的6个节点选择方法在各个传播模型上的表现。其中，横坐标表示初始活跃节点(种子节点)集合的大小，即参数 k 的大小，k 的取值范围从0到50，纵坐标表示最终的影响效果，即最终活跃节点数目。

图 1(a)是 5 个不同算法选出的种子节点在线性阈值模型下的传播效果($t = 0.5$)，可以看出，本文算法(MVCFVS)虽不如 Degree 方法和 LC 方法，但和他们接近，且比 Rand 方法和 MDD 方法好得多。

图 1(b)是 6 个不同算法(包括 DDIC 算法)选出的种子节点在独立级联模型下的传播效果($p = 0.01$)，可以看出，本算法(MVCFVS)是表现最好的。

图 1(c)是 5 个不同算法选出的种子节点在加权级联模型下的传播效果($p = 1/d_w$，d_w 是 w 节点的度)，本算法(MVCFVS)和按照度选择方法(Degree)结果相似。

总的来说，MVCFVS 算法在 CA-HepTh 网络上表现不错，特别是对于独立级联模型和加权级联模型，它能够通过较小的种子节点集合去影响更多的节点，和按照度选择方法(Degree)效果相似是因为本算法在进行节点选择时用到了最大度的思想，且由于 CA-HepTh 网络本身的局部中心性，导致本算法和 Degree 方法选出的种子节点有较大的重合。

图2比较了在Email-Enron网络上6个节点选择方法在各个传播模型上的表现。图2(a)是5个不同算法选出的种子节点在线性阈值模型下的传播效果($t = 0.5$)，可以看出，本文算法(MVCFVS)在 $k < 12$ 时不如Degree方法和MDD方法，但是在 $k \geq 12$ 时和他们接近，且比Rand方法和MDD方法好得多。通过仔细比较这些方法选出的不同节点，发现造成这种结果是因为Degree方法和MDD方法一开始就将度最大的节点选入了种子节点集合，使得他们能够在早期快速影响较多的节点，而本算法是在稍晚些时候才将这些度居榜首的节点选入种子节点集合；后期随着Degree方法和MDD方法的缺陷逐渐显现，本算法的表现越来越好，开始追上Degree方法，赶超MDD方法。

图 2(b)是 6 个不同算法选出的种子节点在独立级联模型下的传播效果($p = 0.01$)，可以看出，本算法(MVCFVS)是最先收敛的，即在 k 较小时比起 Degree 方法和 MDD 方法本算法能影响更多的节点。

图 2(c)是 5 个不同算法选出的种子节点在加权级联模型下的传播效果($p = 1/d_w$，d_w 是 w 节点的度)，可以看出，本文算法(MVCFVS)也是最先收敛的，即在 k 较小时比起 Degree 方法和 MDD 方法本文算法能影响更多的节点。

总的来说，MVCFVS 算法在 Email-Enron 网络上 3 种模型下的表现比其他算法都好，特别是对于独立级联模型和加权级联模型，它收敛的速度最快，和 Degree 算法和 MDD 算法比较起来，本算法能够通过较小的种子节点集合去影响更多的节点。

图3比较了在Facebook-Combined网络上6个节点选择方法在各个传播模型上的表现。

图 3(a)是 5 个不同算法选出的种子节点在线性阈值模型下的传播效果($\theta = 0.5$)，可以看出，本文算法(MVCFVS), Degree 方法和 MDD 方法接近。

图 3(b)是 6 个不同算法选出的种子节点在独立级联模型下的传播效果($p = 0.01$)，可以看出，本文算法(MVCFVS)和 Degree 方法、MDD 方法接近。

图 3(c)是 5 个不同算法选出的种子节点在加权级联模型下的传播效果($p = 1/d_w$，d_w 是 w 节点的度)，可以看出，本文算法(MVCFVS)和 Degree 方

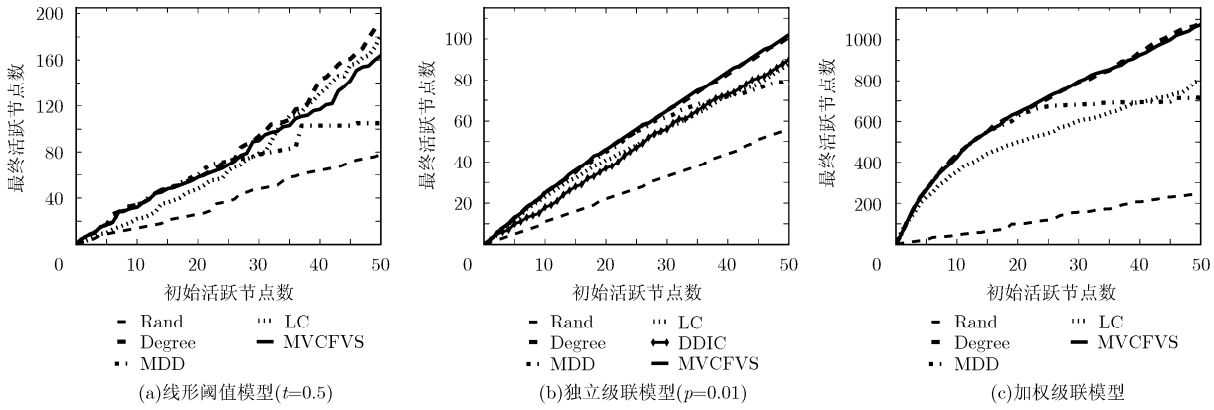

图1 CA-HepTh网络上各个算法在3个模型下的实验对比

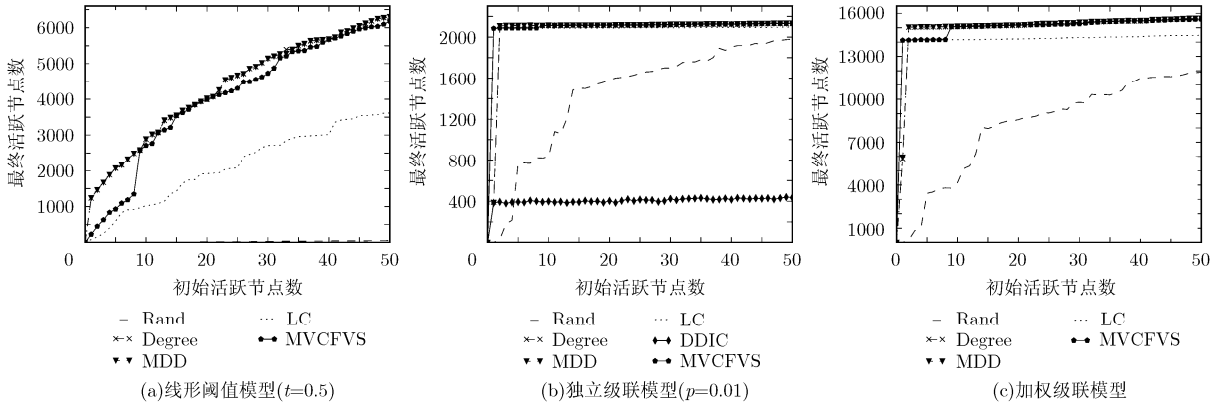

图2 Email-Enron网络上各个算法在3个模型下的实验对比

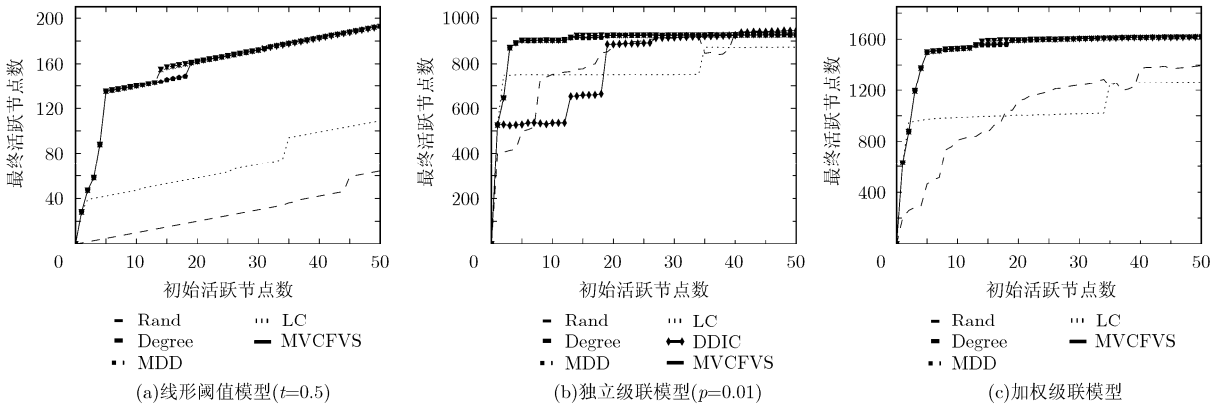

图3 Facebook-Combined网络上各个算法在3个模型下的实验对比

法、MDD 方法接近。

总的来说，在这 3 种模型下，MVCFVS 算法在 Facebook-Combined 网络上表现和 Degree 算法、MDD 算法相似，但比其他文献的算法好。这是因为 Facebook-Combined 网络是一个 ego 网络，大多数节点会团结在某个节点周围，使得本文算法的影响力函数和 Degree 算法、MDD 算法效果相似，选出的节点集合也相似，最终的传播效果也相似。

5 结束语

本文提出了一种求解社交网络影响最大化的算法MVCFVS。通过和不同的节点选择算法在不同的数据集上实验比较发现，算法MVCFVS比较适用于独立级联模型和加权级联模型，并且收敛速度很快，在 k 取较小值时，就能影响非常多的节点。在规模较大且不是那么集中的网络中采用本算法选取种子

节点，效果会更好。当然，阈值越大，最终活跃节点数越少；影响概率越大，最终活跃节点数越多。并且概率对于传播模型的影响非常显著。

另一方面，FC 算法在求解反馈点集时只考虑了顶点度的影响，并没有考虑图的整体结构，故得到的反馈点集可能会有一定的局限性，即反馈点集中包含的顶点可能会多一些，从而影响了算法 FC 选出的 k 个顶点的传播效果。为此，我们需要对 FC 算法在求解反馈点集这一步进行更深入的研究，这将是我们后续探索的工作。

参 考 文 献

[1] WATTS D J and STROGATZ S H. Collective dynamics of 'small-world' networks[J]. *Nature*, 1998, 393(6684): 440–442.

[2] BARABASI A L and ALBERT R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509–512.

[3] SAITO K, NAKANA R, and KIMURA M. Prediction of information diffusion probabilities for independent cascade model[J]. *Lecture Notes in Computer Science*, 2008, 5179: 67–75.

[4] TANG J, SUN J, and YANG Z. Social influence analysis in large-scale networks[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2009: 807–816.

[5] GOYAL A, BONCHI F, and LASKHMANAN L. Learning influence probabilities in social networks[C]. Proceedings of the Third ACM International Conference on Web Search & Data Mining, New York, USA, 2010: 241–250.

[6] WANG C, TANG J, SUN J, *et al*. Dynamic social influence analysis through time-dependent factor graphs[C]. Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining, Washington, DC, USA, 2011: 239–246.

[7] DOMIGOS P and RICHARDSON M. Mining the network value of customers[C]. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA, 2001: 57–66.

[8] RICHARDSON M and DOMINGOS P. Mining knowledge-sharing sites for viral marketing[C]. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, 2002: 61–70.

[9] KEMPE D, KLEINBERG J, and TARDOS E. Influential nodes in a diffusion model for social networks[J]. *International Colloquium on Automata, Languages and Programming*, 2005, 32: 1127–1138.

[10] KEMPE D, KLEINBERG J, and TARDOS E. Maximizing the spread of influence in a social network[C]. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, USA, 2003: 137–146.

[11] LESKOVEC J, KRAUSE A, GUESTRIN C, *et al*. Cost-effective outbreak detection in networks[C]. Proceedings of the Kdd 07 ACM SIGKDD International Conference on Knowledge Discovery and Data, Pittsburgh, PA, USA, 2007: 420–429.

[12] ZHOU C and GUO L. A note on influence maximization in social networks from local to global and beyond[C]. Proceedings of the 1th International Conference on Data Science (ICDS), Beijing, China, 2014: 27–28.

[13] ZHOU C, ZHANG P, GUO J, *et al*. An upper bound based greedy algorithm for mining top-k influential nodes in social networks [C]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data, Seoul, Korea, 2014: 421–422.

[14] CHENG S, SHEN H, HUANG J, *et al*. IMRank: influence maximization via finding self-consistent ranking[C]. Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14), New York, NY, USA, 2014: 475–484.

[15] CHEN W, WANG Y, and YANG S. Efficient influence maximization in social networks[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009: 199–208.

[16] JIANG Q, SONG G, CONG G, *et al*. Simulated annealing based influence maximization in social networks[C]. Proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2011: 127–132.

[17] ZOU F, ZHANG Z, and WU W. Latency-bounded minimum influential node selection in social networks[C]. Proceedings of the Wireless Algorithms, Systems, and Applications, 4th International Conference, Boston, MA, USA, 2009: 519–526.

[18] ZOU F, WILLSON J, ZHANG Z, *et al*. Fast information propagation in social networks[J]. *Discrete Mathematics Algorithms & Applications*, 2010, 2(1): 125–141.

[19] COHEN E, DELLING D, PAJOR T, *et al*. Sketch-based influence maximization and computation: scaling up with guarantees[C]. Proceedings of Conference on Information and Knowledge Management, CIKM, Shanghai, China, 2014: 629–638.

[20] JIANG F, JIN S, WU Y, *et al*. A uniform framework for community detection via influence maximization in social

networks[C]. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Beijing, China, 2014: 27–32.

[21] CAO J, DAN D, XU S, *et al*. A k-core based algorithm for influence maximization in social networks[J]. *Chinese Journal of Computers*, 2015, 38(2): 238–248.

[22] LUCIER B, OREN J, and SINGER Y. Singer influence at scale: distributed computation of complex contagion in networks[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, 2015: 735–744.

[23] GOLDENBERG J, LIBAI B, and MULLER E. Talk of the network: a complex systems look at the underlying process of word-of-mouth [J]. *Marketing Letters,* 2001, 12(3): 211–223.

[24] KARP R M. Reducibility among Combinatorial Problems, Complexity of Computer Computations[M]. New York, USA, Plenum Press, 1972: 85–103.

[25] SCHULZ A. Correctness-proof of a greedy-algorithm for minimum vertex cover of a tree[OL]. http.//cs.stakexchange.com, 2013.

[26] LESKOVEC J, KLEINBERG J, and FALOUTSOS C. Graph evolution: densification and shrink diameters[J]. *ACM Transactions on Knowledge Discovery from Data*, 2007, 1(1): 1–41.

[27] LESKOVEC J, LANG K, DASGUPTA A, *et al*. Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters[J]. *Internet Mathematics*, 2009, 6(1): 29–123.

[28] MCAULEY J and LESKOVEC J. Learning to discover social circles in ego networks[C]. Proceedings of the 26th Annal Conference on Information Processing Systems, Lake Tahoe, NeVada, USA, 2012: 539–547.

许宇光： 男，1984 年生，博士生，研究方向为计算机软件与理论.

潘惊治： 女，1992 年生，硕士生，研究方向为社交网络.

谢惠扬： 女，1963 年生，教授，研究方向为应用数学.