doi:10.11887/j.cn.201903025 http://journal.nudt.edu.cn

复杂网络影响力极大化快速评估算法*

王潇杰[1],赵城利[1],张 雪[1],易东云[1,2]
(1. 国防科技大学 文理学院,湖南 长沙 410073;
2. 国防科技大学 高性能计算国家重点实验室,湖南 长沙 410073)

摘 要:分析复杂网络中影响力极大化问题,设计一种新的启发式算法框架。针对信息传递中节点的交互方式进行分析,给出节点在任意时刻处于信息接收态的概率。通过期望计算得到种子节点集传播影响力的近似估计,实现集群影响力快速计算,进而得到基于序列采样的影响力极大化快速评估算法。特别地,对于六个来自不同领域的真实网络上的影响力极大化问题进行了研究,仿真结果表明:该方法能够高效识别网络中具有重要传播影响力的节点集,在三种常见度量准则下的表现均明显优于三种影响力极大化问题基准算法。

关键词:复杂网络;传播动力学;影响力极大化;序列采样;启发式算法
中图分类号:N94 **文献标志码**:A **文章编号**:1001-2486(2019)03-166-08

Maximizing spread of influence in complex networks through fast evaluation

WANG Xiaojie[1], *ZHAO Chengli*[1], *ZHANG Xue*[1], *YI Dongyun*[1,2]
(1. College of Liberal Arts and Sciences, National University of Defense Technology, Changsha 410073, China;
2. State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China)

Abstract: In order to study the influence maximization problem in complex networks, a heuristic framework was developed. Based on the in-depth analysis of information transmit process between node pairs, the probability of a node being in the informed state was obtained, and then an approximation of spreading influence of seed nodes was conducted through expectation calculation. A fast evaluation algorithm was proposed based on sequential seeding strategy. Specifically, simulation results on six real networks from various fields all show that the proposed algorithm is able to distinguish a small set of influential seed nodes. Moreover, the influence scope of the seed nodes selected by the method is significantly better than three benchmark influence maximization algorithms under three common measurements.

Keywords: complex network; spreading dynamics; influence maximization; sequential seeding; heuristic algorithm

研究网络科学的主要目的之一是解决复杂网络上的动力学问题,对于网络上动力学特性的研究一直是网络科学研究领域的重点与难点。特别地,对于网络传播动力学的研究更是具有极为重要的现实意义。传播现象在现实生活中无处不在,例如谣言在社交媒体上的传播[1],传染病在人群中的传播[2],以及电力网络的级联故障[3]等。对于传播动力学的研究可以揭示复杂网络中的传播机理以及动力学行为,从而提供对这些行为的切实可行的控制方法,创造巨大的经济价值和社会价值。在现实生活中,人们经常面临的一个实际问题就是高效地寻找少部分具有重要影响力的初始传播者。例如,对于一个新产品的市场营销而言,选取少量的种子用户作为产品推广人,利用口碑营销的方式迅速打开市场、提高产品知名度,是十分重要的。在网络科学的领域中,这种通过选取少量节点作为初始节点,以极大化这些节点在整个网络中的传播影响力的问题,称为影响力极大化问题[4]。

关于影响力极大化问题的研究可以追溯到Domingos等的工作[5],他们第一次研究了如何将影响力极大化问题表述为一个算法问题,并提出了一种基于概率的算法用于近似求解。此后,Kempe等[6]利用社会网络分析的方法系统地研究了这个问题,他们发现,网络中的影响力极大化问题的实质是NP-hard的组合优化问题,其精确

* 收稿日期:2018-03-07
基金项目:国家重点基础研究发展计划资助项目(2017YCF1200301)
作者简介:王潇杰(1990—),男,江苏无锡人,博士研究生,E-mail:wangxiaojie0817@gmail.com;
易东云(通信作者),男,教授,博士,博士生导师,E-mail:dongyun.yi@gmail.com

求解非常困难。进一步地,他们给出了一个基于贪婪思想的算法以近似求解影响力极大化问题,并证明了该算法的精确度下界。然而,由于该算法十分耗时,往往只能应用在不大的网络上进行求解。基于类似的贪婪思想,Leskovec 等[7]通过对影响力极大化问题的子模块性质进行研究,提出了高效贪婪选择(Cost-Effective Lazy Forward selection, CELF)算法,提高了 Kempe 算法的效率。在 CELF 算法的基础上,Goyal 等[8]对于算法步骤进一步优化,提出 CELF + + 算法,极大提高了 CELF 算法的效率。

虽然基于贪婪思想的算法大多可以取得较为满意的结果,较高的算法复杂度往往成了限制它们应用的一个关键因素。因此,越来越多的学者开始尝试提出启发式算法来近似求解影响力极大化问题。Narayanam 等[9]另辟蹊径,通过引入博弈论中的 Shapely 值的概念,提出了基于 Shapely 值的重要节点选择(ShaPley value based Influential Nodes,SPIN)算法。Zhao 等[10]受到地图着色问题的启发,提出了一种基于着色的算法。Zhang 等[11]提出了基于迭代的投票排名算法,可以有效地识别一组影响力较大的离散节点。通过分析节点的相对关系,Chen 等[12]提出了折扣度算法,很好地平衡了算法的计算效率和精确度,取得了不弱于贪婪算法的结果,已成为现在影响力极大化问题的标准算法之一。Lü 等[13]分析了度中心、H 指数以及核数的关系,验证了 H 指数在描述节点重要性的良好表现。近年来,Morone 等[14]研究了网络中的级联失效问题,提出了基于最优渗流理论的集群影响算法,可以有效识别级联失效问题中的重要节点。

通常而言,大多数启发式算法往往通过某种重要性指标来间接反映节点的影响力。本文结合网络上具体的传播动力学分析,提出快速评估算法,通过期望计算的方式直接估计节点的传播影响力,并进一步运用序列采样的策略进行种子集的快速选择,在保证算法效率的基础上极大提高了算法的精度。

1　影响力极大化问题分析

沿用图论中的相关记号,网络可以用图 $G(V, E)$来表示[15],其中节点 V 代表网络中的个体,边 E 代表个体之间的联系。例如对一个在线社交网络而言,用户就是网络中的节点,用户之间的好友关系自然而然地构成了网络中的边。

在一个网络中,影响力极大化问题可以描述为:如何寻找网络中的 L 个节点 $S \subset V$ 作为种子节点(即信息的初始传播者),将信息传播到网络中尽可能大的范围。

对于影响力极大化问题,一个直观的想法是,如果可以对节点在网络中的重要性进行排序,依次选取排名靠前的节点,它们在整个网络中的集群影响力自然会大。例如,可以利用网络中的各种中心性指标[15]对节点进行排序,依次选取排序中靠前的部分节点作为种子节点来极大化它们在整个网络中的影响力。事实上,这种选取方式存在一个严重的问题——由于节点在网络中的影响力存在着相互重叠的区域,节点间的相对位置必然会对它们的集群影响力造成重要的影响,这也是影响力极大化问题的难点所在。

通常而言,针对网络上不同的信息传播方式,最优的初始传播者会有所不同,很难找到一种统一的算法适用于所有传播动力学下的影响力极大化问题。在下面的分析中,将主要讨论一种简单的信息传递模型。假定网络中所有的节点都可能具有两种状态,即接收态与未接收态。处于未接收态的节点,代表网络中还没有接收到信息的普通个体;而处于接收态的节点,则代表那些接收到信息的个体,它们会以概率 p 向周围邻居广播消息,进而将消息扩散到整个网络中。

2　影响力极大化快速评估模型

记 I_v^r 为节点 v 在第 r 轮信息传递过程中处于接收态的概率。初始时刻,只有种子节点处于接收态($I_v^0 = 1, v \in S$),其他节点均处于未接收态($I_v^0 = 0, v \in V - S$)。显然,所有的种子节点 $v \in S$ 在任意时刻均处于接收态($I_v^r \equiv 1$)。

当 $r = 1$ 时,对于所有的普通节点 $v \in V - S$,它们会受到周围种子节点的影响而转化为接收态,其概率为:

$$I_v^1 = 1 - (1 - p)^{t_v} \tag{1}$$

式中:p 为信息传递概率;Γ_v 为节点 v 的邻居,$t_v = |\Gamma_v \cap S|$ 为节点 v 的种子节点邻居数量。

当 $r \geqslant 2$ 时,信息传递由所有处于接收态的节点进行,这包含两部分节点:初始时刻的种子节点,以及在之前时刻转化为接收态的普通节点。因此,对于一个非种子节点 $v \in V - S$,它在第 r 轮信息传递时处于接收态的概率为:

$$I_v^r = 1 - (1 - I_v^{r-1}) \prod_{u \in \Gamma_v \cap S} (1 - p) \prod_{u \in \Gamma_v - S} (1 - I_u^{r-1} p) \tag{2}$$

式中,$1 - I_v^{r-1}$ 代表节点 v 在 $r - 1$ 时处于未接收态

的概率，$\prod_{u\in\Gamma_v\cap S}(1-p)$ 与 $\prod_{u\in\Gamma_v-S}(1-I_u^{r-1}p)$ 分别代表节点 v 周围的种子节点及处于接收态的非种子节点未能将信息传递给节点 v 的概率。

注意到对于非种子节点 $v=V-S$ 有 $I_v^0=0$，而对于种子节点 $v\in S$ 有 $I_v^r\equiv 1$，可以对式(2)进行进一步简化。

$$I_v^r = 1-(1-I_v^{r-1})\prod_{u\in\Gamma_v}(1-I_u^{r-1}p) \tag{3}$$

综合上述分析，当种子节点集为 S 时，在第 r 轮信息传递中，网络中的某节点 v 处于信息接收态的概率为：

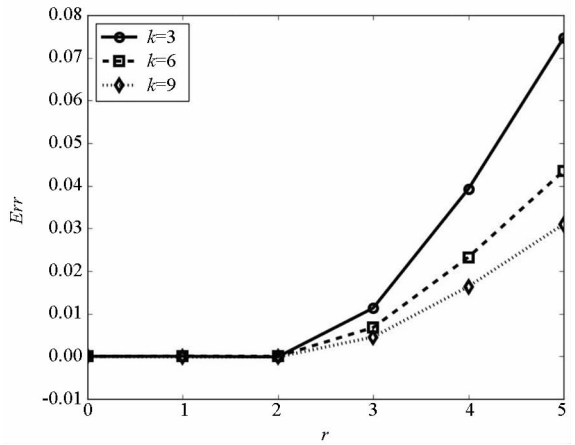
$$I_v^r=\begin{cases}1 & v\in S\\ 0 & v\in V-S, r=0\\ 1-(1-I_v^{r-1})\prod_{u\in\Gamma_v}(1-I_u^{r-1}p) & v\in V-S, r\geqslant 1\end{cases} \tag{4}$$

记 E_S^r 为由种子节点集 S 引起的信息传递在 r 轮所能到达的范围，得到近似估计：

$$E_S^r=\sum_{v\in V}I_v^r \tag{5}$$

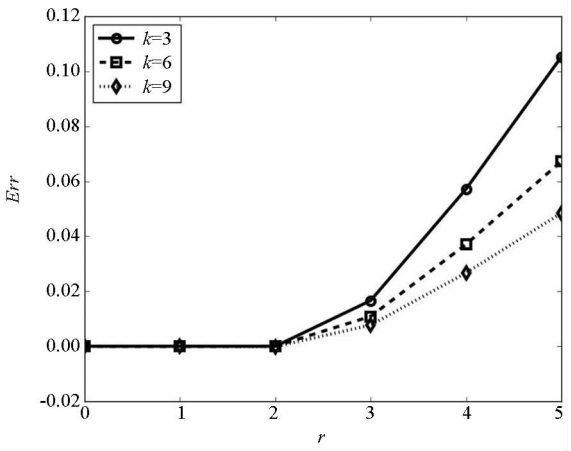
记 $Err_S^r=(E_S^r-T_S^r)/T_S^r$ 为近似估计 E_S^r 与真实传播范围 T_S^r 的相对误差。为了验证上述估计，在人工网络上进行传播验证。由于现实网络中节点的度分布大多为幂律分布 $P(k)\sim k^{-\gamma}$，构造满足幂律分布的配置网络进行实验验证。

图1显示了不同幂指数时的近似估计与真实传播范围的误差。仿真网络的规模均为10 000个节点，平均度分别为3,6,9，初始传播者为随机选取的100个节点，传播概率为 $p=1.1p_c$，其中 p_c 为疾病阈值。图1中的线代表真实传播范围，符号代表近似估计结果。从图1中可以看出，在传播的早期，近似估计结果与真实传播范围符合得很好，随着传播过程的进行，二者的差距逐渐增大，这种趋势在平均度较小的稀疏网络中表现得更为明显。由于影响力极大化问题通常关注的是网络上的早期传播行为，本估计基本可以满足精度方面的需求。



(a) 幂指数为2.1时近似误差图

(a) Errors of the approximations when $\gamma=2.1$



(b) 幂指数为3.0时的近似误差

(b) Errors of the approximations when $\gamma=3.0$

图1 不同幂指数时的近似误差图

Fig. 1 Errors of the approximations with different γ

通过式(5)，可以快速计算种子节点集 S 在整个网络中的集群影响力，特别地，可以用来评估将一个新的节点 v 加入种子节点集所带来的集群影响力增量。

$$FE_v^r=E_{S+\{v\}}^r-E_S^r \tag{6}$$

下面具体分析在信息传递的早期 $r\leqslant 2$ 时节点 v 的影响力增量 FE_v^r 的具体形式。

当 $r=1$ 时，由式(1)可知，非种子节点 $u\notin S$ 转化为接收态的概率 I_u^1 仅与它的种子节点邻居数量 t_u 有关。当感染概率 p 较小时，根据一阶Taylor公式可得 $(1-p)^s=1-sp+o(p)$，从而

$$I_u^1=1-(1-p)^{t_u}\approx 1-(1-t_up)=t_up$$

注意到添加 v 作为种子节点后，邻居节点 $u\in\Gamma_v-S$ 的种子节点邻居数量增加了1，变为了 $t_u'=t_u+1$，此时 u 转化为接收态的概率为：

$$I_u^{1'}\approx(t_u+1)p$$

而对于与节点 v 不相邻的节点 $u\notin\Gamma_v-S$，它们转化为接收态的概率不会发生变化。

综合两方面的分析，添加一个新的节点 v 作为种子节点，其所带来的集群影响力增量 FE_v^1 为：

$$\begin{aligned}FE_v^1&=\sum_{u\in V}(I_u^{1'}-I_u^1)\\&=1-I_v^1+\sum_{u\in V-S-\{v\}}(I_u^{1'}-I_u^1)\\&=1-I_v^1+\sum_{u\in\Gamma_v-S}(I_u^{1'}-I_u^1)\end{aligned}$$

$$\approx 1 - t_v p + \sum_{u \in \Gamma_v - S} p$$
$$= 1 + (d_v - 2t_v)p \quad (7)$$

当 $r = 2$ 时,由式(3)可知,节点 $u \notin S$ 转化为接收态的概率为:

$$I_u^2 = 1 - (1 - I_u^1) \prod_{w \in \Gamma_u} (1 - I_w^1 p)$$
$$= 1 - (1-p)^{2t_u} \prod_{w \in \Gamma_u - S} \{1 - [1 - (1-p)^{t_w}]p\}$$

通过近似和简化,可以得到

$$I_u^2 \approx 2t_u p - t_u(2t_u - 1)p^2 + \sum_{w \in \Gamma_u - S} t_w p^2$$

由上述可知,该近似过程仅需用到节点 u 自身的种子节点邻居数量以及邻居节点的种子节点邻居数量。

将 I_u^2 分解为三项 $I_u^2 = \sum_{k=1}^{3} I_u^2|_k$,其中

$$I_u^2|_1 = 2t_u p$$
$$I_u^2|_2 = -t_u(2t_u - 1)p^2$$
$$I_u^2|_3 = \sum_{w \in \Gamma_u - S} t_w p^2$$

从而添加 v 为种子节点所能带来的影响力增量 FE_v^2 为:

$$FE_v^2 = 1 - I_v^2 + \sum_{u \in V - S - \{v\}} (I_u^{2'} - I_u^2)$$

记 $FE_v^2|_k = \sum_{u \in V-S-\{v\}} (I_u^{2'}|_k - I_u^2|_k)(k = 1,2,3)$,则 $FE_v^2 = 1 - I_v^2 + \sum_{k=1}^{3} FE_v^2|_k$,下面依次进行分析。

对于 $FE_v^2|_1$,仿照前述分析得到:

$$FE_v^2|_1 = \sum_{u \in V-S-\{v\}} (I_u^{2'}|_1 - I_u^2|_1)$$
$$= \sum_{u \in \Gamma_v - S} (I_u^{2'}|_1 - I_u^2|_1)$$
$$= \sum_{u \in \Gamma_v - S} [2(t_u + 1)p - 2t_u p]$$
$$= 2(d_v - t_v)p$$

类似得到 $FE_v^2|_2 = -\sum_{u \in \Gamma_v - S} (4t_u + 1)p^2$。

对于 $FE_v^2|_3$ 的推导则较为复杂。

$$FE_v^2|_3 = \sum_{u \in V-S-\{v\}} \left(\sum_{w \in \Gamma_u - S - \{v\}} t'_w p^2 - \sum_{w \in \Gamma_u - S} t_w p^2 \right)$$
$$= \sum_{u \in V-S-\{v\}} \left[\sum_{w \in \Gamma_u - S} (t'_w - t_w)p^2 - A_{uv} t_v p^2 \right]$$
$$= \sum_{u \in V-S-\{v\}} \sum_{w \in \Gamma_u - S} A_{wv} p^2 - (d_v - t_v) t_v p^2$$
$$= \sum_{u \in \Gamma_v - S} (d_u - t_u - 1)p^2 - (d_v - t_v) t_v p^2$$

综合上述分析并进行合并简化,可以得到 FE_v^2 表达式。

$$FE_v^2 = 1 + 2(d_v - 2t_v)p + (3t_v^2 - d_v t_v - t_v)p^2 + \sum_{u \in \Gamma_v - S} (d_u - 4t_u - 2)p^2 \quad (8)$$

由此提出一种启发式算法,用以近似求解网络中的影响力极大化问题。算法采用序列采样的方式,按照种子集扩展的方式,每次从所有候选节点中选取一个对种子节点集影响力增量最大的节点加入种子集,直到足够数量的种子节点被选出,构成影响力极大化问题中的初始传播者。假定信息传递概率为 p,传递轮数为 r,种子节点数目为 L,如算法 1 所示。

算法 1　影响力极大化快速评估算法

Alg. 1　Influence maximization via fast evaluation

已知:网络 $G(V,E)$,信息传递概率 p,传递轮数 r,种子节点数量 L

1. 初始化种子集 $S = \varnothing$
2. **For** $i = 1 \cdots L$ **do**
3. 　**For all** $v \in V - S$ **do**
4. 　　计算影响力增量 $FE_v^r = E_{S+\{v\}}^r - E_S^r$
5. 　**End For**
6. 　选取最优节点 $v^* = \arg\max_{v \in C} FE_v^r$
7. 　加入初始传播者集合 $S = S + \{v^*\}$
8. **End For**

在传统的基于贪婪思想的影响力极大化算法中,需要通过大量的随机仿真来计算种子节点的集群影响力,这导致算法复杂度激增,即便只是在一个中等规模的网络中寻找极少数量的种子节点也需要花费非常长的时间,难以用来求解现代大规模商业和社交网络上的影响力极大化问题。与传统算法不同,本算法运用近似估计的方式,可以直接计算候选节点的集群影响力增量,加快了影响力极大化问题的求解速度。通常而言,传播轮数越多,快速评估算法的计算复杂度就越高,为了兼顾算法的效率和精度,在后面的实验中,固定参数 $r = 2$。

3　算例分析

3.1　传播模型

当选出种子节点之后,需要通过传播模型来度量这些节点在网络上的传播能力,这里考虑一个更为符合现实中消息传播模式的易感 - 感染 - 恢复(Susceptible-Infected-Recovered, SIR)模型[16]。该模型最早被用来研究传染病在人群中的传播行为,在该模型中,每个节点可以处于三个状态:易感态 S、感染态 I 以及恢复态 R。易感节点指的是那些尚未被感染的个体,可能以概率 p 被周围处于感染态的个体感染。对

于感染节点,它们代表那些感染了疾病的个体,在下一个时间片上,以概率 q 康复,即转为恢复节点。

在经典的 SIR 模型中,每个感染节点可以同时尝试感染所有的易感邻居。然而,在现实生活中,一个更加常见的现象是一个处于感染状态的节点仅仅可以尝试感染一个处于易感状态的邻居(例如握手行为)。因此,这里使用有限感染的 SIR 模型[17]进行仿真。在初始时刻,种子节点被标记为感染状态,其他节点均为易感状态,此后,在每一个时刻,每个感染者以概率 p 随机选取一个邻居节点进行感染,然后以概率 q 恢复。定义有效传播率为感染概率与恢复概率的比值。当网络中没有感染者时,整个传播过程结束,最终感染的节点越多,说明种子节点的影响力越大。

3.2 数据描述

为了验证快速评估算法在影响力极大化问题中的表现,选取 6 个不同类型的真实网络进行验证,分别为:邮件网络[18]——反映了 Enron 公司内部近 50 万封邮件的收发关系;合作网络[19]——隶属于计算机科学的科学家合作网络;购物网络[19]——一个共同购买网络,来自购物网站 Amazon 上的共同购买信息;分享网络[20]——文件分享网站 Gnutella 上的 p2p 网络;信任网络[21]——一个在线评价网站 Epinions 上的用户信任网络;社交网络[22]——在线社交网站 Twitter 上的用户间好友关系网络。上述网络的简要介绍见表 1。

表 1 六个真实网络数据基本性质表

Tab. 1 Brief introduction of six real networks

网络名称	网络类型	节点数	边数	平均度
邮件网络	无向网络	33 696	180 811	10.73
合作网络	无向网络	317 080	1 049 866	6.62
购物网络	无向网络	334 863	925 872	5.53
分享网络	有向网络	8104	26 008	3.21
信任网络	有向网络	75 877	508 836	6.71
社交网络	有向网络	81 306	870 161	10.70

3.3 度量准则

选取以下 3 个度量准则衡量算法的优劣性:种子节点传播范围,种子节点间的平均距离,以及冗余覆盖率。

传播范围是度量种子节点传播能力最直观的度量,定义种子节点 S 的传播范围 F_S 为这些节点所能感染的节点数目。由于传播具有一定的随机性,需要通过多次数值仿真计算感染节点数目的均值。

集合内部节点间的平均距离可以从侧面反映这个集合的结构特性,定义种子节点间的平均距离为:

$$dist(S) = \frac{1}{|S|(|S|-1)}\sum_{u,v\in S} dist(u,v)$$

式中:$|S|$代表种子节点数目;$dist(u,v)$ 代表节点 u 到节点 v 的距离,即从 u 到 v 的最短路径跳数。

为了衡量种子节点传播范围的重叠程度,定义种子节点的冗余覆盖率为:

$$AC(S) = 1 - \frac{F_S}{\sum_{v\in S} F_{\{v\}}}$$

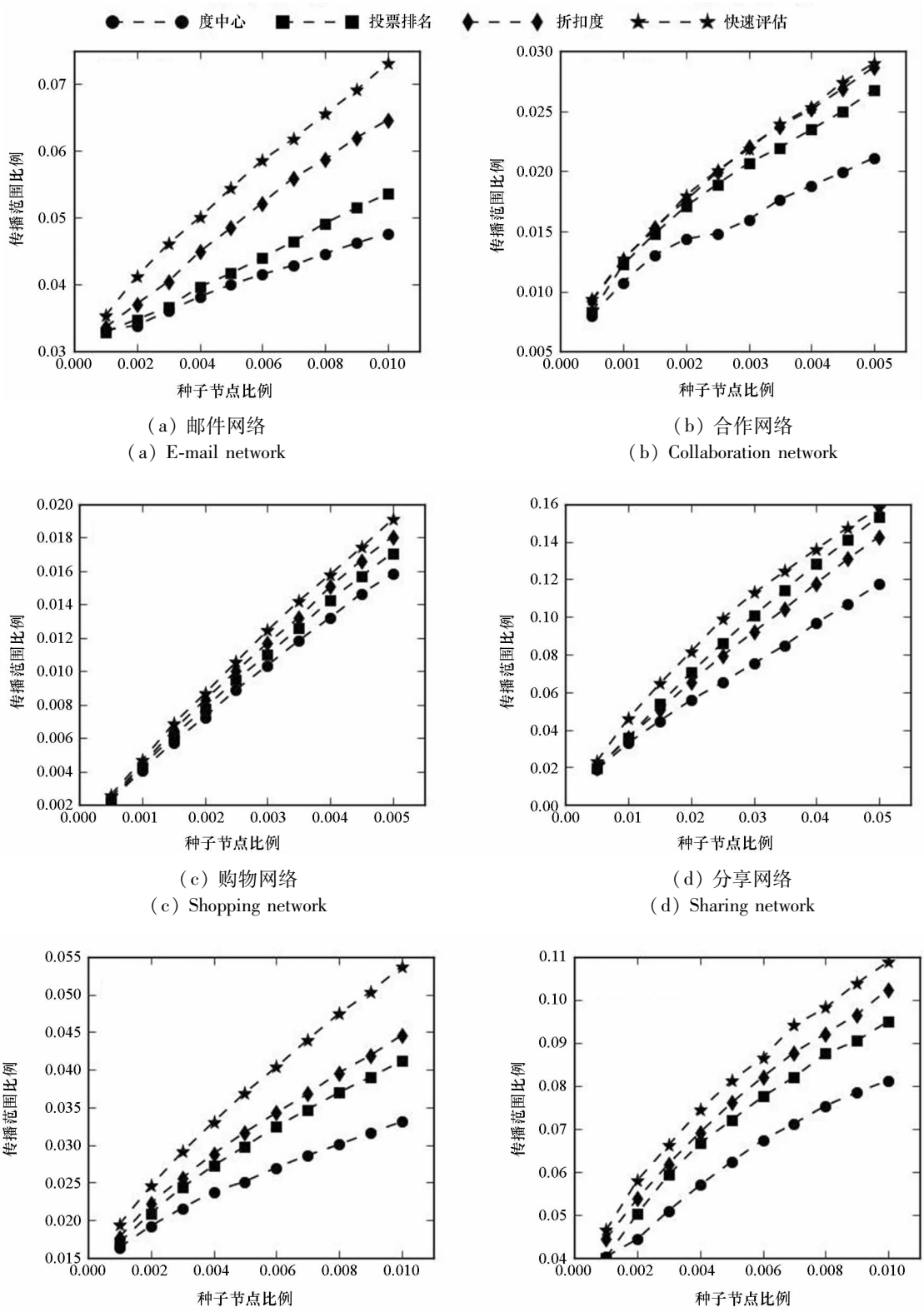
式中,F_S 为以 S 作为初始传播者时所能达到的传播范围,$F_{\{v\}}$ 为仅以节点 v 作为初始传播者的传播范围。

3.4 实验结果

在实验中,设定恢复率 $q = 1/k$,感染率 $p = 1.5q$,其中 k 为网络的平均度。不同种子节点比例时的影响力传播范围如图 2 所示。从图 2 可以看出,快速评估算法的表现一致优于其他三种基准算法,并且随着种子节点数目的增加,快速评估算法的优势越来越明显。在 6 个不同类型的网络中,由快速评估算法选出的种子节点都具有更强的传播能力,体现了该算法的广泛适用性。

种子节点间的平均距离随着种子节点数量的变化如图 3 所示。由图 3 可知,种子节点间的相对距离越近,它们的传播范围相互重叠得越明显,会造成严重冗余,这对于信息传播是不利的。因此,一个好的算法找出的种子节点应当相互分散,即种子节点间的平均距离较高。在该指标下,快速评估算法依旧可以取得较为满意的结果,尤其是在邮件网络、信任网络、社交网络中,快速评估算法明显优于其他基准算法。在分享网络中,度中心找到的种子节点间的平均距离随着种子节点数目的增加呈现先减后增的趋势。事实上,这个结果并不反常,种子节点平均距离与种子节点数目之间并没有确定性的关系,而是与网络结构以及具体算法有关。分享网络是一个高度中心化的网络,由少数紧密相连的中心节点与大量边缘节点构成,呈现出

一种明显的中心 - 边缘趋势。度中心算法会将网络的中心节点作为种子节点,由于中心节点是紧密相连的,因此随着节点数目的增加,种子节点间的平均距离会减少;当中心节点全部找出之后,度中心算法才会加入边缘节点,导致种子节点平均距离的逐渐增加。



(a) 邮件网络
(a) E-mail network

(b) 合作网络
(b) Collaboration network

(c) 购物网络
(c) Shopping network

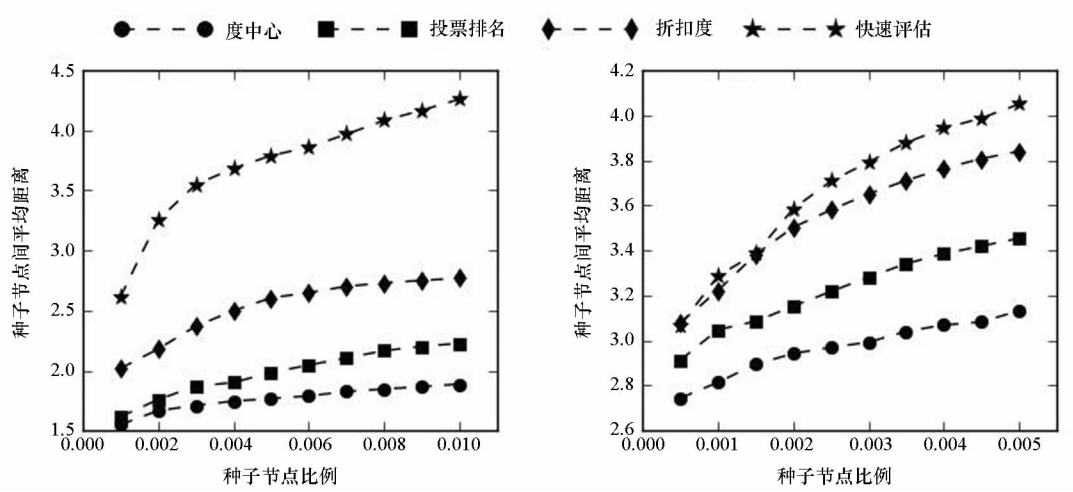
(d) 分享网络
(d) Sharing network

(e) 信任网络
(e) Trust network

(f) 社交网络
(f) Social network

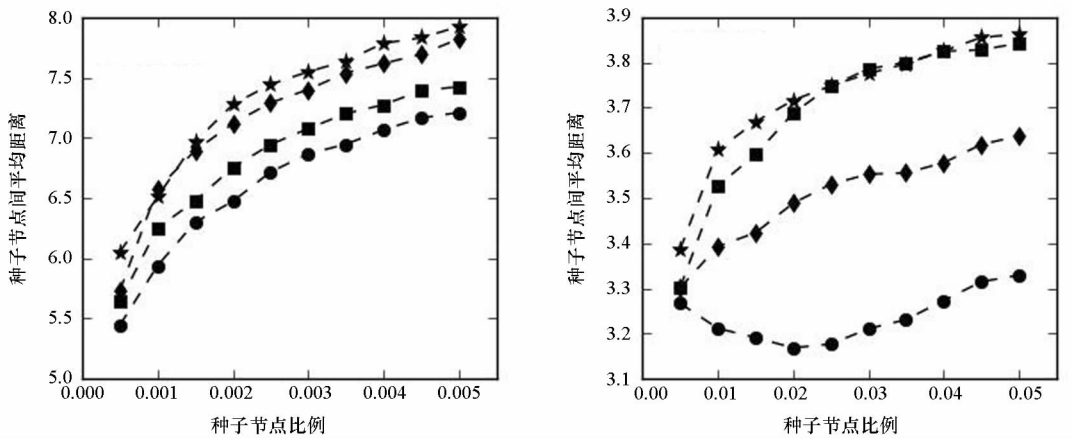
图 2　不同种子节点比例时的影响力传播范围

Fig. 2　Spreading scope with different fractions of seed nodes



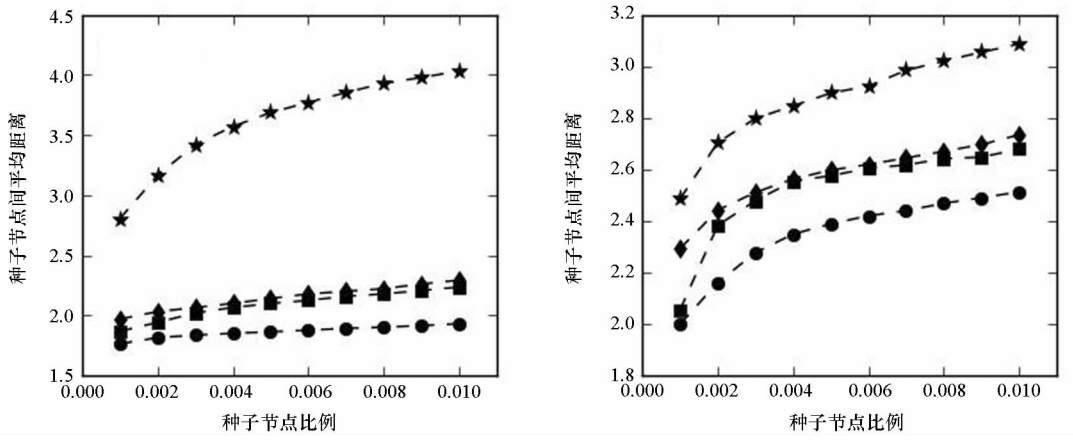
(a) 邮件网络
(a) E-mail network

(b) 合作网络
(b) Collaboration network



(c) 购物网络
(c) Shopping network

(d) 分享网络
(d) Sharing network



(e) 信任网络
(e) Trust network

(f) 社交网络
(f) Social network

图 3　不同种子节点比例时的种子节点平均距离

Fig. 3　Average distance among seed nodes with different fractions of them

种子节点冗余覆盖率是衡量种子节点传播重叠程度的直观指标,冗余覆盖率越大,说明种子节点相互间的传播范围重叠程度越大,它们构成集群时的影响力损耗也越多。各算法在不同网络上的种子节点冗余覆盖率结果见表 2,除了在分享网络上快速评估算法的表现略差于投票排名以外,在其他网络中该算法均能得到最优的结果。

表 2　种子节点冗余覆盖率

Tab. 2　Abundant coverage of seed nodes

网络名称	度中心	投票排名	折扣度	快速评估
邮件网络	0.982	0.977	0.971	0.927
合作网络	0.919	0.860	0.829	0.797
购物网络	0.431	0.362	0.284	0.229
分享网络	0.483	0.247	0.314	0.297
信任网络	0.989	0.985	0.985	0.969
社交网络	0.985	0.979	0.977	0.968

4　结论

本文实现基于种子集扩展的影响力极大化快速评估算法。在 6 个真实网络数据集上进行实证分析,验证了快速评估算法的有效性。而且快速评估算法具有很好的可拓展性。例如,对于节点传播能力异质的情况,可以将式(3)中的传播概率 p 替换为节点传播概率 p_u;当所考虑的网络为时序网络时[23-24],也可以将式(3)中的邻居集 Γ_v 替换为不同时间片 r 时的邻居集 $\Gamma_v(r)$。仿照本算法中的推导过程,可以得到前述情况下的快速评估算法。在下一步的工作中,将对算法进行进一步优化,从图 1 中可以看出,算法对于传播范围的估计精度偏低,还有较大的改进空间。

参考文献(References)

[1] Deb S, Medard M, Choute C, et al. Algebraic gossip: a network coding approach to optimal multiple rumor mongering[J]. IEEE Transactions on Information Theory, 2006, 52(6): 2486 – 2507.

[2] Goh K I, Cusick M E, Valle D, et al. The human disease network[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(21): 8685 – 8690.

[3] Albert R, Albert I, Nakarado G L, et al. Structural vulnerability of the north American power grid[J]. Physical Review E, 2004, 69(2): 025103.

[4] Richardson M, Domingos P M. Mining knowledge-sharing sites for viral marketing[C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002: 61 – 70.

[5] Domingos P, Richardson M. Mining the network value of customers [C]//Proceedings of seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001: 57 – 66.

[6] Kempe D, Kleinberg J M, Tardos E, et al. Maximizing the spread of influence through a social network [C]// Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003: 137 – 146.

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

[8] Goyal A, Lu W, Lakshmanan L V S. CELF + +: optimizing the greedy algorithm for influence maximization in social networks [C]//Proceedings of the 20th International Conference on World Wide Web, 2011: 47 – 48.

[9] Narayanam R, Narahari Y. A shapley value-based approach to discover influential nodes in social networks[J]. IEEE Transactions on Automation Science and Engineering, 2011, 8(1): 130 – 147.

[10] Zhao X Y, Huang B, Tang M, et al. Identifying effective multiple spreaders by coloring complex networks [J]. Europhysics Letters, 2014, 108(6): 68005.

[11] Zhang J X, Chen D B, Dong Q, et al. Identifying a set of influential spreaders in complex networks [J]. Scientific Reports, 2016, 6(1): 27823.

[12] Chen W, Wang Y J, Yang S Y, et al. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009: 199 – 208.

[13] Lü L Y, Zhou T, Zhang Q M, et al. The H-index of a network node and its relation to degree and coreness[J]. Nature Communications, 2016, 7: 10168.

[14] Morone F, Makse H A. Influence maximization in complex networks through optimal percolation [J]. Nature, 2015, 524(7563): 65 – 68.

[15] Newman M. Networks: an introduction [M]. UK: Oxford University Press, 2010.

[16] Hethcote H W. The mathematics of infectious diseases[J]. Siam Review, 2000, 42(4): 599 – 653.

[17] Yang R, Wang B H, Ren J, et al. Epidemic spreading on heterogeneous networks with identical infectivity[J]. Physics Letters A, 2007, 364(3/4): 189 – 193.

[18] Leskovec J, Lang K J, 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.

[19] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth [J]. Knowledge and Information Systems, 2015, 42(1): 181 – 213.

[20] Ripeanu M, Foster I T. Mapping the Gnutella network: macroscopic properties of large-scale peer-to-peer systems[C]//Proceedings of International Workshop on Peer-to-Peer Systems, 2002: 85 – 93.

[21] Richardson M, Agrawal R, Domingos P M, et al. Trust management for the semantic web [C]//Proceedings of International Semantic Web Conference, 2003: 351 – 368.

[22] Leskovec J, McAuley J. Learning to discover social circles in ego networks [C]//Proceedings of Neural Information Processing Systems, 2012: 539 – 547.

[23] Holme P, Saramäki J. Temporal networks [J]. Physics Reports, 2011, 519(3): 97 – 125.

[24] Li A, Cornelius S P, Liu Y Y, et al. The fundamental advantages of temporal networks [J]. Science, 2017, 358(6366): 1042 – 1046.