计算机科学
COMPUTER SCIENCE

社交网络中的虚假信息经加边修正最大化问题

宋新月，帅天平，陈彬

引用本文

宋新月，帅天平，陈彬. 社交网络中的虚假信息经加边修正最大化问题[J]. 计算机科学, 2022, 49(11): 316-325.

SONG Xin-yue, SHUAI Tian-ping, CHEN Bin. Misinformation Correction Maximization Problem with Edge Addition in Social Networks[J]. Computer Science, 2022, 49(11): 316-325.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000043

社交网络中的虚假信息经加边修正最大化问题

宋新月　帅天平　陈　彬
北京邮电大学理学院　北京 100876
(xysong@bupt.edu.cn)

摘　要　在线社交网络如微信等的普及,使人们更加关注信息传播的问题。虚假信息在社交网络中进行传播可能会造成很严重的后果,比如经济损失或者公众恐慌等。因此,需要采取相关的措施来控制虚假信息的传播。传统的虚假信息控制方法主要通过向网络中的部分节点传播真实信息,让真实信息和虚假信息进行竞争来减小虚假信息的影响。文中将传播真实信息和加边的方式相结合,提出了一个虚假信息修正最大化问题。该问题是NP-难的,其目标函数值的计算是#P-难的。由于目标函数既不是次模的也不是超模的,因此采用三明治近似策略来求解该问题。为此,构造目标函数的次模的上界和下界函数,利用反向影响采样技术在基数约束下求解上界和下界函数,最终得到原问题的一个数据相关的近似解。通过在3个真实网络的数据集上进行仿真实验,验证了所提算法的有效性。

关键词:社交网络;信息传播;影响力最大化;虚假信息控制;三明治近似策略
中图法分类号　TP391;O221

Misinformation Correction Maximization Problem with Edge Addition in Social Networks

SONG Xin-yue,SHUAI Tian-ping and CHEN Bin
School of Science,Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract　The popularity of online social networks such as Wechat has aroused people's more attention to information diffusion. The spread of misinformation in social networks may lead to serious consequences,such as economic losses and public panic. Therefore,relevant measures need to be taken to control the spread of misinformation. The classical misinformation containment problem aims to reduce the impact of misinformation by launching a set of nodes as real information seeds to compete against misinformation. This paper combines the spread of real information with the edge addition,proposes a misinformation correction maximization problem. The proposed problem is NP-hard and the computation of its objective function is #P-hard. Then,we use the sandwich approximation strategy to get an approximation solution since the objective function is neither submodular nor supermodular. We first find submodular lower and upper bound functions of the objective function,and then get corresponding approximation solutions under the cardinality constraint by the reverse influence sampling technique. At last,combined with lower and upper bound function's solution,we obtain an approximation solution for misinformation correction maximization problem with a data-dependent approximation ratio. Experiments on three realistic data sets indicate that the proposed algorithm is efficient and performs better than classical heuristic methods.

Keywords　Social network,Information diffusion,Influence maximization,Misinformation containment,Sandwich approximation strategy

1　引言

随着互联网的快速发展,在线社交网络越来越受到人们的欢迎,并且已经成为了主要的社交工具。人们在在线社交网络平台上进行交流,传播信息。影响最大化问题作为社交网络信息传播研究中的一个关键算法问题,由于具有潜在的商业价值,近年来得到了广泛的研究[1-9]。影响力最大化问题的目标是在给定的传播模型下,从社交网络中选择k个最具影响力的用户作为种子节点,使得影响传播最大化。

在线社交网络在为信息交流提供有效途径的同时,也会造成虚假信息的传播。虚假信息的传播速度非常快,并且可能造成很严重的负面影响[10]。例如,在新冠疫情期间,经常有人发布与疫情相关的不实信息,这些信息在网络上快速传播,造成了群众的恐慌。为了尽可能减小虚假信息传播带来

到稿日期:2021-10-08　返修日期:2022-04-22
基金项目:国家自然科学基金(12171051,12171052);中央高校基本科研业务费(500421358)
This work was supported by the National Natural Science Foundation of China(12171051,12171052) and Fundamental Research Funds for the Central Universities of China(500421358).
通信作者:帅天平(tpshuai@bupt.edu.cn)

的负面影响，进行虚假信息的控制成为一个关键问题，该问题近年来受到了学者的广泛研究[11-18]。现有的关于虚假信息控制问题的研究主要有向网络中传播真实信息[11-13]、阻断节点[14-15]和阻断边[16-17]等方式。

以新浪微博为例，微博中的每个用户都可以接收信息和传播信息。当微博中有人发布了虚假消息时，我们可以通过以下几种方式来控制虚假消息的传播。第一，可以注销掉部分影响力较大的用户的账号（即阻断节点），使他们既不能接收到虚假信息，也不能向其他人传播这些信息；第二，可以对部分影响力较大的账号进行禁言处理（即阻断边），使得这些用户虽然可以接收到虚假信息，但是不能向其他人传播；第三，还可以将真实信息告知部分用户，让他们将这些信息在微博中进行发布，然后传播给其他用户，从而减少相信虚假信息的人数。

在现实生活中，直接注销账号或者禁言处理容易引起用户的不满，尤其是在用户未传播虚假信息时，因此，这两种方式存在诸多不足。而对于第三种方式，用户的不满程度相对较低，对其进行研究具有重要的意义。由于真实信息在网络中的传播速度比虚假信息慢[19]，因此需要采取其他手段来加快真实信息的传播。通过在用户之间加边来加快传播是一种很有效的方式[20-22]。本文将传播真实信息和加边的方式相结合，提出了虚假信息修正最大化问题。目的是将那些已经接受了真实信息的用户推荐给其他用户，并且向他们传播真实信息，最终使得尽可能多的人接受真实信息，从而减小虚假信息的影响。

本文的主要贡献如下：首先，提出了虚假信息修正最大化问题，并且分析了问题的复杂性；其次，提出了一个基于三明治策略的求解算法，并且分析了算法的近似比；最后，通过在3个不同的数据集上进行实验，验证了所提算法的有效性。

本文第2节介绍相关工作；第3节给出了问题的定义，并且分析了问题的复杂性；第4节给出了虚假信息修正最大化问题的求解算法；第5节通过实验验证了算法的有效性；最后总结全文。

2 相关工作

2.1 影响力最大化问题

2003年，Kempe等[1]将影响力最大化问题作为一个组合优化问题来进行考虑，并且证明了在独立级联模型下影响力最大化问题是NP-难问题，所以需要利用近似算法或启发式算法进行求解。Kempe等[1]证明了应用贪心算法求解影响力最大化问题可以得到一个 $1-1/e$ 的近似解。在用贪心算法进行求解时，每一次迭代都需要计算目标函数的值。Chen等[2]证明了在独立级联模型下计算目标函数是#P-难的，他们利用蒙特卡洛模拟来进行估计，但大大增加了贪心算法求解的时间，不适用于大型网络。基于Brogs等[3]提出的反向影响采样的技术，Tang等提出了TIM/TIM$^+$算法[4]和IMM (Influence Maximization Via Martingales)算法[5]，在至少 $1-1/n^l$ 的概率下得到的近似解不小于最优解的 $1-1/e-\varepsilon$ 倍。IMM算法有一个近线性的时间复杂度，是目前较好的算法。

在影响力最大化问题中，除了直接进行求解外，还可以通过在网络中加边来提高影响力的传播。Chaoji等[20]考虑了通过向网络中加入 k 条边，使得被影响的节点个数的期望值最大化的问题，并且证明了这个问题的目标函数是非次模的。他们提出了一个限制的最大概率路径模型，并且证明了在这个模型下，目标函数是次模的。D'Angelo等[21]考虑了在成本约束下通过加边的方式来促进影响传播的问题，与其他研究不同的是这篇论文只考虑了和种子集中的节点相连的边。Lei等[22]利用了文献[20]中提出的限制的最大概率路径模型来考虑在网络中加边的问题，并且提出了一个改进的贪婪算法。

2.2 虚假信息控制问题

本文只考虑了向网络中传播真实信息来控制虚假信息的方式，因此只介绍与传播真实信息相关的研究工作。Budak等[11]提出了一个竞争的独立级联模型，即让虚假信息和真实信息作为两个竞争的活动同时在网络中进行传播。他们证明了在这种竞争的独立级联模型下，虚假信息控制问题是次模问题。He等[12]证明了在竞争的线性阈值模型下，虚假信息控制问题是次模最大化问题。因此，利用贪心算法求解可以得到 $1-1/e$ 的近似比。由于贪心算法的时间复杂度很高，Tong等[13]提出了一个基于反向影响采样技术的算法来求解虚假信息控制问题。

3 模型和问题描述

3.1 影响传播模型

社交网络中常见的影响传播模型有独立级联模型和线性阈值模型等，本文采用独立级联模型。下面给出独立级联模型的定义。

定义1（独立级联模型[1]） 用图 $G=(V,E,P)$ 表示网络，节点集 V 表示网络中的用户，边集 E 表示用户间的关系，每条边 $e=(u,v)$ 有一个对应的影响概率 p_{uv}，p_{uv} 表示 u 被激活后独立激活 v 的概率。初始时被激活的节点被称为种子节点，基于影响概率，从种子节点集 S_0 开始，独立级联模型下的动态传播过程在离散的时间步长下以如下形式完成：初始时，集合 S_0 中的节点被激活，而其他节点都处于不活跃状态；用 S_{t-1} 表示在 $t-1$ 时刻前被激活的节点的集合，在 t 时刻，$S_{t-1}\backslash S_{t-2}$ 中的所有节点 u 以概率 p_{uv} 激活它的非活跃的邻居节点 v，并且 u 只有一次机会去激活 v，如果 v 被激活，则将 v 加入集合 S_t 中；当没有新的节点可以被激活时，传播过程结束。

3.2 活跃边图

定义2（活跃边图） 活跃边图 $X=(V,E_X)$ 是图 G 的一个子图，X 的节点集合与 G 相同，而边集 E_X 是 E 的一个子集。活跃边图 X 发生的概率是在 X 中的边都被选中而不在 X 中的边都没有被选中的概率，即 $\Pr[X]=\prod_{(u,v)\in E_X} p_{uv} \prod_{(u,v)\in E\backslash E_X}(1-p_{uv})$。对于 E 中的所有边，出现在 X 中的边为活跃边，不在 X 中的边为阻断边，传播可以通过活跃边进行，但不能通过阻断边进行。

令 X_G 为图 G 的所有活跃边图的集合。

3.3 问题定义

考虑一个真实信息和虚假信息都在其上传播的社交

网络，称被虚假信息激活的节点为被感染的节点，称初始拥有(接受)称真实信息的节点的集合为真实信息的种子集，那些原本被虚假信息感染但又接受了真实信息的节点为被修正的节点。虚假信息修正最大化就是通过某种策略(如选择传播真实信息)使得修正的期望节点数最多。

以往研究主要考虑如何控制虚假信息传播，本文主要考虑修正已"感染"节点。为简便起见，假设虚假信息的传播过程已经结束。由于影响力最大化问题的相关研究成果丰富，因此本文不考虑真实信息种子集的选取问题，即假设真实信息的种子集已知。本文的核心是考虑如何通过加边来加速真实信息传播，使得尽可能多的"感染"者被修正，即虚假信息修正最大化，问题描述如下。

虚假信息修正最大化问题：给定一个社交网络 $G=(V,E,P)$，被感染的节点的集合 I，真实信息的种子集 S 和候选边集 C。给定一个正整数 k，从候选边集 C 中选择一个有 k 条边的集合 R^* 加入 G 中，使得被修正的节点的期望值的增量最大化，即：

$$R^*=\arg\max_{R\subseteq C,|R|\leqslant k}\Delta f(S,R) \tag{1}$$

其中，$\Delta f(S,R)=f(S,R)-f(S,\varnothing)$，$f(S,R)$ 表示加入边集 R 后被修正的节点的期望值。

3.4 问题的复杂性

前面给出了问题的描述，下面分析问题的复杂度。

定理 1 在独立级联模型下，虚假信息修正最大化问题是 NP-难的问题。

证明：将集合覆盖问题[23]多项式归约到虚假信息修正最大化问题来进行证明。给定一个基础集合 $U=\{u_1,\cdots,u_n\}$，$T=\{T_1,\cdots,T_m\}$ 是 U 的子集的集合，集合覆盖问题就是要判断在 T 中是否存在 k 个子集，使得它们的并集等于 U。

考虑集合覆盖问题的任意一个实例，我们构造一个相应的有 $1+m+n$ 个节点的有向的三部图 G，如图 1 所示。节点 s 为真实信息的种子节点，$A=\{a_1,\cdots,a_m\}$ 中的节点 a_i 对应 T 中的集合 T_i，$B=\{b_1,\cdots,b_n\}$ 中的节点 b_j 对应 U 中的元素 u_j。如果 $u_j\in T_i$，则从 a_i 到 b_j 有一条有向边，且激活概率为 1，从节点 s 到 A 中的 l 个节点有概率为 1 的有向边。我们的问题中只考虑已经被感染的节点，所以令 $I=B$。

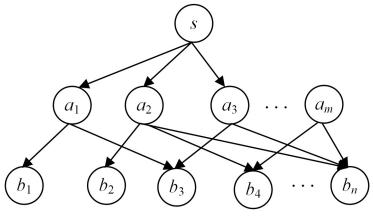


图 1 NP-难图构造实例

Fig. 1 Illustration of NP-hard graph construction

在 A 中选择 k 个节点，将 s 指向这 k 个节点的 k 条边或者这 k 个节点指向 B 中的节点的 k 条边的集合记为集合 R。集合覆盖问题等价于确定在 A 中是否存在这样的 k 个节点，得到集合 R，并且将集合 R 加入图 G 后，可以使得：

$$f(S,R)=f(S,\varnothing)+\Delta f(S,R)=n \tag{2}$$

因为集合覆盖问题是 NP-完全问题，所以虚假信息修正最大化问题是 NP-难的问题。

定理 2 给定 S 和 R 后，计算 $\Delta f(S,R)$ 是 #P-难的。

证明：将有向图中的 s-t 连接数问题[24]多项式归约到 $\Delta f(S,R)$ 的计算问题来进行证明。考虑 s-t 连接数问题的任意一个实例，有向图 $G_1=(V,E)$，如图 2 所示。图中有两个节点 s 和 t，s-t 连接数问题就是计算在 G_1 的所有子图中 s 和 t 相连的子图的数量。假设 G_1 中的每条边上的概率为 0.5，则这个问题等价于计算在 G_1 中节点 s 和 t 相连的概率[2]。

令 t' 为图 G_1 外的一点，构造一个新的有向图 $G_2=(V\cup t',E)$，如图 2 所示。在 G_2 中，令 $S=\{s\}$，$I=\{t'\}$，$R=\{(t,t')\}$，$p_{tt'}=0.5$。假设在 G_2 中已经计算出 $\Delta f(S,R)$，则在 G_1 中 s 和 t 相连的概率为 $\Delta f(S,R)/p_{tt'}=2\Delta f(S,R)$。因此在 G_2 中计算 $\Delta f(S,R)$ 就转化为在 G_1 中求解 s-t 连接数问题。因为 s-t 连接数问题是 #P-完全问题，所以计算 $\Delta f(S,R)$ 是 #P-难的问题。

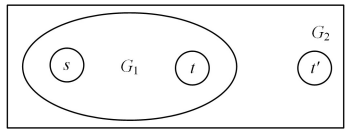


图 2 #P-难图构造实例

Fig. 2 Illustration of #P-hard graph construction

3.5 目标函数的性质

注意到，在给定 S 后，$f(S,\varnothing)$ 的值是确定的，目标函数 $\Delta f(S,R)=f(S,R)-f(S,\varnothing)$ 的性质依赖于 $f(S,R)$。因此，可以通过分析 $f(S,R)$ 的性质来得到 $\Delta f(S,R)$ 的性质。首先证明 $f(S,R)$ 的非负性和单调性。

定理 3 $f(S,R)$ 是非负且单调递增的。

证明：因为 $f(S,R)$ 是被修正的节点的个数的期望值，所以非负性显然。

下面证明 $f(S,R)$ 关于 R 是单调递增的。

对于任意 $R\subseteq C$ 和 $e\in C\backslash R$，可以通过证明 $f(S,R\cup\{e\})-f(S,R)\geqslant 0$ 来证明单调性。

对于 $X_{G\cup R}$ 中的任意一个活跃边图 X，$f(S,R)$ 可以表示为：

$$f(S,R)=\sum_{X\in X_{G\cup R}}\Pr[X]\cdot f_X(S,R) \tag{3}$$

其中，$f_X(S,R)$ 表示在活跃边图 X 中被修正的节点的个数。

对于 X，在 $X_{G\cup R\cup\{e\}}$ 中有两个相应的活跃边图 X' 和 X''，它们的边集分别为 $E_{X'}=E_X$，$E_{X''}=E_X\cup\{e\}$，则 X' 和 X'' 发生的概率分别为：

$$\Pr[X']=\Pr[X]\cdot(1-p_e) \tag{4}$$

$$\Pr[X'']=\Pr[X]\cdot p_e \tag{5}$$

在活跃边图 X' 和 X'' 中，被修正的节点的个数分别为：

$$f_{X'}(S,R\cup\{e\})=f_X(S,R) \tag{6}$$

$$f_{X''}(S,R\cup\{e\})\geqslant f_X(S,R) \tag{7}$$

因此，有：

$$\begin{aligned}
&f(S,R\cup\{e\})-f(S,R)\\
&=\sum_{X\in X_{G\cup R\cup\{e\}}}\Pr[X]\cdot f_X(S,R\cup\{e\})-\sum_{X\in X_{G\cup R}}\Pr[X]\cdot f_X(S,R)\\
&=\sum_{X\in X_{G\cup R}}\{\Pr[X']\cdot f_{X'}(S,R\cup\{e\})+\Pr[X'']\cdot f_{X''}(S,R\cup\{e\})\}-\sum_{X\in X_{G\cup R}}\Pr[X]\cdot f_X(S,R)\\
&=\sum_{X\in X_{G\cup R}}\{\Pr[X]\cdot(1-p_e)\cdot f_{X'}(S,R\cup\{e\})+
\end{aligned}$$

$$\Pr[X]\cdot p_e\cdot f_{X''}(S,R\cup\{e\})-\Pr[X]\cdot f_X(S,R)\}$$
$$=\sum_{X\in X_{G\cup R}}\Pr[X]\cdot p_e\cdot\{X''(S,R\cup\{e\})-f_X(S,R)\}$$
$$\geqslant 0 \tag{8}$$

所以 $f(S,R)$是单调递增的。

下面证明 $f(S,R)$的次模性。

定理 4 $f(S,R)$既不是次模的，也不是超模的。

证明：通过构造两个反例来进行证明，图 3(a)为次模性的反例，图 3(b)为超模性的反例，图中每条边上的概率为 1。

图 3 次模性和超模性反例

Fig. 3 Counter example of submodular and supermodular

在图 3(a)中，令 $C=\{(s,v_1),(v_2,v_4),(v_3,v_4)\}$，$S=\{s\}$，$I=\{v_1,v_2,v_3,v_4\}$，并且令 $R=\emptyset$，$D=\{(v_3,v_4)\}$，$e=(s,v_1)$，则有 $f(S,R)=1$，$f(S,D)=1$。将边 e 分别加入 R 和 D 中，有 $f(S,R\cup\{e\})=3$，$f(S,D\cup\{e\})=4$。因此有 $f(S,R\cup\{e\})-f(S,R)\leqslant f(S,D\cup\{e\})-f(S,D)$，所以 $f(S,R)$是非次模的。

在图 3(b)中，令 $C=\{(s,v_1),(v_2,v_1),(v_2,v_4)\}$，$S=\{s\}$，$I=\{v_1,v_2,v_3,v_4\}$，并且令 $R=\emptyset$，$D=\{(v_2,v_4)\}$，$e=(s,v_1)$，则有 $f(S,R)=1$，$f(S,D)=3$。将边 e 分别加入 R 和 D 中，有 $f(S,R\cup\{e\})=3$，$f(S,D\cup\{e\})=4$。因此有 $f(S,R\cup\{e\})-f(S,R)\geqslant f(S,D\cup\{e\})-f(S,D)$，所以 $f(S,R)$是非超模的。

3.6 目标函数的界

对于次模函数最大化问题，用贪婪算法和 IMM 算法求解可以得到 $1-1/e$ 和 $1-1/e-\varepsilon$ 的近似比[1,5]；而对于非次模情形，用这两个算法求解时不能得到常数近似比。因此，Lu 等[25]提出了一种三明治近似策略来解决非次模和非超模函数的优化问题，并给出了一个数据相关的近似比。利用三明治近似策略求解时，需要先找到目标函数的次模的上界函数和下界函数，然后求解对应于上界函数和下界函数下的同一问题，称为上界问题和下界问题。下面来构造目标函数的次模的上界函数和下界函数。

3.6.1 目标函数的下界函数

为得到问题的次模的下界函数，首先给出如下限制条件：假设边集 R 已知，在网络中加入边集 R，生成一个活跃边图 X。在 X 中，当 S 到 v 的所有路径中有一条路径至多包含一条 R 中的边时，节点 v 就可以被 S 激活。在目标函数中加入这个限制条件后，得到目标函数的下界函数 $L(S,R)$。

下面给出限制条件的示例，如图 4 所示。给定一个网络 G，令 $S=\{s\}$，$I=\{v_3,v_4\}$，$R=\{(s,v_1),(v_1,v_3),(v_1,v_4),(v_2,v_4)\}$。图 4 表示在网络中加入边集 R 后，生成的一个活跃边图 X。从 s 到 v_3 只有一条路径，包含了两条 R 中的边，所以 v_3 不能被激活。从 s 到 v_4 有两条路径，$P_1=sv_1v_4$ 和 $P_2=sv_2v_4$。路径 P_1 中包含了两条 R 中的边，路径 P_2 中只包含了一条 R 中的边，满足限制条件，所以节点 v_4 可以被激活。

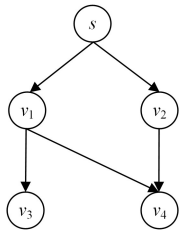


图 4 限制条件实例(1)

Fig. 4 Illustration with restricted conditions(1)

下面分析 $L(S,R)$的性质，有如下定理。

定理 5 $L(S,R)$是非负单调并且次模的。

证明：非负性显然。

下面证明单调性。对于任意 $R\subseteq C$ 和 $e\in C\backslash R$，$X_{G\cup R}$中的任意一个活跃边图 X，以及在 $X_{G\cup R\cup\{e\}}$ 中与 X 对应的 X''，由 $f(S,R)$的单调性的证明类似可得：

$$L(S,R\cup\{e\})-L(S,R)$$
$$=\sum_{X\in X_{G\cup R}}\Pr[X]\cdot p_e\cdot\{L_{X''}(S,R\cup\{e\})-L_X(S,R)\}$$
$$\geqslant 0 \tag{9}$$

所以 $L(S,R)$是单调递增的。

最后证明次模性。对于任意 $R\subseteq D\subseteq C$ 和 $e\in C\backslash D$，式(9)成立；并且对于 $X_{G\cup D}$ 中的任意一个活跃边图 Y 以及在 $X_{G\cup D\cup\{e\}}$ 中对应的 Y''，由单调性的证明类似可得：

$$L(S,D\cup\{e\})-L(S,D)$$
$$=\sum_{Y\in X_{G\cup D}}\Pr[Y]\cdot p_e\cdot\{L_{Y''}(S,D\cup\{e\})-L_Y(S,D)\}$$
$$\geqslant 0 \tag{10}$$

因为 $R\subseteq D$，所以对于 $X_{G\cup D}$中的任意一个活跃边图 Y，在 $X_{G\cup R}$中总有一个活跃边图 X 与之对应，X 是 Y 的一个子图，并且 $E_Y\backslash E_X$ 中只包含 $D\backslash R$ 中的边。

对于 $X_{G\cup R}$中的任意一个活跃边图 X，在 $X_{G\cup D}$中可能有多个活跃边图 Y 与 X 相对应，用 $X_{G\cup D}(X)$表示在 $X_{G\cup D}$中与 X 相对应的所有活跃边图 Y 的集合，则有：

$$L(S,D\cup\{e\})-L(S,D)=\sum_{X\in X_{G\cup R}}\sum_{Y\in X_{G\cup D}(X)}\Pr[Y]\cdot p_e\cdot\{L_{Y''}(S,D\cup\{e\})-L_Y(S,D)\} \tag{11}$$

又因为 $\sum_{Y\in X_{G\cup D}(X)}\Pr[Y]=\Pr[X]$，所以：

$$L(S,R\cup\{e\})-L(S,R)=\sum_{X\in X_{G\cup R}}\sum_{Y\in X_{G\cup D}(X)}\Pr[Y]\cdot p_e\cdot\{L_{X''}(S,R\cup\{e\})-L_X(S,R)\} \tag{12}$$

因此，对于 $X_{G\cup R}$中的任意一个活跃边图 X，可以通过证明下面的不等式对于 $X_{G\cup D}$中与 X 对应的所有活跃边图 Y 均成立来证明 $L(S,R)$是次模的。

$$L_{X\cup\{e\}}(S,R\cup\{e\})-L_X(S,R)\geqslant L_{Y\cup\{e\}}(S,D\cup\{e\})-L_Y(S,D) \tag{13}$$

对于 $X_{G\cup D}$中与 X 对应的任意一个活跃边图 Y，以及 I 中的节点 v，如果在 Y 中 v 不能被 S 激活，但是加入边 e 后其可以被 S 激活，则加入边 e 后 S 到 v 一定存在一条路径，且路径中只包含一条 C 中的边也就是边 e。又因为 $E_Y\backslash E_X$ 中只

包含 C 中的边，所以在 X 中加入边 e 后 v 也可以被 S 激活。反之，依然成立。

因为 $E_X \subseteq E_Y$，所以在 Y 中不能被 S 激活的节点在 X 中也一定不能被 S 激活，而在 X 中不能被 S 激活的节点在 Y 中有可能被 S 激活。

所以对于 $X_{G\cup D}$ 中与 X 对应的任意一个活跃边图 Y，式(13)均成立，即 $L(S,R)$ 是次模的。

3.6.2 目标函数的上界函数

为得到问题的次模的上界函数，给出如下限制条件：假设边集 R 已知，在网络中加入候选边集 C，生成一个活跃边图 X。在 X 中，当 S 到 v 的所有路径中有一条路径满足不包含 C 中的边或者至少包含一条 R 中的边时，节点 v 也可以被 S 激活。在目标函数中加入这个限制条件后，得到目标函数的上界函数 $U(S,R)$。

下面构造一个限制条件的实例，如图 5 所示。给定网络 G，令 $S=\{s\}$，$I=\{v_1,v_3,v_4\}$，$C=\{(s,v_2),(v_1,v_3),(v_1,v_4),(v_2,v_3),(v_2,v_4)\}$，并且令 $R=\{(v_2,v_4)\}$。图 5 表示在网络中加入边集 C 后，生成的一个活跃边图 X。从 s 到 v_1 只有一条路径，并且不包含 C 中的边，满足限制条件，所以 v_1 可以被激活。从 s 到 v_3 有两条路径：$P_1=sv_1v_3$ 和 $P_2=sv_2v_3$。路径 P_1 和路径 P_2 中都包含了 C 中的边并且没有包含 R 中的边，所以节点 v_3 不能被激活。从 s 到 v_4 只有一条路径，并且包含了一条 R 中的边，满足限制条件，所以节点 v_4 可以被激活。

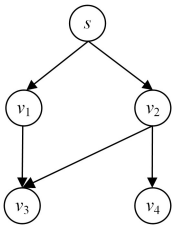


图 5 限制条件实例(2)

Fig. 5 Illustration with restricted conditions(2)

下面分析 $U(S,R)$ 的性质，有如下定理。

定理 6 $U(S,R)$ 是非负单调并且次模的。

证明：非负性显然。

下面证明单调性。将 C 中的边加入图 G 中，得到一个活跃边图的集合 $X_{G\cup C}$。对于任意 $R\subseteq C$ 和 $e\in C\backslash R$，以及 $X_{G\cup R}$ 中的任意一个活跃边图 X，$U(S,R)$ 可以表示为：

$$U(S,R)=\sum_{X\in X_{G\cup R}}\Pr[X]\cdot U_X(S,R) \tag{14}$$

$U_X(S,R)$ 的值与在 $X_{G\cup C}$ 中与 X 相对应的活跃边图 Z 有关。$X_{G\cup C}$ 中可能有多个活跃边图 Z 与 X 相对应，X 是 Z 的一个子图，并且 $E_Z\backslash E_X\subseteq C\backslash R$。

用 $X_{G\cup C}(X)$ 表示在 $X_{G\cup C}$ 中与 X 相对应的所有活跃边图 Z 的集合，则有：

$$\sum_{Z\in X_{G\cup D}(X)}\Pr[Z]=\Pr[X] \tag{15}$$

对于 $X_{G\cup C}$ 中的任意一个活跃边图 Z，在 $X_{G\cup R}$ 中有且仅有一个活跃边图 X 与之对应，用 $X(Z)$ 表示在 $X_{G\cup R}$ 中与 Z 相对应的 X，则：

$$\begin{aligned}U(S,R)&=\sum_{X\in X_{G\cup R}}\Pr[X]\cdot U_X(S,R)\\&=\sum_{X\in X_{G\cup R}}\sum_{Z\in X_{G\cup C}(X)}\Pr[Z]\cdot U_{X(Z)}(S,R)\\&=\sum_{Z\in X_{G\cup C}}\Pr[Z]\cdot U_{X(Z)}(S,R)\end{aligned} \tag{16}$$

要证 $U(S,R)$ 是单调递增的，需要证明 $U(S,R\cup\{e\})-U(S,R)\geqslant 0$。

对于 $X_{G\cup C}$ 中的任意一个活跃边图 Z，在 $X_{G\cup\{e\}}$ 中有且仅有一个活跃边图 $\bar{X}$ 与之对应，用 $\bar{X}(Z)$ 表示在 $X_{G\cup R\cup\{e\}}$ 中与 Z 相对应的 $\bar{X}$，则同理可得：

$$U(S,R\cup\{e\})=\sum_{Z\in X_{G\cup C}}\Pr[Z]\cdot U_{\bar{X}(Z)}(S,R) \tag{17}$$

所以：

$$\begin{aligned}U(S,R\cup\{e\})-U(S,R)=\sum_{Z\in X_{G\cup C}}\Pr[Z]\cdot\{&U_{\bar{X}(Z)}(S,R)-\\&U_{X(Z)}(S,R)\}\end{aligned} \tag{18}$$

若 $e\in E_Z$，则 $\bar{X}(Z)=X(Z)\cup\{e\}$，并且 $U_{\bar{X}(Z)}(S,R\cup\{e\})\geqslant U_{X(Z)}(S,R)$；若 $e\notin E_Z$，则 $\bar{X}(Z)=X(Z)$，$U_{\bar{X}(Z)}(S,R\cup\{e\})=U_{X(Z)}(S,R)$。为了便于区分，令 Z' 表示不包含边 e 的 Z，Z'' 表示包含边 e 的 Z，那么：

$$\begin{aligned}U(S,R\cup\{e\})-U(S,R)=\sum_{Z''\in X_{G\cup C}}\Pr[Z'']\cdot\{&U_{\bar{X}(Z')}(S,R\cup\\&\{e\})-U_{X(Z')}(S,R)\}\geqslant 0\end{aligned} \tag{19}$$

所以 $U(S,R)$ 是单调递增的。

最后证明次模性。对于任意 $R\subseteq D\subseteq C$ 和 $e\in C\backslash D$，式(19)成立。对于 $X_{G\cup C}$ 中的任意一个活跃边图 Z，用 $Y(Z)$ 表示在 $X_{G\cup D}$ 中与 Z 相对应的 Y，$\bar{Y}(Z)$ 表示在 $X_{G\cup D\cup\{e\}}$ 中与 Z 相对应的 $\bar{Y}$，Z'' 表示包含边 e 的 Z，由单调性的证明类似可得：

$$\begin{aligned}U(S,R\cup\{e\})-U(S,R)=\sum_{Z''\in X_{G\cup C}}\Pr[Z'']\cdot\{&U_{\bar{Y}(Z')}(S,R\cup\\&\{e\})-U_{Y(Z'')}(S,R)\}\geqslant 0\end{aligned} \tag{20}$$

要证 $U(S,R)$ 是次模的，只需要证明对于 $X_{G\cup C}$ 中的任意一个 Z''，下面的不等式均成立。

$$\begin{aligned}&U_{X(Z'')\cup\{e\}}(S,R\cup\{e\})-U_{X(Z'')}(S,R)\geqslant\\&U_{Y(Z'')\cup\{e\}}(S,D\cup\{e\})-U_{Y(Z'')}(S,D)\end{aligned} \tag{21}$$

对于 $X_{G\cup C}$ 中的任意一个 Z''，对于 I 中的节点 v，如果在 $Y(Z'')$ 中 v 不能被 S 修正，但是加入边 e 后可以被 S 修正，则说明在 Z'' 中 S 到 v 存在一条路径，并且路径中包含边 e 且不包含 D 中的边。所以，在 $X(Z'')$ 中加入边 e 后 v 也可以被 S 修正。反之，依然成立。

因为 $E_{X(Z'')}\subseteq E_{Y(Z'')}$，所以在 Y 中不能被 S 修正的节点在 X 中也一定不能被 S 修正，而在 X 中不能被 S 修正的节点在 Y 中有可能被 S 修正。

所以对于 $X_{G\cup C}$ 中的任意一个活跃边图 Z''，式(21)均成立，即 $U(S,R)$ 是次模的。

4 虚假信息修正最大化问题的求解算法

IMM 算法[5]是目前求解影响力最大化问题最有效的算法之一。在求解次模函数最大化问题时，利用 IMM 算法可以得到一个近似比为 $1-1/e-\varepsilon$ 的近似解。因此，我们通过修改 IMM 算法来求解下界问题和上界问题。

4.1 IMM 算法的框架

IMM 算法采用了由 Brogs 等[3]提出的反向影响力采样的技术，主要思想是生成足够多的反向可达集，从中找出影响力最大的节点。

定义 3（反向可达集） 从图 G 中采样得到它的活跃边图 X，称在活跃边图 X 中所有可以到达 v 的节点的集合为 v 的反向可达集，称从 V 中均匀随机选取的节点的反向可达集为随机反向可达集。

IMM 算法包含两个过程：采样过程和节点选择过程。采样过程生成足够多的随机反向可达集，使得影响传播的估计值是“足够精确”的；节点选择过程贪婪地选择 k 个可以覆盖最多的反向可达集的节点作为种子节点。

4.2 下界问题的求解算法

为了将 IMM 算法应用到下界函数，将反向可达集概念进行推广。

定义 4（随机反向加入后可修正的边集） 均匀随机选取 I 中的一个节点 v，将边集 C 加入图 G 中，生成一个活跃边图 X；将 X 中所有属于集合 C 的边去掉，得到一个新的活跃边图 X'；如果在 X' 中节点 v 不能被 S 激活，但在 X 中节点 v 可以被 S 激活，则在 X 中存在从 S 到 v 的至多包含一条 C 中的边的路径。将这些路径上的所有属于 C 的边的集合称为 v 的反向加入后可修正的边集，用 $T(v)$ 表示。

引理 1 给定集合 S 和 I，对于任意的集合 R 和 I 中的任意一个节点 v，节点 v 在加入 R 之前不能被 S 激活但在加入 R 后可以被 S 激活的概率 $\Pr[S,v]$ 等于 $R\cap T(v)\neq\emptyset$ 的概率。

给定 S 和 I，在 $G\cup C$ 中生成一个集合 $\mathcal{T}=\{T_1,T_2,\cdots,T_\theta\}$。对于任意的 $T_i\in\mathcal{T}$，定义随机变量 $x_i=\min\{1,|R\cap T_i|\}$。

令 $C_{\mathcal{T}}(R)=\frac{|\{T\in\mathcal{T}|R\cap T\neq\emptyset\}|}{|\mathcal{T}|}$，从而可得 $C_{\mathcal{T}}(R)=\frac{1}{\theta}\sum_{i=1}^{\theta}x_i$。

引理 2 给定集合 S 和 I，对于任意的集合 R，有 $\Delta L(S,R)=E[|I|\cdot C_{\mathcal{T}}(R)]$。

证明：给定集合 S 和 I，对于任意的集合 R，由引理 1 可得：

$$\begin{aligned}\Delta L(S,R)&=L(S,R)-L(S,\emptyset)\\&=\sum_{v\in I}\Pr[S,v]\\&=\sum_{v\in I}\Pr[R\cap T(v)\neq\emptyset]\end{aligned}\tag{22}$$

对于任意反向加入后可修正的边集 T，因为节点 v 是均匀随机选取的，所以有：

$$\Pr[R\cap T\neq\emptyset]=\frac{1}{|I|}\sum_{v\in I}\Pr[S\cap T(v)\neq\emptyset]\tag{23}$$

因此：

$$\begin{aligned}\Delta L(S,R)&=|I|\cdot\Pr[R\cap T\neq\emptyset]\\&=\frac{|I|}{\theta}\cdot\sum_{i=1}^{\theta}\Pr[R\cap T_i\neq\emptyset]\\&=\frac{|I|}{\theta}\cdot\sum_{i=1}^{\theta}\Pr[x_i=1]=\frac{|I|}{\theta}\cdot\sum_{i=1}^{\theta}E[x_i]\\&=E[\frac{|I|}{\theta}\sum_{i=1}^{\theta}x_i]=E[|I|\cdot C_{\mathcal{T}}(R)]\end{aligned}\tag{24}$$

类似于 IMM 算法的思想，我们设计了 Modified-IMM-L 算法来求解下界问题。算法包含两个过程：边选择过程和采样过程。首先给出算法的边选择过程，具体算法描述如算法 1 所示。

算法 1 EdgeSelection-L 算法

输入：$\mathcal{T}$，k

输出：R

1. / * 边选择过程 * /
2. 初始化：R←∅ / * R 为加入的边的集合 * /
3. for i←1 to k do
4. $\quad e^*\leftarrow\arg\max_{e\in C}C_{\mathcal{T}}(R\cup\{e\})-C_{\mathcal{T}}(R)$
5. $\quad R\leftarrow R\cup\{e^*\}$
6. $\quad C\leftarrow C\backslash\{e^*\}$
7. end for
8. 返回 R

算法 1 是求解最大覆盖问题的标准算法。令 R^* 为最优解，$OPT=\Delta L(S,R^*)$ 为对应的最优值。对于大小为 k 的任意边集 $R\subseteq C$，如果 $C_{\mathcal{T}}(R)$ 关于 R 是单调非递减并且次模的，则对于算法 1 返回的解 R'，可以保证 $C_{\mathcal{T}}(R')\geqslant(1-1/e)\cdot C_{\mathcal{T}}(R^*)$。

引理 3 给定集合 S 和 I，对于任意的集合 R，$C_{\mathcal{T}}(R)$ 关于 R 是单调递增并且次模的。

证明：首先证明 $C_{\mathcal{T}}(R)$ 关于 R 是单调递增的。对于任意的边集 $R\subseteq C$ 和任意的一条边 $e\in C\backslash R$ 以及任意的 $T_i\in T$，定义随机变量 $x_i(R)=\min\{1,|R\cap T_i|\}$ 和 $x_i(R\cup\{e\})$，则有：

$$C_{\mathcal{T}}(R\cup\{e\})-C_{\mathcal{T}}(R)=\frac{1}{\theta}\cdot\sum_{i=1}^{\theta}(x_i(R\cup\{e\})-x_i(R))\tag{25}$$

通过证明 $x_i(R\cup\{e\})-x_i(R)\geqslant0$ 来证明 $C_{\mathcal{T}}(R\cup\{e\})-C_{\mathcal{T}}(R)\geqslant0$。

当 $x_i(R)=1$ 时，可知 $R\cap T_i\neq\emptyset$。因此，有 $(R\cup\{e\})\cap T_i\neq\emptyset$，即 $x_i(R\cup\{e\})=1$，结论成立。

下面证明 $C_{\mathcal{T}}(R)$ 关于 R 是次模的。对于任意的边集 $R_1\subseteq R_2\subseteq C$ 和任意的一条边 $e\in C\backslash R_2$，以及任意的 $T_i\in T$，定义随机变量 $x_i(R_1)$，$x_i(R_2)$，$x_i(R_1\cup\{e\})$ 和 $x_i(R_2\cup\{e\})$。

要证次模性，需证：

$$C_{\mathcal{T}}(R_1\cup\{e\})-C_{\mathcal{T}}(R_1)\geqslant C_{\mathcal{T}}(R_2\cup\{e\})-C_T(R_2)\tag{26}$$

通过证明下面的不等式成立来证明式(26)成立。

$$x_i(R_1\cup\{e\})-x_i(R_1)\geqslant x_i(R_2\cup\{e\})-x_i(R_2)\tag{27}$$

当 $x_i(R_2\cup\{e\})-x_i(R_2)=1$ 时，可以得到 $(R_2\cup\{e\})\cap T_i\neq\emptyset$，$R_2\cap T_i=\emptyset$，因此有 $\{e\}\cap T_i\neq\emptyset$。又因为 $R_1\subseteq R_2$，$R_2\cap T_i=\emptyset$，所以有 $R_1\cap T_i=\emptyset$，$(R_1\cup\{e\})\cap T_i\neq\emptyset$。即 $x_i(R_1\cup\{e\})-x_i(R_1)=1$，结论成立。

其次，考虑 Modified-IMM-L 算法的采样过程，具体算法描述见算法 2。在采样过程中，需要先确定采样个数 θ 的值。定理 7 在保证近似的前提下，给出了 θ 的范围。

定理 7 给定任意 $\varepsilon_1\leqslant\varepsilon$ 和任意 $\delta_1,\delta_2\in(0,1)$，满足 $\delta_1+\delta_2\leqslant1/n^l$，令：

$$\theta_1=\frac{2|I|\cdot\log(1/\delta_1)}{OPT\cdot\varepsilon_1^2}\tag{28}$$

$$\theta_2=\frac{(2-2/e)\cdot|I|\cdot\log\left(\log\binom{|C|}{k}/\delta_2\right)}{OPT\cdot(\varepsilon-(1-1/e)\varepsilon_1)^2}\tag{29}$$

当 $\theta \geqslant \max\{\theta_1, \theta_2\}$ 时，可以保证算法 1 在至少 $1-1/n^l$ 的概率下返回一个 $1-1/e-\varepsilon$ 的近似解。

证明：略。类似于文献[5]中定理 1 的证明。

算法 2 Sampling-L 算法

输入：G,S,I,C,k,ε,l

输出：$\mathcal{T}$

1. /* 采样过程 */
2. 初始化：$\mathcal{T}\leftarrow\emptyset$，LB$\leftarrow$1，$\varepsilon'\leftarrow\sqrt{2}\cdot\varepsilon$ /* $\mathcal{T}$为反向加入后可修正的边集的集合，LB 为 L(S,R)的下界 */
3. /* 确定 θ 的值 */
4. for k$\leftarrow$1 to $\log_2|I|-1$ do
5. x$\leftarrow|I|/2^i$
6. $\lambda'\leftarrow\dfrac{\left(2+\frac{2}{3}\varepsilon'\right)\cdot\left(\log\binom{|C|}{k}+l\cdot\log n+\log\log_2|I|\right)\cdot|I|}{\varepsilon'^2}$
7. $\theta_i\leftarrow\lambda'/x$
8. while $|\mathcal{T}|\leqslant\theta_i$ do
9. 从 I 中均匀随机选择一个节点 v
10. 生成一个集合 T(v)，将这个集合加入到 $\mathcal{T}$ 中
11. end while
12. $R_i\leftarrow$EdgeSelection$-$L($\mathcal{T}$,k)
13. if $|I|\cdot C_{\mathcal{T}}(R_i)\geqslant(1+\varepsilon')\cdot x$ then
14. LB$\leftarrow|I|\cdot C_{\mathcal{T}}(R_i)/(1+\varepsilon')$
15. break
16. end if
17. end for
18. $\alpha\leftarrow\sqrt{l\cdot\log n+\log 2}$
19. $\beta\leftarrow\sqrt{(1-1/e)\cdot\left(\log\binom{|C|}{k}+\log 2+l\log n\right)}$
20. $\lambda\leftarrow 2|I|\cdot((1-1/e)\cdot\alpha+\beta)^2\cdot\varepsilon^{-2}$
21. $\theta\leftarrow\lambda/$LB
22. /* 采样 */
23. while $|\mathcal{T}|\leqslant\theta$ do
24. 从 I 中均匀随机选择一个节点 v
25. 生成一个集合 T(v)，将这个集合加入到 $\mathcal{T}$ 中
26. end while
27. 返回 $\mathcal{T}$

下面确定 θ 的值。令 $\delta_1=\delta_2=1/(2n^l)$，当 $\theta_1=\theta_2$ 时，θ 取最小值。当且仅当 $\varepsilon_1=\varepsilon\cdot\frac{\alpha}{(1-1/e)\cdot\alpha+\beta}$ 时，有 $\theta_1=\theta_2$。其中：

$$\alpha=\sqrt{\log 2+l\log n} \tag{30}$$

$$\beta=\sqrt{(1-1/e)\cdot\left(\log\binom{|C|}{k}+\log 2+l\log n\right)} \tag{31}$$

此时，$\theta=\frac{2|I|\cdot((1-1/e)\cdot\alpha+\beta)^2}{OPT\cdot\varepsilon^2}$。令 $\lambda=2|I|\cdot((1-1/e)\cdot\alpha+\beta)^2\cdot\varepsilon^{-2}$，当 $\theta\geqslant\lambda/OPT$ 时，算法 1 可以在至少 $1-1/n^l$ 的概率下返回一个 $1-1/e-\varepsilon$ 的近似解。

由于问题是 NP-难问题，因此直接计算 OPT 是非常困难的，所以希望通过确定 OPT 的一个下界 LB 来近似 θ，即令 $\theta=\lambda/LB\geqslant\lambda/OPT$。下界 LB 的确定过程参考了 IMM 算法[5]的思想，具体过程为算法 2 中的第 3－17 行。下界 LB 的值确定后，可以得到 θ 的值。然后进行采样，算法 2 中的第 23－26 行为采样过程。

定理 8 在至少 $1-1/n^l$ 的概率下，算法 2 可以返回一个集合 $\mathcal{T}$，并且 $|\mathcal{T}|\geqslant\lambda/OPT$。

证明：略。类似于文献[5]中定理 2 的证明。

首先利用算法 2 采样得到集合 $\mathcal{T}$，然后利用算法 1 选择一个有 k 条边的集合作为最终解。将这两个过程放在一起，可以得到最后的 Modified-IMM-L 算法。具体算法框架见算法 3。

算法 3 Modified-IMM-L 算法

输入：G,S,I,C,k,ε,l

输出：R

1. $G'\leftarrow G\cup C$
2. l$\leftarrow$l$+\log 2/\log n$
3. $\mathcal{T}\leftarrow$Sampling$-$L(G',S,I,C,k,ε,l)
4. R$\leftarrow$EdgeSelection$-$L($\mathcal{T}$,k)
5. 返回 R

下面分析算法 3 的性能。

定理 9 对于下界问题，算法 3 可以在至少 $1-1/n^l$ 的概率下返回一个 $1-1/e-\varepsilon$ 的近似解。

证明：根据定理 7 和定理 8 和联合界，算法 3 在至少 $1-2/n^l$ 的概率下返回一个 $1-1/e-\varepsilon$ 的近似解。令 $l=l+\log 2/\log n$，则概率可以改为 $1-1/n^l$。

4.3 上界问题的求解算法

类似于下界问题情形，推广反向可达集概念。

定义 5(随机反向加入后可修正的边集) 任取 I 中的一个节点 v，将边集 C 加入图 G 中，生成一个活跃边图 X；将 X 中所有属于集合 C 的边去掉，得到一个新的活跃边图 X'；如果在 X' 中节点 v 不能被 S 激活，但在 X 中节点 v 可以被 S 激活，则将在 X 中从 S 到 v 的路径上所有属于 C 的边的集合称为 v 的随机反向加入后可修正的边集，用 $W(v)$ 表示。

引理 4 给定集合 S 和 I，对于任意的集合 R 和 I 中的任意一个节点 v，节点 v 在加入 R 之前不能被 S 激活但在加入 R 后可以被 S 激活的概率 $\Pr[S,v]$ 等于 $R\cap W(v)\neq\emptyset$ 的概率。

给定 S 和 I，在 $G\cup C$ 中生成一个集合 $\mathcal{W}=\{W_1,W_2,\cdots,W_\theta\}$。对于任意的 $W_i\in\mathcal{W}$，定义随机变量 $y_i=\min\{1,|R\cap W_i|\}$。

令 $C_{\mathcal{W}}(R)=\frac{|\{W\in\mathcal{W}|R\cap W\neq\emptyset\}|}{|\mathcal{W}|}$，从而可得 $C_{\mathcal{W}}(R)=\frac{1}{\theta}\sum_{i=1}^{\theta}y_i$。

引理 5 给定集合 S 和 I，对于任意的集合 R，有 $\Delta U(S,R)=E[|I|\cdot C_{\mathcal{W}}(R)]$。

证明：略。与引理 2 证明类似。

引理 6 给定集合 S 和 I，对于任意的集合 R，$C_{\mathcal{W}}(R)$ 关于 R 是单调递增并且次模的。

证明：略。与引理 3 证明类似。

将算法 1、算法 2 和算法 3 中的 $\mathcal{T}$ 替换为 $\mathcal{W}$，可以得到 EdgeSelection-U 算法(算法 4)、Sampling-U 算法(算法 5)和 Modified-IMM-U 算法(算法 6)。由于算法 4、算法 5 和算法 6

与算法 1、算法 2 和算法 3 只进行了将 $\mathcal{T}$ 替换为 $\mathcal{W}$ 的改动，因此算法框架不具体写出。

定理 10 对于上界问题，Modified-IMM-U 算法可以在至少 $1-1/n^l$ 的概率下返回一个 $1-1/e-\varepsilon$ 的近似解。

证明：证明过程与 Modified-IMM-L 算法求解下界问题的近似比的证明过程类似。

4.4 三明治近似策略

得到上界问题和下界问题的近似解后，我们利用三明治近似策略对原问题进行求解，具体算法描述如算法 7 所示。

算法 7 Sandwich 算法

输入：G，S，I，C，k，ε，l

输出：R

1. 令 R_L 为算法 3 求解下界问题得到的集合
2. 令 R_U 为算法 6 求解上界问题得到的集合
3. 令 R_O 为求解原问题得到的集合
4. $R\leftarrow \arg\max_{R'\in\{R_U,R_L,R_O\}}\Delta f(S,R')$
5. 返回 R

下面分析 Sandwich 算法的近似比。

定理 11 令 R_L^* 为下界问题的最优解，R_U^* 为上界问题的最优解，R^* 为原问题的最优解，R 为 Sandwich 算法返回的解，则至少在 $1-1/n^l$ 的概率下有：

$$\Delta f(S,R)\geqslant \max\left\{\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)},\frac{\Delta L(S,R_L^*)}{\Delta f(S,R^*)}\right\}\cdot\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta f(S,R^*) \tag{32}$$

证明：由定理 9 和定理 10 可得，至少在 $1-1/n^l$ 的概率下有：

$$\begin{aligned}\Delta f(S,R_U)&=\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}\cdot\Delta U(S,R_U)\\&\geqslant\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}\cdot\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta U(S,R_U^*)\\&\geqslant\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}\cdot\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta U(S,R^*)\\&\geqslant\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}\cdot\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta f(S,R^*)\end{aligned} \tag{33}$$

和

$$\begin{aligned}\Delta f(S,R_L)&\geqslant\Delta L(S,R_L)\\&\geqslant\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta L(S,R_L^*)\\&\geqslant\frac{\Delta L(S,R_L^*)}{\Delta f(S,R^*)}\cdot\left(\left(1-\frac{1}{e}-\varepsilon\right)\cdot\Delta f(S,R^*)\right)\end{aligned} \tag{34}$$

令 $R=\arg\max_{R'\in\{R_U,R_L,R_O\}}\Delta f(S,R')$，则至少在 $1-1/n^l$ 的概率下可以得到式(32)。因此，结论成立。

5 实验与结果分析

5.1 实验设置

5.1.1 数据集

为了验证所提出算法的有效性，在 3 个不同大小的真实网络的数据集 ca-netscience，soc-wiki-Vote 和 email-univ 上对算法进行仿真实验，并将其与其他启发式算法进行对比。第一个数据集是合作关系网络，表示研究者在进行理论和实验研究时的合作关系。第二个数据集是维基百科投票网络，包含了从维基百科成立到 2008 年 1 月以来维基百科上的所有投票数据。第三个数据集是西班牙某大学的电子邮箱交流网络。表 1 列出了 3 个数据集的信息[26]。其中，n 为网络中节点的个数，m 为网络中边的条数。

表 1 数据集

Table 1 Statistics of datasets

Dataset	n	m
Ca-netscience	379	914
Soc-wiki-Vote	889	2914
Email-univ	1133	5451

5.1.2 参数设置

对于这 3 个网络，边上的激活概率 p 在(0，1)上均匀随机选取。在网络中随机选择部分节点作为被感染的节点的集合 I，选择出度最大的节点的集合作为真实信息的种子集 S，选择图 G 的补图中的边的集合作为候选边集 C，边上的激活概率在(0，1)上均匀随机选取。对于第一个数据集，令 $|S|=30$，$|I|=100$，$|C|=200$；对于第二个和第三个数据集，令 $|S|=30$，$|I|=200$，$|C|=500$。由于 I 中的节点是随机选取的，为了保证实验的准确性，在所有的实验中，选择 5 个集合 I，对得到的结果取平均值。在我们的算法中，令 $l=1$。在计算目标函数值时，蒙特卡洛模拟的次数为 10000。

5.1.3 算法比较

由于本文考虑的是一类新的模型，没有直接相关算法，因此我们将所提出的算法和一些经典启发式算法进行比较，这些经典的启发式算法包括 Random 算法[1]、Degree 算法[27]和 Weight 算法[28]等。Random 算法是在候选边集中随机选择 k 条边，Degree 算法在候选边集中选择边的头节点的出度最大的 k 条边，Weight 算法是在候选边集中选择权重最大的 k 条边。

5.2 实验结果与分析

首先分析 Sandwich 算法中 ε 的取值，令 $k=50$。图 6 给出 ε 从 0.1 到 0.5 变化时，Sandwich 算法在 3 个数据集上进行实验得到的被修正的节点的期望值的增量的变化情况。其中，横轴表示 ε 的大小，纵轴表示被修正的节点的期望增量。图 7 给出 ε 从 0.1 到 0.5 变化时，Sandwich 算法在 3 个数据集上求解的时间。其中，横轴表示 ε 的大小，纵轴表示求解时间。

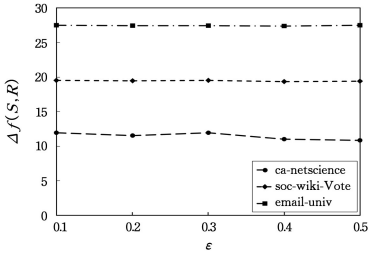


图 6 Sandwich 算法求得的被修正的节点的期望值的增量 $\Delta f(S,R)$ 随 ε 的变化情况

Fig. 6 Increased expected number of corrected nodes $\Delta f(S,R)$ by Sandwich algorithm varies with ε

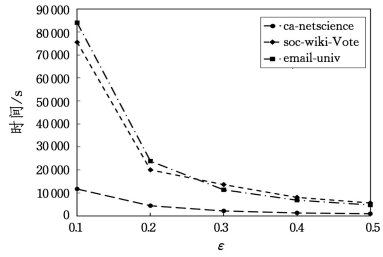
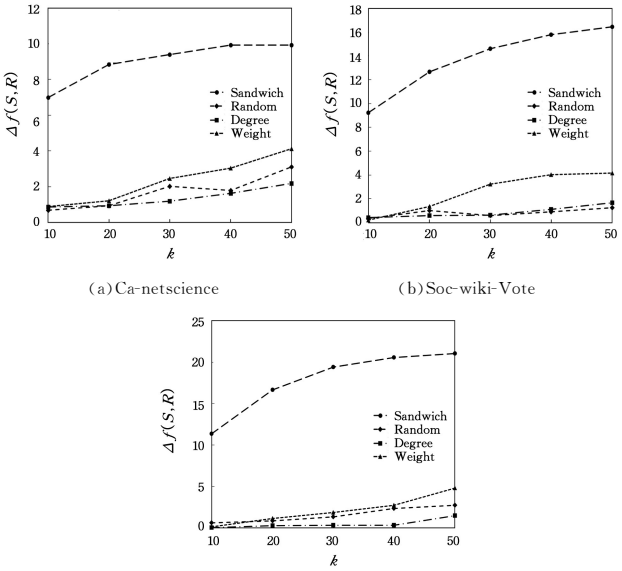


图 7 Sandwich 算法的运行时间随 ε 的变化情况

Fig. 7 Run time of the Sandwich algorithm varies with ε

从图 6 和图 7 可以看出，在这 3 个数据集上，ε 的取值对目标函数值基本没有影响，但是求解时间受 ε 取值的影响。在这 3 个数据集上，随着 ε 的增大，求解时间均不断减少，并且当 $\varepsilon=0.5$ 时，求解时间最短。因此，在后面的实验中，将 ε 设为 0.5，这样既可以保证解的质量，又可以缩短算法的运行时间。

图 8 为不同算法在 3 个数据集上进行实验得到的结果，每一个子图分别展示了在每一个数据集上当 k 从 10 到 50 进行变化时，4 个算法求得的被修正的节点的期望值的增量的变化情况。其中，横轴表示选择的边集的大小，纵轴表示被修正的节点的期望增量。从图中可以看出，当网络的规模增加时，求得的结果也会增大。在这 3 个数据集上，所提算法得到的结果均明显优于其他 3 个启发式算法。



(a) Ca-netscience (b) Soc-wiki-Vote

(c) Email-univ

图 8 不同算法求得的被修正的节点的期望值的增量随 k 的变化情况

Fig. 8 Increased expected number of corrected nodes in different algorithms varies with k

表 2 列出当 $k=50$ 时，不同算法在 3 个数据集上的运行时间。从表中可以看出，当网络的规模增加时，运行时间也会增加。在这 3 个数据集上，其他启发式算法的时间均优于文中所提算法。

结合表 2 和图 8 可以看出，虽然启发式算法的时间优于文中所提的算法，但是其结果却比所提算法差很多。因为启发式算法求解是基于直观或经验构造出问题的一个可行解，虽然求解时间短，但没有理论保证。文中所提算法对上下界问题和原问题进行求解，虽然求解时间较长，但是可以得到一个有质量保证的解。我们后期准备在保证解的质量的前提下，考虑缩短算法的运行时间，以使算法更加完善。

表 2 $k=50$ 时的运行时间

Table 2 Run time when $k=50$

(单位：s)

	Ca-netscience	Soc-wiki-Vote	Email-univ
Sandwich	253.31	1 317.95	1 669.27
Random	7.47	26.81	44.21
Degree	7.41	25.39	44.40
Weight	7.55	25.47	46.58

在 4.4 节中我们证明了文中算法有一个数据相关的近似比，我们可以计算近似比的一个下界 $\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}\cdot\left(1-\frac{1}{e}-\varepsilon\right)$。在实验中，我们分别计算了在 3 个数据集上求解得到的 $\frac{\Delta f(S,R_U)}{\Delta U(S,R_U)}$ 的值，如图 9 所示。从图中可以看出，在 3 个数据集上，比值都随着加入边的个数的增加而增加。

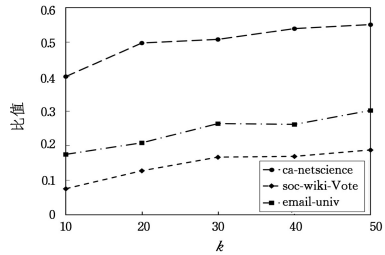


图 9 数据相关的近似比

Fig. 9 Data-dependent approximation ratio

结束语 虚假信息在在线社交网络中的传播速度非常快，容易造成严重的影响，因此需要制定相关的措施来降低虚假信息的影响。本文引入了虚假信息修正最大化问题，通过向网络中传播真实信息，使得那些已经被虚假信息感染的节点尽可能多地被修正，转而接受真实信息。然后通过加边的方式来促进真实信息的传播，从而减少虚假信息的影响。虚假信息修正最大化问题的目标函数既不是次模的，也不是超模的，因此采用三明治近似策略进行求解。为此，构造了目标函数的次模的上界函数和下界函数，得到对应的上界问题和下界问题。通过修改 IMM 算法对上界问题和下界问题进行求解，利用所得的上下界问题的解来求得原问题的近似解。最后，在 3 个真实网络的数据集上进行了仿真实验，实验结果表明所提算法具有较好的性能。在本文中，我们假设虚假信息传播已经结束的情形，具有一定的局限性。未来将进一步研究虚假信息传播尚未结束情形下的虚假信息修正最大化问题，即虚假信息和真实信息同时在网络中传播的具有竞争性质的虚假信息修正，这更有挑战和实际意义。

参 考 文 献

[1] KEMPE D, KLEINBERG J, 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. ACM, 2003: 137-146.

[2] CHEN W,WANG C,WANG Y J. 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. ACM,2010:1029-1038.

[3] BROGS C,BRAUTBAR M,CHAYES J,et al. Maximizing social influence in nearly optimal time [C]// Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms. SIAM,2014:946-957.

[4] TANG Y Z,XIAO X K,SHI Y C. Influence maximization:near-optimal time complexity meets practical efficiency[C]// Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. ACM,2014:75-86.

[5] TANG Y Z,SHI Y C,XIAO X K. Influence Maximization in Near-Linear Time:A Martingale Approach[C]// Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM,2015:1539-1554.

[6] HU Q C,ZHANG Y,XING C X. K-clique Heuristic Algorithm for Influence Maximization in Social Network [J]. Computer Science,2018,45(6):32-35.

[7] KONG F,LI Q Z,LI S. Survey on Online Influence Maximization [J]. Computer Science,2020,47(5):7-13.

[8] TAN Q,ZHANG F L,ZHANG Z Y,et al. Modeling Methods of Social Network User Influence [J]. Computer Science,2021,48(2):76-86.

[9] TAN Q,ZHANG F L,WANG T,et al. Social Network User Influence Evaluation Algorithm Integrating Structure Centrality [J]. Computer Science,2021,48(7):124-129.

[10] DOERR B,FOUZ M,FRIEDRICH T. Why rumors spread so quickly in social networks [J]. Communications of the ACM,2012,55(6):70-75.

[11] BUDAK C,AGRAWAL D,ABBADI E A. Limiting the spread of misinformation in social networks[C]// Proceedings of the 20th International Conference on World Wide Web. ACM,2011:665-674.

[12] HE X R,SONG G J,CHEN W,et al. Influence blocking maximization in social networks under the competitive linear threshold model[C]// Proceedings of the 2012 SIAM International Conference on Data Mining. SIAM,2012:463-474.

[13] TONG G M,WU W L,GUO L,et al. An efficient randomized algorithm for rumor blocking in online social networks[J]. IEEE Transactions on Network Science and Engineering,2020,7(2):845-854.

[14] WANG S Z,ZHAO X J,CHEN Y,et al. Negative influence minimizing by blocking nodes in social networks [C]// Proceedings of the 17th AAAI Conference on Late-Breaking Developments in the Field of Artificial Intelligence. ACM,2013:134-136.

[15] YAN R D,LI D Y,WU W L,et al. Negative influence minimizing by blocking nodes in social networks [J]. IEEE Transactions on Network Science and Engineering,2020,7(3):1067-1078.

[16] KIMURA M,SAITO K,MOTODA H. Minimizing the spread of contamination by blocking links in a network[C]// Proceedings of the 23rd National Conference on Artificial Intelligence-Volume 2. ACM,2008:1175-1180.

[17] KUHLMAN C J,TULI G,SWARUP S,et al. Blocking simple and complex contagion by edge removal [C]// IEEE 13th International Conference on Data Mining. IEEE,2013:399-408.

[18] CHEN J Y,ZHANG D J,LIN X,et al. False Message Propagation Suppression Based on Influence Maximization [J]. Computer Science,2020,47(S1):17-23.

[19] VOSOUGHI S,ROY D,ARAL S. The spread of true and false news online [J]. Science,2018,359(6380):1146-1151.

[20] CHAOJI V,RANU S,RASTOGI R,et al. Recommendations to boost content spread in social networks[C]// Proceedings of the 21st International Conference on World Wide Web. ACM,2012:529-538.

[21] D'ANGELO G,SEVERINI L,VELAJ Y. Recommending links through influence maximization [J]. Theoretical Computer Science,2019,764:30-41.

[22] YU L,LI G H,YUAN L. Maximizing Boosted Influence Spread with Edge Addition in Online Social Networks[J]. ACM/IMS Transactions on Data Science,2020,1(2):1-21.

[23] KARP R M. Reducibility among combinatorial problems[C]// Complexity of Computer Computations. Springer,1972:85-103.

[24] VALIANT L G. The complexity of enumeration and reliability problems [J]. SIAM Journal on Computing,1979,8(3):410-421.

[25] LU W,CHEN W,LAKSHMANAN L V S. From competition to complementarity:Comparative influence diffusion and maximization [J]. Proceedings of the VLDB Endowment,2015,9(2):60-71.

[26] ROSSI R A,AHMED N K. The network data repository with interactive graph analytics and visualization[EB/OL]. http://networkrepositoy.com.

[27] GAO C,LIU J,ZHONG N. Network immunization and virus propagation in email networks experimental evaluation and analysis[J]. Knowledge and Information Systems,2011,27(2):60-71.

[28] KHALIL E,DIKINA B,SONG L. Scalable diffusion-aware optimization of network topology [C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2014:1226-1235.

SONG Xin-yue, born in 1998, postgraduate. Her main research interests include combinatorial optimization and social network.

SHUAI Tian-ping, born in 1977, Ph.D, associate professor. His main research interests include combinatorial optimization and social network.

(责任编辑:柯颖)