面向不确定性影响源的社会网络影响力传播抑制方法

李 劲[①②] 岳 昆[*③] 尤 洁[①②] 谢潇睿[①②] 张云飞[③]
[①](云南大学软件学院 昆明 650500)
[②](云南省软件工程重点实验室 昆明 650500)
[③](云南大学信息学院 昆明 650500)

摘 要：社会网络中影响力传播的有效抑制是社会网络影响力传播机制研究所关注的问题之一。该文针对未知影响传播源，或传播源信息具有不确定性的情况，提出面向不确定性影响源的影响力传播抑制问题。首先，为有效提高抑制算法的执行效率，讨论竞争线性阈值传播模型下影响源传播能力的近似估计方法，进而提出有限影响源情况下，期望抑制效果最大化的抑制种子集挖掘算法。其次，对于大尺寸不确定性影响源的情况，考虑算法运行效率和抑制效果之间的有效折中，提出基于抽样平均近似的期望抑制效果最大化的抑制种子集挖掘算法。最后，在真实的社会网络数据集上，通过实验测试验证了所提出方法的有效性。

关键词：社会网络；不确定性影响源；影响力传播抑制；竞争线性阈值模型；抽样平均近似
中图分类号：TP393; TP311 **文献标识码**：A **文章编号**：1009-5896(2017)09-2063-08
DOI: 10.11999/JEIT161360

Uncertain Influence Sources Oriented Influence Blocking Maximization in Social Networks

LI Jin[①②] YUE Kun[③] YOU Jie[①②] XIE Xiaorui[①②] ZHANG Yunfei[③]
[①](*School of Software*, *Yunnan University*, *Kunming* 650500, *China*)
[②](*Key Laboratory of Software Engineering*, *Yunnan Province*, *Kunming* 650500, *China*)
[③](*School of Information Science and Engineering*, *Yunnan University*, *Kunming* 650500, *China*)

Abstract: Influence blocking maximization is currently a focused issue in the research area of social networks. This paper considers the issue of influence blocking maximization with uncertain negative influence sources. First, in order to increase efficiency of blocking seeds mining algorithms, the approximate estimation method of influence propagation of negative seeds under the competitive linear threshold model is discussed. Based on the estimation, a blocking seeds mining algorithm for finite uncertain negatively influence sources is proposed to maximize expected influence blocking utility. Second, for the case of huge amount of negatively influence sources with uncertainty, a blocking seeds mining algorithm based on the sampling average approximation approach is proposed to balance the tradeoffs between scalability and effectiveness of the influence blocking maximization. Finally, experiments are carried on real data sets of social networks to verify the feasibility and scalability of the proposed algorithms.

Key words: Social networks; Uncertain influence sources; Influence blocking maximization; Competitive linear threshold model; Sampling average approximation

1 引言

近年来，许多学者针对面向多信息源发布的信息全局传播机制及其预测模型、以影响力传播范围及传播速度最大化为目标的关键结点集选取等问题开展了积极探索，并取得了一系列成果，极大地促进了社会网络影响力传播机制及相关问题的研究[1–8]。然而，由于社会网络本身的开放性和虚拟性，各种不良、虚假信息、反动言论可以跨地域、跨国界地散布和传播，严重危害社会稳定及国家安全。因此，为有效抑制万维网环境下社会网络负面影响传播，以影响力传播抑制最大化为目标的关键结点集选取、给定影响力传播范围条件下的最小化抑制结点集选取等问题引起关注，取得了一些初步的研究成果[9–15]。

例如，文献[9]扩展经典的线性阈值模型(Linear

收稿日期：2016-12-13；改回日期：2017-04-11；网络出版：2017-05-26
*通信作者：岳昆 kyue@ynu.edu.cn
基金项目：国家自然科学基金(61562091, 61472345)，云南省应用基础研究计划，(2014FA023, 2016FB110)，云南大学中青年骨干教师培养计划项目，云南大学青年英才培育计划(XT412003)，云南省软件工程重点实验室开放项目(2012SE303, 2012SE205)
Foundation Items: The National Natural Science Foundation of China (61562091, 61472345), The Natural Science Foundation of Yunnan Province (2014FA023, 2016FB110), The Foundation of Backbone Teacher Development of Yunnan University, The Program for Excellent Young Talents of Yunnan University (XT412003), The Open Foundation of Key Laboratory of Software Engineering of Yunnan Province (2012SE303, 2012SE205)

Threshold model, LT)，给出了用于描述社会网络中正、负影响竞争传播现象的竞争线性阈值模型(Competitive Linear Threshold model, CLT)，提出影响力传播的新问题：影响力传播抑制最大化问题(Influence Blocking Maximization, IBM)，证明了IBM的目标函数满足子模性(submodularity)，给出了解决IBM问题的高效近似算法。文献[10]从另一角度研究了影响力传播抑制问题，即限定影响力传播范围，求解最小的抑制种子集。文献[11]基于影响力传播的独立级联模型(Independent Cascade model, LC)建立了竞争级联模型(Multi-Campaign Independent Cascade Model, MCICM)，提出影响传播抑制问题(Eventual Influence Limitation, EIL)，给出MCICM模型下EIL问题目标函数满足子模性的充分条件，以及EIL问题的近似求解算法。

此外，由于影响力传播与抑制行为的本质是传播方与抑制方之间的一种博弈关系，文献[12]以传播方和抑制方的种子集作为策略空间，将传播与抑制决策建模为零和博弈，为避免对指数级的策略空间进行枚举，提出了基于策略生成技术的博弈Nash均衡求解算法。最近，文献[13,14]针对社会网络中影响力传播抑制问题，扩展了经典的线性阈值模型，提出了一种新型的影响力传播、抑制模型，并给出了此模型上的影响传播的最优抑制方法。文献[15]研究了针对策略性传播源的影响力传播抑制问题。该文将影响力传播抑制建模为一个极小极大(minmax)优化问题，其主要目标是最小化最坏情况下的负影响力传播范围。与该文不同，本文主要目标是针对不确定性(或未知)传播源的情况下最小化期望负影响力传播范围。

上述研究工作极大地推进了社会网络中影响力传播抑制问题的研究，但是，仍存在挑战，具体而言：已有工作假设在挖掘抑制种子集时，传播种子集是已知的。然而，在实际社会网络应用环境中，在作出抑制决策时，抑制方很难掌握传播源的准确信息，从而无法对影响力传播进行有效抑制。面对不确定性影响源，如何有效进行影响力传播抑制，这是我们面临的新的研究挑战。

针对这一研究挑战，本文提出面向不确定性影响源的影响力传播抑制问题(influence blocking maximization with uncertain influence sources)，研究期望抑制效果最大化的抑制种子集挖掘算法。具体而言，首先，用影响源上的概率分布来描述抑制方对于影响源信息的不确定性，例如，当抑制方未知影响源时，用可能的影响源上的均匀分布描述影响源的不确定性；或者抑制方根据经验知识，确定可能的影响源，及影响源上的概率分布。其次，对于有限影响源，采用基于局部有向无环图(Local Directed Acyclic Graph, LDAG)结构的估计方法对影响源传播范围进行高效的近似估计，基于此，提出了针对不确定传性播源的抑制种子集挖掘算法。然后，对于影响源数目是结点集数目的组合数的情况，为达到算法的运行效率和抑制质量之间的有效折中，提出了基于抽样平均近似(sampling average approximation)[16]的期望抑制最大化的抑制种子集挖掘算法。最后，在真实的社会网络数据集上通过实验验证了本文方法的可行性和有效性。

2 问题定义

在社会网络影响力传播抑制问题中，存在决策目标对立的双方：传播方，目的是通过选择相应的负种子集 $D \subseteq V$，使得负影响力传播的范围最大化；抑制方，通过选取正种子集 $C \subseteq V$，最大限度抑制负影响力传播的范围。负、正种子集的集合分别记为 $[D]$ 和 $[C]$，令 $\boldsymbol{\theta}^+$ 和 $\boldsymbol{\theta}^-$ 分别表示 V 上正、负影响阈值随机向量。正、负影响从种子集 D 和 C 中按照CLT模型[9]开始传播。最终，被负激活的结点集记为 $V^-(D,C|\boldsymbol{\theta}^-,\boldsymbol{\theta}^+)$。由于 $\boldsymbol{\theta}^+$ 和 $\boldsymbol{\theta}^-$ 是随机向量，$|V^-(D,C|\boldsymbol{\theta}^-,\boldsymbol{\theta}^+)|$ 是一个随机变量。在给定 C 的情况下，传播方的传播效用值定义为关于 $\boldsymbol{\theta}^+$ 和 $\boldsymbol{\theta}^-$ 的期望值 $E_{\boldsymbol{\theta}^-,\boldsymbol{\theta}^+}(|V^-\left(D,C|\boldsymbol{\theta}^-,\boldsymbol{\theta}^+\right)|)$，并记作 $\mathrm{Inf}(D,C)$。另外，给定 D 的情况下，定义结点集 C 的抑制效用函数 $\sigma\left(C|D\right)=\mathrm{Inf}(D,\varnothing)-\mathrm{Inf}(D,C)$。于是，影响力传播抑制优化问题描述如下：给定 D，求解 C，使得 $C^*=\arg\max_{C\in[C]}\sigma\left(C|D\right)$。

然而，在实际社会网络环境中，负种子集 D 的准确信息不易获得。设 p 是 $[D]$ 上的一个概率分布，描述了抑制方对于影响源信息的不确定性。本文中，p 也称作传播方的一个随机策略。于是，面向不确定性影响源的影响传播抑制问题定义为：给定 p，选取至多包含 k 个结点的抑制正种子集 C，使得负影响**期望**传播效用尽可能地小，即

$$\begin{aligned}C^*&=\mathop{\arg\max}_{C\in[C],|C|\le k}\sigma\left(C|p\right)=\mathop{\arg\max}_{C\in[C],|C|\le k}E_{D\sim p}\left(\sigma\left(C|D\right)\right)\\&=\mathop{\arg\max}_{C\in[C],|C|\le k}\sum_{D\in[D]}p(D)\sigma\left(C|D\right)\end{aligned}\tag{1}$$

3 抑制种子集挖掘算法

3.1 影响传播效用的近似估计

首先，由文献[9]结论可知，$\sigma(C|D)$ 是 V 集上的子模函数。由式(1)可知，目标函数 $\sigma(C|p)$ 是 $\sigma(C|D)$ 的非负线性组合。根据文献[17]可知，子模函数的非负线性组合仍具有子模性，因此，$\sigma(C|p)$

是结点集 V 上的子模函数。因此，从理论上看，以 $\sigma(C|p)$ 作为优化的目标函数，基于贪心法进行 k 次迭代可求得保下界的近似最优抑制种子集 $\widetilde{C}^*$，即，$\sigma(\widetilde{C}^*|p)\geq(1-1/e)\sigma(C^*|p)$。

然而，在上述贪心算法的每一次迭代过程中，在选择具有最大边际抑制效用的结点时需要计算 $\mathrm{Inf}(D,C)$。由于 $\mathrm{Inf}(D,C)$ 是关于随机向量 $\boldsymbol{\theta}^+$ 和 $\boldsymbol{\theta}^-$ 的一个期望值，准确地估计 $\mathrm{Inf}(D,C)$ 需要进行大量的随机试验，计算代价较高。因此，面对含有大量结点的社会网络来说，基于贪心法求 $\widetilde{C}^*$ 实际上是不可行的。

为此，本文采用文献[5]提出的基于 LDAG 结构对影响传播效用进行近似估计方法。设定 LDAG 结构大小的阈值 $\varphi\in(0,1)$，对于任意结点 $v\in V$，构建有向无环子图 $\mathrm{LDAG}_{\varphi}^{+}(v)$，$\mathrm{LDAG}_{\varphi}^{+}(v)$ 是一个以 v 作为汇点的有向无环子图。对于任意结点 $u\in\mathrm{LDAG}_{\varphi}^{+}(v)$，$u$ 对于 v 有正影响。任意 $u\notin\mathrm{LDAG}_{\varphi}^{+}(v)$，则 u 对于结点 v 没有正影响。类似地，对于任意 $v\in V$，构建 $\mathrm{LDAG}_{\varphi}^{-}(v)$。基于 $\mathrm{LDAG}_{\varphi}^{-}(v)$ 及 $\mathrm{LDAG}_{\varphi}^{+}(v)$，给定负、正种子集 D，C，结点 v 在 CLT 传播模型下的负、正激活概率 $\mathrm{ap}^-(v|D,C)$ 和 $\mathrm{ap}^+(v|D,C)$ 定义如下[9]：

$$\left.\begin{aligned}
P^+(v,t)&=\sum_{u\in\mathrm{LDAG}^+(v)}w_{u,v}^{+}\mathrm{ap}^+(u,t-1)\\
P^-(v,t)&=\sum_{u\in\mathrm{LDAG}^-(v)}w_{u,v}^{-}\mathrm{ap}^-(u,t-1)\\
\mathrm{ap}^+(v,t)&=P^+(v,t)\left(1-\sum_{k=0}^{t}P^-(v,k)\right)\\
\mathrm{ap}^-(v,t)&=P^-(v,t)\left(1-\sum_{k=0}^{t-1}P^+(v,k)\right)
\end{aligned}\right\}\tag{2}$$

由文献[9]可知，基于 $\mathrm{ap}^-(v|D,C)$ 可以对 $\mathrm{Inf}(D,C)$ 进行有效估计，即：$\mathrm{Inf}(D,C)\approx\sum_{v\in V\setminus D\cup C}\mathrm{ap}^-\left(v|D,C\right)$。

需要说明的是，$\mathrm{Inf}(D,C)=E_{\boldsymbol{\theta}^-,\boldsymbol{\theta}^+}(|V^-(D,C|\boldsymbol{\theta}^-,\boldsymbol{\theta}^+)|)$ 和 $\mathrm{Inf}(D,C)\approx\sum_{v\in V\setminus D\cup C}\mathrm{ap}^-\left(v|D,C\right)$ 均是负种子集的影响传播范围的度量标准，但两个公式的值含义不同。前者是负激活结点集大小的期望值，而后者的值是负激活结点的负激活概率值之和。为区分两者不同的含义，将 $\mathrm{Inf}(D,C)\approx\sum_{v\in V\setminus D\cup C}\mathrm{ap}^-\left(v|D,C\right)$ 称作负种子集 D 在抑制种子集 C 下的负影响度。文献[9]结论表明，基于 $\mathrm{ap}^-\left(v|D,C\right)$ 对 $\mathrm{Inf}(D,C)$ 进行估计可实现算法效率和估计质量之间的有效折中。

3.2 面向有限影响源的抑制种子集的挖掘算法

设 $[D]_N=\{D_1,D_2,\cdots,D_N\}$（$\forall D_i\subseteq V$）是包含 N 个负种子集(即影响源)的集合。p_N 是 $[D]_N$ 上的一个概率分布。由 3.1 节可知，随机策略 p_N 的期望负影响度为

$$\begin{aligned}
\mathrm{Inf}\left(p_N,C\right)&=\sum_{D\in[D]_N}p_N(D)\,\mathrm{Inf}(D,C)\\
&\approx\sum_{D\in[D]_N}p_N(D)\sum_{v\in V\setminus D\cup C}\mathrm{ap}^-\left(v|D,C\right)\\
&=\sum_{v\in V\setminus C}\sum_{D\in[D]_N}p_N(D)\,\mathrm{ap}^-\left(v|D,C\right)
\end{aligned}\tag{3}$$

为描述方便，引入记号 $\mathrm{eap}^-\left(v|p_N,C\right)$，即在 p_N 和 C 下，任意结点 v 的负激活概率的期望值：

$$\mathrm{eap}^-\left(v|p_N,C\right)=\sum_{D\in[D]_N}p_N(D)\left(\mathrm{ap}^-\left(v|D,C\right)\right)\tag{4}$$

于是，式(3)可重写为

$$\mathrm{Inf}\left(p_N,C\right)\approx\sum_{v\in V\setminus C}\mathrm{eap}^-\left(v|p_N,C\right)$$

也就是说，$\mathrm{Inf}\left(p_N,C\right)$ 可以用结点的负激活概率的期望值进行估计，由 3.1 节可知，$\mathrm{eap}^-(v|p_N,C)$ 的计算无需大量的随机仿真实验，由式(2)就可以高效求解。在给定 p_N 的情况下，表 1 的算法 1 给出了面向有限影响源的抑制种子集挖掘算法。

表 1　面向有限影响源的抑制种子集挖掘算法

算法 1　面向有限影响源 $[D]_N$ 的抑制种子集挖掘算法

输入： 图 $G=\langle V,E\rangle$，$[D]_N=\{D_1,D_2,\cdots,D_N\}$，$[D]_N$ 上的概率分布 p_N，LDAG 结构控制参数 φ，面向负影响抑制的正种子集大小 k。

输出： $[D]_N$ 下的抑制种子集 C_N^*。

步骤 1　$C\leftarrow\varnothing$；

步骤 2　对图中所有结点 $v\in V$，基于局部结构控制参数 φ 构建 $\mathrm{LDAG}_{\varphi}^{+}(v)$，$\mathrm{LDAG}_{\varphi}^{-}(v)$，$\mathrm{OutLS}_{\varphi}^{+}(v)$；

步骤 3　初始化每个结点 v 的抑制效用，记为 $\mathrm{DecInf}^-(v)\leftarrow 0$

步骤 4　**FOR**　$v\in V$ and　$u\in\mathrm{LDAG}^+(v)$　**DO**

$\mathrm{before}\leftarrow\mathrm{eap}^-\left(v|p_N,\varnothing\right)$；

$\mathrm{after}\leftarrow\mathrm{eap}^-\left(v|p_N,\{u\}\right)$；

$\mathrm{DecInf}^-(u)\leftarrow\mathrm{DecInf}^-(u)+(\mathrm{before}-\mathrm{after})$；

步骤 5　**END FOR**

步骤 6　**FOR**　$i\leftarrow 1$ to　k　**DO**

$c\leftarrow\arg\max_{v\in V\setminus C}\left(\mathrm{DecInf}^-(u)\right)$；//选择抑制效用最大的正种子结点；

步骤 7　**FOR**　$s\in\mathrm{OutLS}^+(c)$　**DO**

FOR　$u\in\mathrm{LDAG}^+(s)$　**DO**　//更新 $\mathrm{LDAG}^+(s)$ 中结点的抑制效用；

步骤 8　$\mathrm{DecInf}^-(u)\leftarrow\mathrm{DecInf}^-(u)$

$-\left(\mathrm{eap}^-\left(s|p_N,C\cup\{c,u\}\right)-\mathrm{eap}^-\left(s|p_N,C\right)\right)$；

$\mathrm{DecInf}^-(u)\leftarrow\mathrm{DecInf}^-(u)$

$+\left(\mathrm{eap}^-\left(s|p_N,C\cup\{c,u\}\right)-\mathrm{eap}^-\left(s|p_N,C\cup\{c\}\right)\right)$

步骤 9　**END FOR**

步骤 10　**END FOR**

步骤 11　$C\leftarrow c$；//添加新的抑制正种子

步骤 12　**END FOR**

步骤 13　**RETURN**　C。

算法 1 的基本思想类似于贪心法挖掘抑制正种子集的过程，两者的最大不同在于在挖掘正种子集的过程中，每次迭代时，对结点的抑制效用的估计方法不同，具体而言，基于贪心法进行挖掘，如果要获得对结点的抑制效用的准确估计，需要大量随机仿真实验，算法代价高。然而，在算法 1 中，结点的抑制效用是基于 $\mathrm{eap}^-\left(v|p_N,C\right)$ 的计算，该计算无需随机仿真，算法执行效率高，同时，能保证具有较好的抑制效用估计质量，不会对最终的挖掘结果造成较大影响。

算法 1 的执行包括两个阶段：(1)初始阶段，即算法 1 中的步骤 1–步骤 5。其中，步骤 2 对每个结点 v 的 3 种有向无环子图进行初始化，子图 $\mathrm{LDAG}_{\varphi}^{+}(v)$ 是能够对结点 v 产生正影响的结点集，$\mathrm{LDAG}_{\varphi}^{-}(v)$ 是能够对结点 v 产生负影响的结点集，$\mathrm{OutLS}_{\varphi}^{+}(v)$ 是结点 v 能够对其有正影响的结点集。步骤 4, 步骤 5 初始化每一个结点单独作为正种子时的抑制效用。$\mathrm{DecInf}^-(v)$ 存储了结点 v 对于当前正种子集 C 的抑制效用的期望值的边际贡献值。(2)种子集挖掘阶段，步骤 6–步骤 12 是正种子集的挖掘过程，每次迭代，选取具有最大边际贡献值 $\mathrm{DecInf}^-(v)$ 的结点 v 加入到正种子集 C 中，当 v 加入 C 后，通过步骤 8 更新相关的结点的 $\mathrm{DecInf}^-(v)$ 值。由于算法 1 与贪心法本质上是一致的，因此，算法的正确性可以得到保证。

3.3 基于抽样平均近似的抑制种子集的挖掘算法

设 p 是 $[D]$ 上的概率分布。当 $[D]$ 中负种子集数目很大的情况下，直接基于算法 1 求解最优抑制种子集 C^* 是不现实的。由抑制效用的目标函数可知，给定 p 的情况下，求具有最优期望抑制效果的抑制种子集问题是一个随机优化问题[18](Stochastic Optimization Problem, SOP)，而抽样平均近似方法[16](Sampling Average Approximation, SAA)是求解 SOP 问题的有效近似方法。

一个 SOP 问题描述为 $\max\limits_{x\in S}\{F(x):=E_{\theta\sim p}f(x,\theta)\}$，其中，参数 θ 是服从概率分布 p 的随机变量，S 是可行解集，$f(x,\theta)$ 是以 x 为自变量、由 θ 决定的实函数。$F(x)=\sum_{\theta}p(\theta)f(x,\theta)$ 是 $f(x,\theta)$ 关于概率分布 p 的期望值函数。SAA 是基于蒙特卡罗模拟的求解随机优化问题的有力方法。令 $\theta_1,\theta_2,\cdots,\theta_N$ 是服从分布 p 的 N 个独立同分布随机样本，SOP 问题的期望目标函数通过以下 N-抽样平均函数(N-sample average function)来近似 $\widehat{F}_N(x)=(1/N)\sum_{i=1}^{N}f(x,\theta_i)$。文献[16]证明在足够取样的情况下，通过求解平均目标函数可得到原问题的近似解。

基于 SAA 方法的算法需解决两个问题：如何实现求解效率和求解质量之间的折中，算法的收敛标准。

首先，讨论折中问题。式(1)定义了一个参数空间 $[D]$ 服从概率分布 p 的 SOP 问题。算法 1 给出了该问题的 N-抽样平均函数的求解方法。一般地，采用较大 N 值求解算法 1 可得到原问题高质量的近似解。随着 N 值的增大，算法 1 的求解效率也随之降低。反之，对于较小的 N 值，算法可以相对快速求解。为有效地在求解质量和求解效率之间进行折中，本文的方法是：针对算法 1 的求解效率，设置相对小的 N 值，进行 M 次抽样。第 $m=1,2,\cdots,M$ 次抽样时，按照分布 p 抽取 N 个负种子集，得到负种子集集合 $[D]_N^m=\left\{D_1^m,D_2^m,\cdots,D_N^m\right\}$。由 $[D]_N^m$ 确定的抽样平均函数记为

$$\hat{\sigma}_N^m(C)=\frac{1}{N}\sum_{i=1}^{N}\sigma\left(D_i^m,C\right) \tag{5}$$

基于算法 1 求解式(5)，得到抑制种子集 $\widehat{C}_N^m=\mathop{\arg\max}\limits_{C\in[C],|C|\le k}\hat{\sigma}_N^m(C)$ 以及抑制效用 $\hat{\sigma}_N^m$。抽样 M 次，可得 $\hat{\sigma}_N^1,\hat{\sigma}_N^2,\cdots,\hat{\sigma}_N^M$，以及 $\widehat{C}_N^1,\widehat{C}_N^2,\cdots,\widehat{C}_N^M$。将 M 次抽样得到的抑制效果求算术平均 $\overline{\sigma}_N^M=(1/M)\sum_{m=1}^{M}\hat{\sigma}_N^m$。由文献[16]可知，$\overline{\sigma}_N^M$ 给出了原问题最优解的一个统计下界。进而，再得到 $\overline{\sigma}_N^M$ 的方差估计，即 $S_M^2/M=\left(1/M(M-1)\right)\sum_{m=1}^{M}\left(\hat{\sigma}_N^m-\overline{\sigma}_N^M\right)^2$。

其次，讨论算法的收敛标准问题。按照 p 抽取 N' $(N<N')$ 个负种子集，得到 $[D]_{N'}=\{D_1,D_2,\cdots,D_{N'}\}$，进而确定平均近似函数：

$$\hat{\sigma}_{N'}(C)=\frac{1}{N'}\sum_{i=1}^{N'}\sigma\left(D_i,C\right) \tag{6}$$

用 $\hat{\sigma}_{N'}(C)$ 作为评价函数，从 $\widehat{C}_N^1$，$\widehat{C}_N^2,\cdots,\widehat{C}_N^M$ 选取具有最大抑制效果的负种子集，记为 $\widehat{C}^*$。$\hat{\sigma}_{N'}(\widehat{C}^*)$ 是对原问题最优解的一个近似估计。求 $\hat{\sigma}_{N'}(\widehat{C}^*)$ 的方差，即

$$S_{N'}^2\frac{\left(\widehat{C}^*\right)}{N'}=\frac{1}{N'\left(N'-1\right)}\sum_{j=1}^{N'}\left(\sigma\left(D_j,\widehat{C}^*\right)-\hat{\sigma}_{N'}\left(\widehat{C}^*\right)\right)^2 \tag{7}$$

由文献[18]的结论可知，式(8)

$$\left(\hat{\sigma}_{N'}\left(\widehat{C}^*\right)-\overline{\sigma}_N^M\right)+z_{\alpha}\left(\frac{S_{N'}^2\left(\widehat{C}^*\right)}{N'}+\frac{S_M^2}{M}\right)^{1/2} \tag{8}$$

给出了近似解 $\widehat{C}^*$ 与原问题解之间的间隙估计(gap estimator)，也就是说，给定一个 $\varepsilon>0$，当式(8)的值小于 ε 时，得到的 $\widehat{C}^*$ 就是原问题的有效近似解。表 2 的算法 2 是基于 SAA 的抑制种子集挖掘算法的完整描述。

表 2 基于 SAA 的抑制种子集挖掘算法

算法 2　基于抽样平均近似的抑制种子集挖掘算法

输入：图 $G=\langle V,E\rangle$，概率分布 $p\sim[D]$，LDAG 结构参数 φ，抑制种子集大小 k，算法终止阈值 ε。

输出：抑制种子集 $\widehat{C}^*$。

步骤 1　确定 N,N',M 参数；

步骤 2　**For** $m=1,2,\cdots,M$ **do**

步骤 3　按照 p，从 $[D]$ 中抽取 N 个种子集 $[D]_N^m=\{D_1^m, D_2^m,\cdots,D_N^m\}$，确定 N-抽样平均函数：

$$\hat{\sigma}_N^m(C)=\frac{1}{N}\sum_{i=1}^{N}\sigma\left(D_i^m,C\right)$$

步骤 4　基于算法 1 求解 $\max\limits_{C}\hat{\sigma}_N^m(C)$，得到 $\hat{\sigma}_N^m,\widehat{C}_N^m$；

步骤 5　**END FOR**

步骤 6　由 $\hat{\sigma}_N^m,\widehat{C}_N^m\ (m=1,2,\cdots,M)$，分别确定均值 $\bar{\sigma}_N^M$，方差 S_M^2/M；

步骤 7　按照 p，从 $[D]$ 中抽取 N' 种子集 $[D]_{N'}=\{D_1,D_2,\cdots,D_{N'}\}$，确定 N'-抽样平均函数：

$$\hat{\sigma}_{N'}(C)=\left(\frac{1}{N'}\right)\sum_{i=1}^{N'}\sigma\left(D_i,C\right)$$

步骤 8　用 $\hat{\sigma}_{N'}(C)$ 作为评价函数，在 $\widehat{C}_N^m\ (m=1,2,\cdots,M)$ 当中选择最大抑制效用种子集：

$$\widehat{C}^*\leftarrow\mathop{\arg\max}_{C\in\left\{\widehat{C}_N^1,\widehat{C}_N^2,\cdots,\widehat{C}_N^M\right\}}\hat{\sigma}_{N'}(C)$$

步骤 9　根据式(6)和式(7)分别确定：$\hat{\sigma}_{N'}\left(\widehat{C}^*\right)$ 和 $S_{N'}^2\left(\widehat{C}^*\right)/N'$；

步骤 10　**IF** 式(8)间隙估计值 $<\varepsilon$ **THEN**；

步骤 11　**RETURN** $\widehat{C}^*$；

步骤 12　**ELSE** 增加 N,N' 尺寸，回到步骤 2 执行算法。

4　实验结果

4.1 实验环境设置

本文在 3 个真实的社会网络数据集 Twitter, Wiki-Vote 和 NetPHY [1)]上进行了算法测试。首先，采用广度优先搜索的方法，从 Twitter, Wiki-Vote 和 NetPHY 中分别抽取了结点数分别为 600, 6000 的连通子图。其次，为使得 CLT 传播模型在下载数据集上进行使用，对于每一个结点 v，将 v 所有入边的边权重进行归一处理，即 v 的每条入边的权重等于该边的初始权重除以 v 的所有入边的初始权重和。最后，与文献[9]中所述方法一致，设置正、负影响传播率 $p^+,p^-\in[0,1]$。对于每一条边，将边权重乘以 p^+,p^-，分别获得该边的正、负影响边权重。实验程序采用 C++语言编写，编译器版本 gcc 4.8.2。实验运行环境：Intel Xeon CPU E5620 2.40 GHz, 8 G 内存的服务器，操作系统 Linux Ubuntu 14.04 LTS。实验程序采用 C++ 11 提供的随机数引擎实现随机数生成及随机抽样功能。

4.2 算法的负影响传播抑制效果

将贪心法(记为 Greedy)、面向有限影响源的抑制算法(算法 1，记为 N-mining)、基于 SAA 的挖掘算法(算法 2，记为 SAA-mining)、随机部署抑制种子集(记为 Random)、以结点度为选取标准的抑制种子集选取算法(记为 Degree) 5 种算法的抑制效果进行了对比测试。对相关算法进行说明：(1)Greedy：基于贪心法挖掘抑制种子集。每次挖掘一个抑制种子，采用 10000 次蒙特卡罗仿真来估计影响传播效果。(2)N-mining：控制 LDAG 结构大小的 $\varphi=0.01$。(3)Random：等概率地从图中选取抑制种子集。(4)Degree：以结点度为概率权重，随机选取抑制种子集。

考虑到 Greedy 方法的运行效率，在比较不同算法的抑制效果时，只在 Twitter-600, NetPHY-600 以及 Wiki-Vote-600 3 个子图上进行对比实验。实验过程：首先，按照均匀分布，随机地从子图结点集 V 中抽取 20 个结点生成负种子集 D_i，一共生成 300 个可能的负种子集 D_i 构成 $[D]_{N=300}$，记 $p_{\text{uf}}=1/N$ 是 $[D]_N$ 上的均匀分布。其次，对抑制种子集大小分别为 30~100 的情形，执行上述 5 种抑制种子集挖掘算法。其中，SAA 算法的参数如下：在 $[D]_{N=300}$ 范围内进行随机采样，初始采样的 N 值为 5，每次迭代 N 值递增 10，迭代的抽样次数 M=10，测试样本的初始大小 $N'=10$，N' 递增值为 20。算法的收敛阈值 $\varepsilon=5$。

图 1、图 2、图 3 分别给出了 5 种算法在 NetPHY-600, Twitter-600 和 Wiki-Vote-600 数据集上的抑制效果对比图。表 3 给出了算法的运行时间对比结果。可以看到，贪心法具有最好的抑制效果，但由于挖掘算法在执行过程中需要通过随机仿真来评价结点的抑制效用，运行时间较长，不适用于大尺寸数据集。本文提出的算法 1 在抑制效果上与 Greedy 差距不大，但其运行时间显著下降；算法 2 抑制效果接近 N-mining，运行时间进一步下降。通过 5 种算法的抑制效果对比图，相比较 Random, Degree 两种简单的抑制种子集选取方法，Greedy, N-mining, SAA-mining 在抑制效果上均有显著的提升，而 N-mining, SAA-mining 在抑制效果和算法运行时间两方面得到了很好的折中。

下面，为验证 SAA-mining 在大尺寸数据集上的抑制效果和运行效率，我们在 Twitter-6000, NetPHY-6000 以及 Wiki-Vote-6000 3 个数据集上进行了测试。实验设置如下：

[1)] Twitter, Wiki-Vote 下载自：http://snap.stanford.edu/, NetPHY 下载自 http://research.microsoft.com/en-us/people/weic/projects.aspx

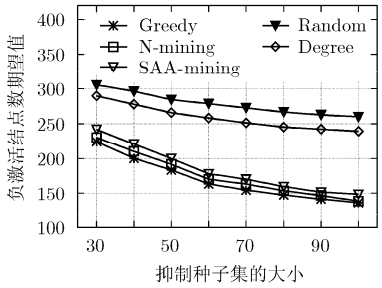

图 1 NetPHY-600 算法抑制效果比较

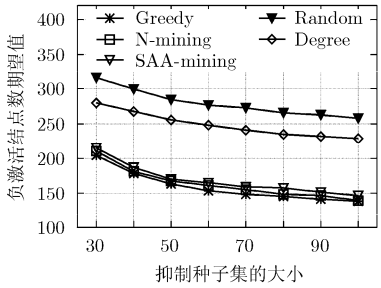

图 2 Twitter-600 算法抑制效果比较

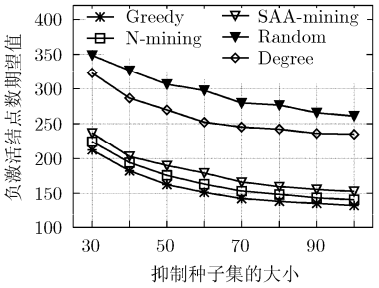

图 3 Wiki-Vote-600 算法抑制效果比较

表 3 算法执行时间比较结果(min)

k	Greedy			N-mining			SAA		
	Wiki	PHY	Twitter	Wiki	PHY	Twitter	Wiki	PHY	Twitter
40	102.3	132.2	236.2	16.4	21.8	45.4	6.3	7.4	10.6
60	126.5	168.5	260.3	17.4	22.3	48.5	7.3	8.6	12.3
80	156.4	195.1	277.4	19.2	25.4	54.8	9.4	10.8	13.5
100	186.5	246.3	298.7	23.5	28.8	60.6	11.6	13.7	14.9

(1) 抑制种子集挖掘阶段，负种子集大小 $|D_i|=100$，初始采样的 N 值为 5，每次迭代 N 值递增 10，迭代的抽样次数 M=10，测试样本的初始大小 $N'=10$，N' 递增值为 20，算法的收敛阈值 $\varepsilon=5$，考虑种子集大小分别为 50, 100, 150, 200 和 250 情形下的挖掘结果。

(2)抑制效果测试阶段，在每次完成抑制种子集挖掘后，对抑制效果进行测试，测试方法是，从 V 中随机选取大小为 100 的负种子集 200 个，每个种子集等概率作为负影响传播源，以抑制种子集 C 在所有负种子集上抑制效用的期望值作为本次测试的抑制效果，重复实验 100 次，重复实验结果的算术平均值作为抑制种子集的最终抑制效果。

图 4、图 5 和图 6 分别给出了 Random, Degree, SAA-mining 在 NetPHY-6000, Twitter-6000 以及 Wiki-Vote-6000 上抑制效果的对比结果。从以上实验结果可知，SAA-mining 算法的结果明显优于 Random 以及 Degree，能够获得很好的期望抑制效果。图 7 给出了 SAA-mining 挖掘算法在 3 个数据集上的运行时间。可以看到，与 Greedy 和 N-mining 算法相比，SAA-mining 算法运行时间显著减少。因此，针对大尺寸数据集，SAA-mining 算法在抑制效果和运行时间上仍能取得很好的折中。

5 结束语

影响力传播的建模和分析及其控制是当前社会网络研究的重要内容。其中，影响力传播的有效抑制是当前研究面临的一个新的挑战。针对现有研究工作未考虑影响源信息未知或具有不确定性的情况，本文提出了面向不确定性影响源的影响力传播抑制问题，以期望抑制效果最大化为目标，对于有限影响源及大尺寸影响源两种情况分别提出了相应的抑制种子集挖掘算法，在实际社会网络数据集上验证了算法，实验结果表明本文提出的方法能够在算法运行效率和影响力传播抑制效果之间得到很好的折衷。

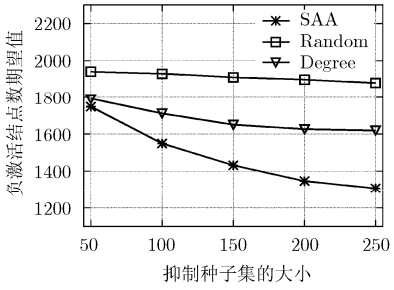

图 4 NetPHY-6000 上算法抑制效果

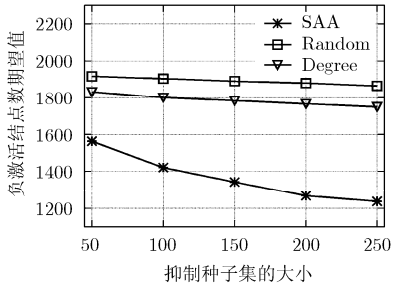

图 5 Twitter-6000 上算法抑制效果

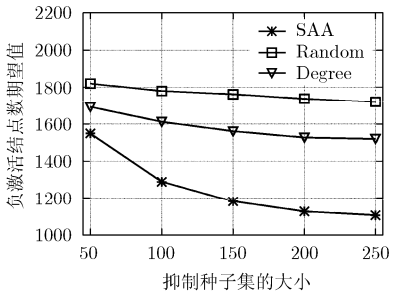

图 6　Wiki-Vote-6000 上算法的抑制效果

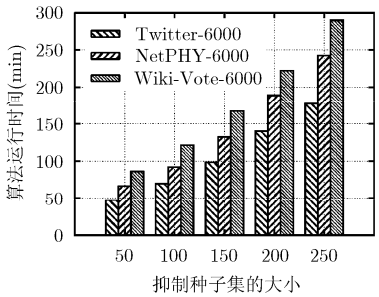

图 7　SAA-mining 算法的运行时间图

在本文研究的基础上，未来可以在以下几个方面继续开展研究。首先，本文中影响力传播模型是对线性阈值模型扩展而得，而对于其他典型的传播模型，例如独立级联模型，进行扩展，进而提出相应的影响力竞争性传播模型，并研究该模型下的具有不确定性影响源的影响传播抑制问题。其次，面对大尺寸不确定性负影响源，本文基于抽样平均近似方法提出了在运行效率和抑制效果之间进行折中的挖掘算法，进一步提高算法的效率和抑制效果，或者基于其他随机优化方法来解决不确定性的影响传播抑制问题也是将来研究的课题。

参 考 文 献

[1] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735–752. doi: 10.3724/SP.J.1016.2014.00735.
WU Xingdong, LI Yi, and LI Lei. Influence analysis of online social networks[J]. *Chinese Journal of Computers*, 2014, 37(4): 735–752. doi : 10.3724/SP.J.1016.2014.00735.

[2] 刘业政, 李玲菲, 姜元春. 社会化营销绩效最大化问题及其扩展研究综述[J]. 电子与信息学报, 2016, 38(9): 2130–2140. doi: 10.11999/JEIT160517.
LIU Yezheng, LI Lingfei, and JIANG Yuanchun. Review of social marketing performance maximization problem and its extension[J]. *Journal of Electronics & Information Technology*, 2016, 38(9): 2130–2140. doi: 10.11999/JEIT 160517.

[3] KEMPE D, KLEINBERG J, and TARDOS É. Maximizing the spread of influence through a social network[C]. Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2003: 137–146. doi: 10.1145/956750.956769.

[4] LESKOVEC J, KRAUSE A, GUESTRIN C, *et al*. Cost-effective outbreak detection in networks[C]. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2007: 420–429. doi: 10.1145/1281192.1281239.

[5] CHEN Wei, WANG Chi, and WANG Yajun. Scalable influence maximization for prevalent viral marketing in large-scale social networks[C]. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2010: 1029–1038. doi: 10.1145 /1835804.1835934.

[6] LU Wei, CHEN Wei, and LAKSHMANAN L V S. From competition to complementarity: Comparative influence diffusion and maximization[J]. *Proceedings of the VLDB Endowment*, 2015, 9(2): 60–71. doi: 10.14778/2850578. 2850581.

[7] SONG Guojie, ZHOU Xiabing, WANG Yu, *et al*. Influence maximization on large-scale mobile social network: A divide-and-conquer method[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2015, 26(5): 1379–1392. doi: 10.1109/ TPDS.2014.2320515.

[8] 许宇光, 潘惊治, 谢惠扬. 基于最小点覆盖和反馈点集的社交网络影响最大化算法[J]. 电子与信息学报, 2016, 38(4): 795–802. doi: 10.11999/JEIT160019.
XU Yuguang, PAN Jingzhi, and XIE Huiyang. Minimum vertex covering and feedback vertex set-based algorithm for influence maximization in social network[J]. *Journal of Electronics & Information Technology*, 2016, 38(4): 795–802. doi: 10.11999/JEIT160019.

[9] HE Xinran, SONG Guojie, CHEN Wei, *et al*. Influence blocking maximization in social networks under the competitive linear threshold model[C]. 9th VLDB Workshop on Secure Data Management, Istanbul, 2012: 463–474. doi: 10.1137/1.9781611972825.40.

[10] NGUYEN N P, YAN G, THAI M T, *et al*. Containment of misinformation spread in online social networks[C] Proceedings of the 3rd Annual ACM Web Science Conference, Evanston, Illinois, 2012: 213–222. doi: 10.1145/2380718. 2380746.

[11] BUDAK C, AGRAWAL D, and EL ABBADI A. Limiting the spread of misinformation in social networks[C] Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 2011: 665–674. doi: 10.1145/ 1963405.1963499.

[12] TSAI J, NGUYEN T H, WELLER N, *et al*. Game-theoretic target selection in contagion-based domains[J]. *The*

Computer Journal, 2014, 57(6): 893–905. doi: 10.1093/comjnl/bxt094.

[13] WU Hong, LIU Weiyi, YUE Kun, *et al*. Maximizing the spread of competitive influence in a social network oriented to viral marketing[C]. Proceedings of the 16th International Conference Web-Age Information Management, Qingdao, China, 2015: 516–519. doi: 10.1007/978-3-319-21042-1_53.

[14] LIU Weiyi, YUE Kun, WU Hong, *et al*. Containment of competitive influence spread in social networks[J]. *Knowledge-Based Systems*, 2016, 109: 266–275. doi: 10.1016/j.knosys.2016.07.008.

[15] 李劲, 岳昆, 张德海, 等. 社会网络中影响力传播的鲁棒抑制方法[J]. 计算机研究与发展, 2016, 53(3): 601–610. doi: 10.7544/issn1000-1239.2016.20148341.
LI Jin, YUE Kun, ZHANG Dehai, *et al*. Robust influence blocking maximization in social networks[J]. *Journal of Computer Research and Development*, 2016, 53(3): 601–610. doi: 10.7544/issn1000-1239.2016.20148341.

[16] KLEYWEGT A, SHAPRIO A, and HOMEM-DE-MELLO T. The sample average approximation method for stochastic discrete optimization[J]. *SIAM Journal on Optimization*, 2002, 12(2): 479–502. doi: 10.1137/S1052623499363220.

[17] FUJISHIGE S. Submodular Functions and Optimization[M]. Amsterdam, Elsevier Science Press, 2005, Chapter 3.

[18] SHAPRIO A, DENTCHEVA D, and RUSZCZYNSKI A. Lectures on Stochastic Programming: Modeling and Theory [M]. SIAM: Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, SIAM Press, 2014, Chapter 1.

李 劲： 男，1975年生，副教授，研究方向为数据挖掘.
岳 昆： 男，1979年生，教授，研究方向为数据挖掘.
尤 洁： 女，1992年生，硕士生，研究方向为数据挖掘.
谢潇睿： 男，1993年生，硕士生，研究方向为数据挖掘.
张云飞： 男，1992年生，硕士生，研究方向为数据挖掘.