软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(8):2161−2174 [doi: 10.13328/j.cnki.jos.005120]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种面向团体的影响最大化方法*

张 平[1,2], 王黎维[3], 彭智勇[1,2], 岳 昆[4], 黄 浩[1,2]

[1](软件工程国家重点实验室(武汉大学),湖北 武汉 430072)

[2](武汉大学 计算机学院,湖北 武汉 430072)

[3](武汉大学 国际软件学院,湖北 武汉 430072)

[4](云南大学 信息工程学院,云南 昆明 650091)

通讯作者: 彭智勇, E-mail: peng@whu.edu.cn

摘 要: 影响最大化旨在从给定的社会网络中寻找出一组影响力最大的子集.现有工作大都在假设实体点(个人或博客等)影响关系已知的情况下,关注于分析单个实体点的影响力.然而在一些实际场景中,人们往往更关注区域或人群等这类团体的组合影响力,如户外广告、电视营销、疫情防控等.研究了影响力团体的选择问题:(1) 基于团体的关联发现,建立了团体传播模型 GIC(group independent cascade);(2) 根据 GIC 模型,给出了贪心算法 CGIM (cascade group influence maximization),搜索最具影响力的 top-*k* 团组合.在人工数据和真实数据上,实验验证了该方法的效果和效率.

关键词: 社会网络;影响最大化;关联模型;影响力团体

中图法分类号: TP311

中文引用格式: 张平,王黎维,彭智勇,岳昆,黄浩.一种面向团体的影响最大化方法.软件学报,2017,28(8):2161−2174. http://www.jos.org.cn/1000-9825/5120.htm

英文引用格式: Zhang P, Wang LW, Peng ZY, Yue K, Huang H. Group-Based method for influence maximization. Ruan Jian Xue Bao/Journal of Software, 2017,28(8):2161−2174 (in Chinese). http://www.jos.org.cn/1000-9825/5120.htm

Group-Based Method for Influence Maximization

ZHANG Ping[1,2], WANG Li-Wei[3], PENG Zhi-Yong[1,2], YUE Kun[4], HUANG Hao[1,2]

[1](State Key Laboratory of Software Engineering (Wuhan University), Wuhan 430072, China)

[2](Computer School, Wuhan University, Wuhan 430072, China)

[3](International School of Software, Wuhan University, Wuhan 430072, China)

[4](School of Information Science and Engineering, Yunnan University, Kunming 650091, China)

Abstract: Influence maximization aims at finding a set of influential individuals (i.e. users, blog etc.) in a social network. Most of the existing work focused on the influence of individuals under the hypothesis that the influence relationship between the individuals is known in advance. Nonetheless, it is often the case that groups (i.e. area, crowd etc.) are only natural targets of initial convincing attempts in many real-world scenarios, such as billboards, television marketing and plague prevention. In this paper, the problem of locating the most influential groups in a network is addressed. (1) Based on the discovery of the group associations, GIC (group independent cascade) model is proposed to simulate the influence propagation process at the group granularity. (2) A greedy algorithm called CGIM (cascade

* 基金项目: 国家自然科学基金(61232002, 61502347, 61202033, 61572376); 中央高校基本科研业务费专项资金(2042015kf0038)

Foundation item: National Natural Science Foundation of China (61232002, 61502347, 61202033, 61572376); Fundamental Research Funds for the Central Universities (2042015kf0038)

收稿时间: 2015-11-09; 修改时间: 2016-03-18; 采用时间: 2016-06-10; jos 在线出版时间: 2016-10-11

CNKI 网络优先出版: 2016-10-12 16:26:35, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1626.004.html

group influence maximization) is introduced to determine the top-k influential groups under GIC model. Experimental results on both synthetic and real datasets verify the effectiveness and efficiency of the presented method.

Key words: social network; influence maximization; association model; influential group

近年来,随着互联网的兴起,由此形成的可观测社会网络为研究信息传播、疾病扩散等现象提供了前所未有的平台和机遇.其中,受舆情控制、病毒式营销、疾病预防等应用的驱动,影响最大化问题受到广泛的关注[1−4].该问题旨在从给定的网络中寻找一组有限子集,并根据影响的级联传递,使得该子集的影响力传播最大.目前,现有研究大都以实体点(如个人或博客)为分析对象,并基于点的影响关系,设计算法搜索最具影响力的 k-点组合[3−12].但在很多现实场景下,人们往往更关注团体(如各类人群、社区等)组合的影响力而并非点组合的影响力.

一个关注于团体影响力的典型例子是:埃博拉病毒在非洲蔓延,世界卫生组织有限的资源建立 k 个防疫站.由于防疫站针对的是一个地区的所有居民,为达到全局最佳的防疫效果,世界卫生组织希望找出期望传染力最强的 k 个区域(而不是针对 k 个居民)建立防疫站.在此,区域传染力可抽象为团体的影响力.类似情况还大量出现在户外广告、电视营销等场景中.由此可见,团体影响力的研究具有广泛的现实意义.本文研究了团体影响最大化问题,即给定大小 k,在网络中寻找 k 个团组成的集合(简称为 k-团体组合),使其影响力传播最大.

通常,一个团体的影响力可视为其内所有“感染”(或采纳谣言、购买产品等)点的影响力之和.因此,我们可以依据点的影响力对团体影响力进行估计,并简单扩展现有点影响最大化方法求解团体影响最大化问题.然而,点影响力估计需要非常高的代价,以至于在使用该方法解决由大量点组成的团体影响传播问题时存在明显的效率缺陷.例如,依据传统的传播模型(如独立级联模型[3]),在一个仅仅由 72 万个点组成的网络中寻找 20 个最有影响力的点就需要数天时间[5].方法效率的低下不仅会增加硬件的投入,而且不适用于诸如疾病防疫等具有时效性的应用.此外,点影响最大化方法需要以点影响关系作为输入,但在多数情况下,我们仅能观测到点的状态而不能观测点的相互影响.例如,防疫站可以判断某人是否染病,但并不能确定他受谁传染.综上所述,面向点影响的最大化方法从效率和应用实际上来看,不能很好地解决团体影响最大化问题.

为了避免点分析法的局限性,我们可以将团体看作整体,通过建立团体粒度上传播模型来解决该问题.由于网络中团体的数量远少于点的数量,这种粗粒度的分析策略将显著提高运行效率.然而,团体粒度传播模型的建立面临以下两个方面的挑战.

(1) 首先,团体间的影响本质上是团体间点的影响,而粗粒度上点影响关系的不可见,使得团体间影响存在不确定性.例如,在点粒度上(如图 1(a)所示),我们可以根据点影响关系(图中边)看出,团体 B 中的点$\{b_1,b_2,b_3,b_4\}$受 A 影响,而$\{b_5,b_6\}$受 C 影响,所以 A,C 对 B 的总影响分别为 4 和 2,但在团体粒度上(如图 1(b)所示),点影响关系的遗失使我们不能确定 B 中的“感染”点(如 b_1)是受 A 或受 C 影响,从而难以通过计数计算 A,C 分别对 B 的总影响.

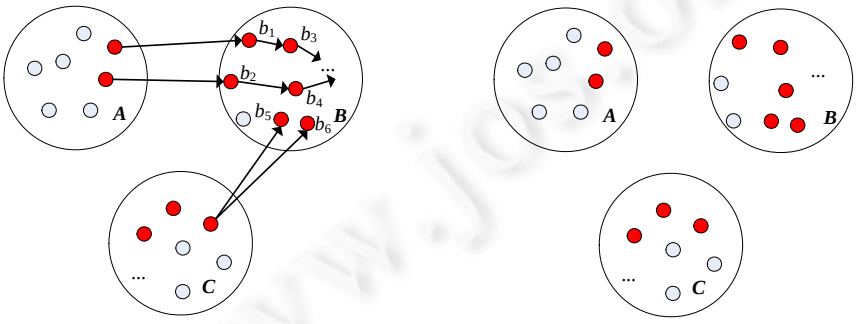


Fig.1 An example of influence relationship of nodes

图 1 点影响关系实例

(2) 其次,作为点的集合,团体可被多个邻居同时影响且状态为连续取值,例如,团体 A 中有 30%的点感染.该特性使得在动态模拟团体影响传递时,需要建立相对于布尔取值的点来说更复杂的规则来计算影响大小.由此可见,如何表达团体的不确定性影响,并描述团体影响传递过程是团体传播模型建立的关键与难点.

为了粗粒度地求解团体影响最大化,以反映团体影响的不确定性为出发点,本文用概率关联的形式描述团体影响的不确定性,并通过对团体历史"感染"数据(见第 3.1 节)进行统计,计算得到团体影响的量值,由此提出了一种基于概率关联的团体传播模型 GIC(group influence cascade).特别指出:由于多个团体影响是一种网状的结构,所以本文定义的概率关联是一种结构化关联.GIC 的核心思想是:将同一团体内点看作"感染"概率相同的随机变量,通过变量集(团体)在历史数据上的条件概率独立描述团体的结构化关联,进而根据关联强弱推测其间不确定性影响,并结合团体"感染"程度动态计算团体影响范围.借助于 GIC 模型,本文给出了 CGIM(cascade group influence maximization)算法快速搜索影响力 top-k 团体组合.该算法利用了 GIC 模型的子模性质,可从理论上保证返回 1–1/e 的最优解.值得指出的是,当我们将一个点看作一个团体时,本文方法同样可以用于解决点影响最大化问题.

本文的主要贡献包括:

(1) 给出了团体影响关联的定义,并用图结构对关联进行建模;

(2) 基于团体关联图,给出了动态计算团体影响力的传播模型 GIC;

(3) 利用 GIC 模型影响函数的单调、子模性,给出贪心算法 CGIM 搜索影响力 k-团组合.

本文第 1 节介绍相关工作.第 2 节定义团体影响关联,并基于关联给出团体传播模型 GIC.第 3 节证明 GIC 模型上影响函数的子模性,并提出 CGIM 算法.第 4 节验证并对比分析 CGIM 算法的效率和效果.第 5 节总结本文工作并探讨未来研究方向.

1 相关工作

与本文相关的工作主要有:点影响最大化方法和实体影响建模.因此,本节将就这两个方面进行简要介绍.

1.1 点影响最大化方法

点影响最大化可定义为:根据特定传播模型,如何从网络中搜寻出一组最具传播力的点集,其大小为 k.该问题被证明是 NP-hard 的.从模型的角度看,IC 模型(independent cascade model,简称 IC)[3]和 LT(linear thresholds)[3]模型是目前研究该问题的两种基础模型.而从算法的角度看,目前主要采用贪心搜索方法[3−7]和启发式搜索方法[8−12]求解该问题.由于本文传播模型的基础是 IC 模型,在此,主要介绍基于 IC 模型的点影响最大化方法.

- 基于贪心搜索的方法

Kempe 等人[3]基于 IC 模型的子模性质,利用贪心算法(又名 KK 算法)来求解该问题.尽管 KK 算法的效果具有理论保证,但由于依赖于大量的蒙特卡洛模拟,该算法的运行十分耗时.此后,Leskovec 等人[4]通过延迟点边际收益计算对 KK 算法进行了优化,提出基于剪枝的 CELF 算法.Wang 等人[5]利用局部搜索降低 KK 算法的开销,并提出了 CGA 算法.研究结果表明:CELF 算法效果好于 KK 算法,但依然存在效率低下的问题[5];CGA 算法依赖的社区划分问题其本身就需要耗费大量计算时间[6],所以它们均无法适用于大规模网络.Borgs 等人[7]给出了一种基于随机抽样的 RIS 算法,并证明了该算法能够在$\Theta(k(n+m)\log n/\varepsilon^3)$的时间复杂度以 $1-\varepsilon$的误差逼近 KK 算法.TIM 算法[8]进一步将时间复杂度降为 $O((k+l)(n+m)\log n/\varepsilon^2)$.虽然 RIS 和 TIM 已逼近线性时间,然而这些算法为记录随机抽样结果需要大量的空间耗费,使其在面对大规模网络时依然存在难以解决的现实问题.作为大量点的集合,求解团体的最大化问题依赖于非常高效的点最大化方法.但就目前的情况来看,现有基于贪心搜索的点最大化方法都难以避免种子集规模增加带来的效率或空间开销问题.因此,使用这些方法求解团体最大化问题并不是一种理想的选择.

- 启发式搜索方法

随着社交网络规模的不断增大,研究者开始关注在大规模环境下影响最大化的计算问题[9−12].这些工作主要采用启发式算法来提高问题求解的效率和可扩展性,但不能在理论上保证准确性.例如,Chen 等人[9]提出了基

于度数的 degreeDiscount 启发算法;Chen 等人[10]根据 IC 模型上的点影响的局部树结构提出了 PMIA 算法; IRIE 算法[11]根据全局影响力排名算法(influence ranking)计算每个点的影响力排名,并基于排序值选择最具影响力点.曹等人[12]将网络中 *K*-core 的作为影响力较大的种子,提出了基于 *K*-core 的最大化算法.相对于基于贪心搜索的方法来说,启发式方法比较简单且具有效率优势,可以在较短时间输出符合其启发规则的团体种子集.但在面对具有整体性质的团体时,面向点的启发式规则是否能够避免团体内点的影响重叠从而准确估计团体影响力,依然是个开放性问题.不仅如此,这些方法大都依赖于点影响关系的获取,所以如何在点影响关系未知的情况下定位最有影响力的团体,仍然是一个亟待解决的问题.

1.2 实体影响建模

目前,一些工作开始关注在实体影响关系不可观测的情况下推测实体影响关系.这些工作主要分为点和点集影响关系建模两个方面.

点影响关系建模的方法[13−16]大都基于 IC 模型,且一般需要点的历史“感染”状态和时间作为输入.它们大多假设一对“感染”间隔小的点其影响可能性大,并通过建立概率生成模型从历史数据中发现点的影响关系.例如,Myers 等人[13]基于点的历史“感染”时间将点的影响关系推测形式化为凸优化问题,并基于生存函数(survival function)提出了 NETRATE 模型.Gomez 等人[15]同样基于点的历史“感染”时间提出期望最大化的 NETINF 模型,用于发现在网络中最强 top-*k* 影响关系.然而,这些基于点“感染”时间间隔的方法难以用于团体影响建模.原因主要有两点.

1) 同一团体中不同点的“感染”时间存在较大差异,以至于团体间的“感染”间隔无法被准确量化;
2) 团体状态取值的连续性,使得概率生成模型难以反映团体状态与影响之间的数量关系,从而给影响推测带来困难.

关于点集影响关系建模的研究有文献[17,18].其中,Mehmood 等人[17]提出的 CSI 模型其实质是点影响关系建模方法的扩展.该模型认为,点集的影响关系即跨集合间点的影响关系.而一个简单的事实是:点集内的点同样会发生影响传递,所以 CSI 模型不能准确描述团体影响关系.Hu 等人[18]提出了 COLD 模型,该模型是一种基于主题分析的团体影响推测方法.然而,COLD 模型所表示的影响不能转化为传统影响最大化问题所依赖的图结构,所以该工作研究的问题实则与本文并不相同.

由第 2.1 节和第 2.2 节的分析可知,现有影响最大化算法和影响建模方法不能很好地用于解决本文研究的问题.由此,本文将从建模和算法两个层面给出适应于团体影响最大化问题的解决办法.

2 团体影响传播建模

本节定义了团体影响关联,并基于关联给出 GIC 传播模型.首先,我们简要介绍方法所依赖的团体历史“感染”数据.

2.1 团体历史“感染”数据

在社会网络中,“疾病”的每次出现引起一次传播过程.我们用 c_l 表示第 l 次“疾病”,并将网络中总共发生的 $|C|$次“疾病”用集合 $C=\{c_1\ldots c_{|C|}\}$表示.当 $c_l\in C$ 传播停止后,网络中团体集 $M=\{m_1,\ldots,m_{|M|}\}$的“感染”程度记为 $S^l=\{S_1^l...S_{|M|}^l\}$,其中, S_i^l 表示团体 m_i“感染”c_l 的比例.直观地,我们可以使用一张$|C|\times|M|$二维表 H 组织整个历史数据,表中 l 行第 i 个元素 $H_{li}=S_i^l$,见表 1 和表 2.

Table 1 Historical records

表 1 历史数据

H	m_1	m_2	…	$M_{\|M\|}$
c_1	13%	9%		15%
c_2	11%	3%		17%
…	…	…	…	…
$c_{\|c\|}$	1%	0%		0%

Table 2 Probability computation of p_l

表 2 p_l 概率计算

D	x_1	x_2	…	x_m	ep
E_1	0	0		0	(1−13%)×(1−9%)×…×(1−5%)
E_2	1	0	…	0	13%×(1−9%)×…×(1−5%)
…	…	…	…	…	…
E_R	1	1		1	13%×…×5%

2.2 IC模型

实体影响力的估计依赖于相应传播模型.IC 模型作为本文模型的基础,是一种点传播模型.在该模型中,点有“未感染”或“感染”两种状态,且点一旦“感染”则状态不再变化.其具体描述如下.

设 S 为种子集,$G(V,E,P)$为社会网络对应的点影响关系图.IC 模型将传播划分为 T 个离散时刻进行模拟:设 A_t 为在 t 时刻的“感染”点集,$A_0=S$.在 $t+1$ 时刻,每个点 $u\in A_t$ 有唯一一次机会尝试以 $p_{u,v}$ 的概率激活 G 中邻居点 $v\notin\bigcup_{i=0}^{t}A_i$,如果激活成功,则有 $A_{t+1}=A_{t+1}\cup v$.重复这一过程,直至 $t=T$ 时刻停止.特别地,当多个“感染”点同时尝试激活同一个点时,激活可以是任意顺序.

2.3 团体影响关联定义

从统计的角度来看,存在影响的团体其间必然存在某种关联,反之亦然.所以团体影响关系对应的关联是一种结构化关联,可用图 $IG(M,I,W)$的形式表示,其中,M,I,W 分别表示团体集、影响关联集和关联程度.本节基于 H 上团体“感染”的概率关系,给出团体影响关联的定义,并用图的形式建模.

在“疾病”c_l 下,网络中点 x_j 的最终状态(是否“感染”)可认为是 c_l 对 x_j 的不确定性影响造成的,则 x_j 可视为二元随机变量,并记 x_j“感染”c_l 的概率为 $p_l(x_j)$,“未感染”c_l 的概率为 $1-p_l(x_j)$.虽然 $p_l(x_j)$的取值无法获得,但将同一团体内的点看作同质时[16](简称为同质性假设),我们可以认为 $p_l(x_j)=H_{li}$,当且仅当 $x_j\in m_i(x)$,其中,$m_i(x)$表示 m_i 对应的点集.例如,在某次埃博拉病暴发时,我们观测到某个区域有 10%的居民感染,那么说明该病有 10%的可能“感染”该区域中的任一个居民.由此,表 1 等价于一张表示点“感染”概率的表,其中,第 l 行第 i 列表示$\forall x_j\in m_i(x)$“感染”c_l的概率.已知点的“感染”概率,我们可以对 c_l下点集状态取值的概率进行计算.设点集 $X=\{x_1,\ldots,x_J\}$的状态取值为 $E_x=(x_1=e_1,\ldots,x_J=e_J)$.那么在 $c_l\in C$ 下,X 以 E_x 出现的概率为

$$p_l(X=E_x)=\prod_{e_j=1}^{J}p_l(x_j)\prod_{e_j=0}^{J}(1-p_l(x_j)) \tag{1}$$

其中,e_j 为二元变量,$e_j=1$ 表示 x_j 的状态为“感染”,$e_j=0$ 表示 x_j 的状态为“未感染”.表 2 给出了一个 $p_l(X=E_x)$的计算实例.根据公式(1),在整个“疾病”集 $C:=\{c_1\ldots c_{|C|}\}$中,X 以状态取值 E_x 出现的概率为

$$p(X=E_x)=\sum_{l=1}^{|C|}p_l(X=E_x)\Big/|C| \tag{2}$$

根据公式(2),我们可以获得同质性假设下 H 上点集状态的完备概率空间 D.由概率论可知,点集对的关联可用点集对的概率独立性进行表达.例如,对于点集 X,Y 的任意状态取值 x,y,都有 $p(x,y)=p(y)p(x)$,则说明 X 和 Y 不存在关联;反之,则存在关联.自然地,作为特定点集,团体对的关联同样可以用概率独立描述.但值得注意的是,概率独立并不具有结构化特征,不能描述团体影响对应的关联图 IG.原因是团体影响有直接和间接之分,而概率独立不能从语义上区别这两种影响所形成的关联.

例 1:设 X,Y 和 Z 为 3 个点集,X 与 Y 存在直接影响记为 $X\sim Y$,且假设有 $X\sim Z\sim Y,X\nsim Y$.因为$\{X,Y,Z\}$中任一团体均可直接或间接地影响其余团体,所以$\{X,Y,Z\}$在其概率空间上的概率独立关系集 $R=\{\varnothing\}$.现在假设影响未知,我们希望根据 R 反推$\{X,Y,Z\}$间的影响.根据 $R=\{\varnothing\}$有,$X\sim Z\sim Y\cup X\nsim Y$ 和 $X\sim Z\sim Y\cup X\sim Y$ 都符合 R 的限定,而 $X\sim Z\sim Y\cup X\sim Y$ 与题设并不相符.由例 1 可以看出,概率独立,不能描述 X 通过 Z 间接关联 Y 的语义,是导致推测错误的关键.为此,我们将引入条件概率独立,以描述团体的直接相关性,从而获得团体影响的结构化关联.首先,我们给出条

件概率独立的定义.

定义 **1**. 设变量集 $U=\{a,b,\ldots\}$的概率分布为 D,且集合 $X,Y,Z\subset U$.X 和 Y 在分布 D 上关于 Z 条件概率独立记作 $I\langle X,Z,Y\rangle$,其中,$I\langle X,Z,Y\rangle$是指对于 X,Y,Z 的任意状态取值 x,y,z,都有 $p(x,y|z)=p(y|z)p(x|z)$.特别地,当 $Z=\varnothing$时,$I\langle X,Z,Y\rangle$退化为概率独立 $I\langle X,Y\rangle$.

定义 1 描述变量间的相对独立性.相对独立性的物理含义反映了排除条件变量(Z)的中介作用后,测试变量(X,Y)的相关性,即反映了测试变量的直接相关性.所以,我们可以通过 $I\langle X,Z,Y\rangle$的真值来判断例 1 中 X 和 Y 是否直接相关来获得其间结构化关联:$I\langle X,Z,Y\rangle$为真,即说明 $X\sim Y$ 为真.由此可见,条件概率独立相对于概率独立具有更丰富的语义,可描述团体影响关联蕴含结构的信息.

根据信息论理论,条件互信息熵常用来定量描述变量的条件独立程度.

定义 **2**. X 和 Y 关于 Z 的条件互信息熵记为 $Inf((X,Y)|Z)$,其中,

$$Inf((X,Y)|Z)=\sum_{X,Y,Z\text{的所有状态取值}}p(x,y,z)\log_2\frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{3}$$

$Inf((X,Y)|Z)$表示 Z 确定之后,Y 与 X 的互信息熵.$Inf((X,Y)|Z)=0$ 表示 $I\langle X,Z,Y\rangle$,且其值越大,说明 X 和 Y 关于 Z 的直接关联性越强.特别指出:当 $Z=\varnothing$时,$Inf((X,Y)|Z)$退化为互信息熵 $Inf(X,Y)$.

根据以上论述,团体直接影响关联的形式化定义如下.

定义 **3**. 设 M 为社会网络 $G(V,E)$上的团体集,团体 $m_i,m_j\in M,m_i$ 对应的点集记为 $m_i(x)$,且有 $V=\bigcup_i^{|M|}m_i(x)$,$m_i(x)\cap m_j(x)=\varnothing$(为简化问题,仅考虑团体不重叠的情况).$m_i$ 和 m_j 存在直接影响关联,当且仅当 $ind(m_i,m_j)>0$.其中,$ind(m_i,m_j)=\begin{cases}w_{ij}, & w_{ij}>\varepsilon\\ 0, & w_{ij}\le\varepsilon\end{cases}$,$w_{ij}=Inf((m_i(x),m_j(x))|(V-(m_i(x),m_j(x))))$,$\varepsilon$为给定阈值.

我们可以根据定义 3 构建 IG.然而,这种直观的构建方法不能运用于较大规模的网络.其原因是:定义 3 中,w_{ij} 的计算需考虑集合 $m_i(x),m_j(x),V-\{m_i(x),m_j(x)\}$的所有状态取值;集合的状态取值数是该集合大小的指数倍,例如,当网络点数量$|V|=n,|m_i(x)|+|m_j(x)|=a$ 时,$V-\{m_i(x),m_j(x)\}$的状态取值就有 2^{n-a} 种,由此可见,w_{ij} 的计算复杂度为 $O(2^n)$.为此,我们将利用同质性假设条件,通过建立团体关联和单点关联的关系,给出一种 w_{ij} 的快速计算方法.在介绍我们的方法之前,我们首先引入条件概率独立相关的数学性质来说明方法的思路.

引理 **1**[19]. 设 X,Y,Z,Q 是分布 D 上两两不相交的变量集,则 $I\langle X,Z,Y\rangle$满足:

(1) 对称律:$I\langle X,Z,Y\rangle\Rightarrow I\langle Y,ZQ,X\rangle$.

(2) 分解律:$I\langle X,Z,Y\rangle\wedge I\langle X,Z,Q\rangle\Leftrightarrow I\langle X,Z,Y\cup Q\rangle$.

在同质性假设下,团体和点的直接影响关联存在如下关系.

定理 **1**. $M=\{m_1,\ldots,m_{|M|}\}$为团体集,$X=\{x_1,\ldots,x_{|M|}\}$为点集,且 $x_i\in m_i(x),i\in 1,\ldots,|M|$.在同质性假设下,在历史观测数据 H 上有 $ind(x_i,x_j)=0$,当且仅当 $ind(m_i,m_j)=0$.

证明:设 $m_i(x)=x_1\cup\ldots\cup x_i$, $m_j(x)=x'_1\cup\ldots\cup x'_j$, $\Upsilon=V-\{m_i(x),m_j(x)\}$.由定义 2 和定义 3 可知,$ind(x_i,x_j)=0$ 当且仅当 $I\langle x_i,\Upsilon,x_j\rangle$,所以原命题即证 $I\langle x_i,\Upsilon,x_j\rangle\Leftrightarrow I\langle m_i(x),\Upsilon,m_j(x)\rangle$成立.显然,在同质性假设下,有 $I\langle x_i,\Upsilon,x_j\rangle\Leftrightarrow I\langle\forall x_i,\Upsilon,\forall x'_j\rangle$ 成立.对 $I\langle\forall x_i,\Upsilon,\forall x'_j\rangle$ 中的$\forall x'$运用分解律有 $I\langle x_i,\Upsilon,x'_1\rangle\wedge\ldots\wedge I\langle x_i,\Upsilon,x'_j\rangle\Leftrightarrow I\langle x_i,\Upsilon,m_j(x)\rangle$ 成立,即 $I\langle m_i,\Upsilon,m_j(x)\rangle$.根据对称律再次运用分解律,同理可证 $I\langle m_i(x),\Upsilon,m_j(x)\rangle$,所以 $I\langle x_i,\Upsilon,x_j\rangle$为 $I\langle m_i(x),\Upsilon,m_j(x)\rangle$的充要条件. □

定理 1 说明点集 X 和团体 M 的关联图同构,所以我们可使用如下的办法快速构建 $IG(M,I,W)$.

- 给定一个以团体为节点(为与社会网络 G 区别,此处将 IG 图中的点称为节点)的完全图 $IG^*(M,I,W)$,如果在分布 D 上有 $Inf(x_i,x_j)\le\varepsilon$(其中,$x_i\in m_i(x)$),说明 m_i,m_j 独立(不存在关联),则直接删去边 $I_{i,j}$;
- 否则,需根据条件概率独立进一步判断关联类型:如果 $ind(x_i,x_j)=0$,则说明在 m_i,m_j 不存在直接关联,删掉该边;如果 $ind(x_i,x_j)>0$,则说明在 m_i,m_j 存在直接关联,将 $I_{i,j}$ 的权值设置为 $w_{i,j}=ind(x_i,x_j)$.

当所有无关联的边被删除后,$IG^*(M,I,W)$即同构于 $IG(M,I,W)$.显然,在该过程中,我们最多需要$|M|(|M|-1)/2$次条件独立计算,且每次计算的复杂性为 $2^{|M|}$.所以在 M 远小于 n 的情况下,该方法可被看作线性时间.算法 1 用伪码的形式清晰地描述了上述过程.

算法 **1**. 团体关联图构建.

输入:团体集 M“感染”的历史观测数据 D.

输出:团体关联图 $IG(M,I,W)$.

1: 初始化图 $IG^*(M,I,W)$为无向完全图;$i\leftarrow 0$

2: **while** $\forall x_i\in m_i$ **and** $m_i.visited$=**false**

3: $m_i.visited\leftarrow$**true**

4: **for each** m_i **and** m_j.visited=**false** //根据条件独立关系删边

If $Inf(x_i,x_j)\geq\varepsilon$

5: **if** $w\leftarrow ind((x_i,x_j)|X-(x_i,x_j))>0$ //$X=\{x_1,\ldots,x_{|M|}\},x_j\in m_j$

6: $w_{ij}\leftarrow w$

7: **else**

8: $I\leftarrow I-e_{ij}$

else

$I\leftarrow I-e_{ij}$

9: i++

10: **return** $IG^*(M,I,W)$

2.4 GIC传递规则

关联程度越高的团体,其间点发生影响传递的可能性越高.根据此特征,本节将基于 IG 给出一种粗粒度的影响传递规则.

给定团体关联图 $IG(M,I,W)$,记团体 $m\in M$ 的邻居集为 $N(m)$.$N(m)$将对 m 产生影响,且如果团体 $u,u'\in N(m)$有 $w_{u,m}>w_{u',m}$,那么点 $x_i\in u(x)$比 $x_j\in u'(x)$有更大可能激活 $x_m\in m$.自然地,我们可以用 $p_{u,m}=\lambda\times w_{u,m}\big/\sum_{k\in N(m)}w_{k,m}$ 来表示 x_i 对 x_m 的影响概率,其中,$\lambda\in[0,1]$为设定激活因子,用以调节影响大小.又由 IC 模型可知,“未感染”的点只受“感染”点影响,且已被“感染”点状态不会发生改变.设 u 的“感染”比例为 η_u,显然,当 $p_{u,m}$ 不变,且 η_u 值越大而 $1-\eta_m$ 值越小时,u 将有更多机会影响 m.所以,$u\rightarrow m$ 的期望传递影响比例 $\eta_{u\rightarrow m}$ 可表示为

$$\eta_{u\rightarrow m}=\eta_u\times p_{u,m}\times(1-\eta_m) \tag{4}$$

特别指出,当 u 中没有点“感染”($\eta_u=0$)或 m 中点已全部被“感染”($1-\eta_m=0$)时,u 对 m 的传递影响比例为 0.

根据公式(4),GIC 模型的动态传递规则如下.

GIC 将影响传播过程划分为 $t(t=0,1,2,\ldots)$个离散时间片进行模拟.为在 G 上激活一次传播,$t=0$ 时刻,我们向种子团体集 S 以广播的方式散布“疾病”,并假设 $m_i\in S$ 以固定比例 R_i“感染”(很多现实场景下,R_i 是可以获得的,例如用户的点击率可被预测).当传播被激活后,S 在 $t>0$ 时刻将迭代的把影响传递给其邻居,乃至其邻居的邻居.IC 模型约定:t 时刻的“感染”点仅在 $t+1$ 时刻具有“传染”性.类似地,在动态模拟团体影响传播过程中,我们将记录团体 m_i 在 t 时刻“感染”的比例 $\eta_i^t(\eta_i^0=R_i)$,并根据 η_i^t 计算团体间的传递影响.设 m_i 的邻居为 $N(m_i)$,我们将分 $|N(m_i)|=1$,$|N(m_i)|>1$ 两种情况给出 η_i^{t+1} 的计算式:当 $N(m_i)=\{u_j\}$时,根据公式(4),直接有 $\eta_i^{t+1}=\eta_{j\rightarrow i}^{t+1}=\eta_j^t\times p_{j,i}\times(1-\eta_i^t)$;当 $N_{in}(m_i)=\{u_j,u_k\}$时,由以上规则可知,u_k 可能重叠激活 m_i 中已被 u_j 部分.为避免重叠激活的部分的累加,我们将 u_k 对 m_i 的影响比例表示为 $\eta_{k\rightarrow i}^{t+1}=\eta_k^t\times p_{k,i}\times(1-\eta_j^t-\eta_{j\rightarrow i}^{t+1})$.类似地,我们可以计算 $N(m_i)$中任意团体对 m_i 的期望影响比例,并对其求和得到 $t+1$ 时刻 m_i 获得的总影响比例 $\eta_i^{t+1}=\sum_{u_j\in N(m_i)}\eta_{j\rightarrow i}^{t+1}$.迭代重复这一过程,直至对任意团体 m_d 都有 $\eta_d^t\times|m_d|<1$(即没有点再被“感染”,其中,$|m_d|$表示 m_d 包含的点数)时停止.特别指出:由于团体激活过程可能成环,为简化问题,我们忽略团体间的回传影响[20],即若在 $t=\alpha$时刻 $u_j\rightarrow m_i$,那么在 $t\geq\alpha$时刻 $u_j\nleftarrow m_i$.

例 2:在图 2 中,设 $S=\{a\}$,$\lambda=1$,且 $\eta_a^0=R_a=30\%$.

- 当 $t=1$ 时,a 将影响邻居 b,c.其中,

$$\eta_b^1=\eta_{a\to b}^1=\eta_a^0\times p_{a,b}\times(1-\eta_b^0)=30\%\times\frac{3}{3+3+4}\times(1-0)=9\%,$$

$$\eta_c^1=\eta_{a\to c}^1=30\%\times\frac{1.5}{1.5+4.5}\times(1-0)=7.5\%.$$

- 在 t=2 时刻,b,c 将共同影响 d,有 $\eta_d^2=\eta_{b\to d}^2+\eta_c^1\times p_{c\to d}\times(1-\eta_d^1-\eta_{b\to d}^2)\approx 7.8\%$.

可以看出,GIC 是 IC 模型的扩展,其主要区别有两点.

1) IC 模型中点的状态为布尔取值,最多能够被 1 个邻居影响,而 GIC 模型将节点(团体)状态扩展为连续取值(比例),能够被多个邻居共同影响.
2) IC 模型中点间的影响的用概率表示,而 GIC 模型中节点间的影响用激活点数的期望比例表示.

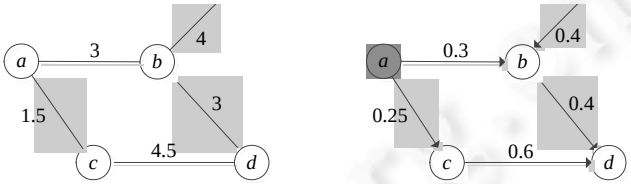


Fig.2 Example of influence diffusion

图 2 影响传播示例

3 团体影响最大化算法

基于 GIC 模型,本节从算法角度定义了团体影响最大化问题,并给出了一种贪心算法快速求解该问题.

3.1 问题定义

给定网络 G 上的团体关联图 $IG(M,I,W)$,整数 k 和团体的初始"感染"率 R_i 作为输入,根据 GIC 模型的模拟传播结果,在 IG 中搜索 k 个团体 $S\subset M$ 作为种子集,使其期望的影响范围$\sigma(S)$最大.

定理 **2**. 团体影响最大化是 NP-hard 问题.

证明:已知点影响最大化问题是 NP-hard.当网络中的每个团体仅包含 1 个点时,团体影响最大化即退化为点影响最大化问题,所以团体影响最大化是 NP-hard 问题. □

由定理 2 可知,在 $P\neq$NP 的假设下,不存在多项式时间的算法能够得到团体最大化问题的精确解.

3.2 函数子模性与贪心选择

定义 **4**(子模性)[21]. F 是定义域为集合 U 的函数,给定 S_1,S_2 为 U 上的 2 个子集.F 具有子模性,如果 F 满足:

$$F(S_1\cup\{u\})-F(S_1)\geq F(S_2\cup\{u\})-F(S_2),$$

其中,$u\subset U$,$S_1\subseteq S_2\subseteq U$.

引理 **2**[21]. 如果一个最优问题的优化目标(函数)F 满足单调性和子模性,那么使用贪心策略求解该问题所获得的结果能够保证返回(1−1/e)的最优.

引理 2 为近似求解特殊优化问题提供了理论支持.本节通过证明团体最大化问题的影响函数$\sigma(S)$满足单调、子模性,给出了一种具有保证的贪心算法.首先,我们基于 GIC 模型给出$\sigma(S)$的函数表达式.

在 GIC 模型中,$\forall m_j\in V/S$ 能够被 S 影响,当且仅当 S 到 m_j(在 IG 中)至少存在 1 条轨(如果路径 $u\to v$ 中的顶点各不相同,则该路径称为 u 到 v 的一条轨).记$\Gamma=\{\Gamma^1,\Gamma^2,\ldots\}$为 S 到 m_j 的轨集,$\Gamma^q=\langle m_i=w_1,w_2,\ldots,w_m=m_j\rangle$表示$\Gamma$中第 q 条轨,其中,$w_1\in S$ 且 $w_u\notin S$,$2\leq u\leq m$.由此,我们可基于轨给出一种特殊的树结构(简称为 Tr)来辅助分析 m_j 受 S 影响的大小,如图 3 所示.

在 Tr 中,m_j 和 S 分别作为树的根和叶子,如果$\Gamma^q\in\Gamma$中 m_j 的直接前驱为 m_k,则 m_k 为 m_j 的儿子节点,依此类推.

显然,在 Tr 中,m_j 只能被其儿子节点 $m_k\in chlid(Tr,j)$直接影响.当不考虑重复激活时,m_k 对 m_j 的期望影响为 $\eta_{k\to j}=\eta_{S\to k}\times p_{k,j}\times(1-0)$.又由于 $chlid(Tr,j)$中团体对 m_j 的影响相互独立,所以我们可以合并所有 $chlid(Tr,j)$对 m_j 的

影响,得到 $\eta_{S\to j}$ 的递推式:

$$\eta_{S\to j}=\begin{cases}R_j, & o_j\in S\\ 1-\prod_{k\in chlid(Tr,j)}(1-\eta_{S\to k}\times p_{k,j}), & o_j\notin S\end{cases} \tag{5}$$

设$|m_j|$表示 m_j 所包含的点数,根据公式(5),影响范围函数$\sigma(S)$可表示为

$$\sigma(S)=\sum_{m_j\in M}|m_j|\times\eta_{S\to j} \tag{6}$$

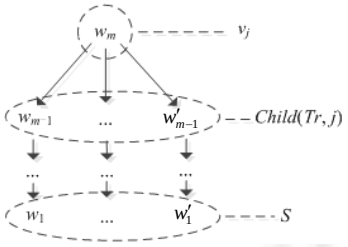


Fig.3 Tr tree

图 3 Tr 树

推论 **1**. 影响函数$\sigma(S)$具有单调性.

证明:给定 $IG(V,E,W)$,$S\subseteq W\subseteq V$.$W-S$ 到$\forall m_j\in M$ 的轨集数量$|\Gamma_{w-s}|\geq 0$,所以有$\Gamma_w\supseteq\Gamma_s$,$chlid(Tr_W,j)\supseteq chlid(Tr_S,j)$.我们首先归纳证明 $\eta_{x\to y}$ 的单调性.当 $m_j\in S$ 时,根据公式(5)直接有 $\eta_{W\to j}=\eta_{S\to j}=R_j$,所以 $\eta_{W\to j}\geq\eta_{S\to j}$ 成立.当 $m_j\notin S$ 时,假设对于$\forall m_k\in chlid(Tr_W,j)$都有 $\eta_{W\to k}\geq\eta_{S\to k}$ 成立.因为 $chlid(Tr_W,j)\supseteq chlid(Tr_S,j)$,对于 $m_j\notin S$ 有:

$$\eta_{W\to j}=1-\prod_{k\in chlid(Tr_W,j)}(1-\eta_{W\to k}\times p_{k,j})\geq 1-\prod_{k\in chlid(Tr_S,j)}(1-\eta_{W\to k}\times p_{k,j}).$$

又已假设 $\eta_{W\to k}\geq\eta_{S\to k}$,有 $\prod_{k\in chlid(Tr_S,j)}(1-\eta_{W\to k}\times p_{k,j})\geq 1-\prod_{k\in chlid(Tr_S,j)}(1-\eta_{S\to k}\times p_{k,j})=\eta_{S\to j}$.所以 $\eta_{W\to j}\geq\eta_{S\to j}$ 成立,即,$\eta_{x\to y}$ 为单调递增函数.又已知单调增函数的和函数依然为单调增函数,$\sigma(S)$的单调性得证. □

推论 **2**. 影响函数$\sigma(S)$具有子模性.

为了证明$\sigma(S,T)$满足子模性,我们首先归纳证明 $\eta_{x\to y}$ 满足子模性,即证:对于任意 $m_i\in M$,有 $\eta_{S'\to j}-\eta_{S\to j}\geq\eta_{W'\to j}-\eta_{W\to j}$ 成立,其中 $S'=S\cup m_i$,$W'=W\cup m_i$,$S\subset W$.当 $m_j\in S$ 时,显然有 $\eta_{S'\to j}-\eta_{S\to j}=\eta_{W'\to j}-\eta_{W\to j}=0$,子模性成立.当 $m_j\notin S$ 时,假设对于 S 的任意前驱 $m_k\in chlid(T_{W'}r,j)$都有 $\eta_{S'\to k}-\eta_{S\to k}\geq\eta_{W'\to k}-\eta_{W\to k}$.根据公式(5)有:

$$\eta_{S'\to j}-\eta_{S\to j}=1-p_{k,j}\times\prod_{k\in chlid(Tr_{S'},j)}((1-\eta_{S'\to k})-(1-\eta_{S\to k}))=p_{k,j}\times\prod_{k\in chlid(Tr_{S'},j)}(\eta_{S'\to k}-\eta_{S\to k}).$$

等式两边取对数得 $\log(\eta_{S'\to j}-\eta_{S\to j})=\log p_{k,j}+\sum_{k\in chlid(Tr_{S'},j)}\log(\eta_{S'\to k}-\eta_{S\to k})$.

又 $\eta_{S'\to k}-\eta_{S\to k}\geq\eta_{W'\to k}-\eta_{W\to k}$,有 $\log(\eta_{S'\to j}-\eta_{S\to j})-\log(\eta_{W'\to j}-\eta_{W\to j})=\sum_{k\in chlid(Tr_{W'},j)}\log\frac{(\eta_{S'\to k}-\eta_{S\to k})}{(\eta_{W'\to k}-\eta_{W\to k})}\geq 0$,即 $\eta_{S'\to j}-\eta_{S\to j}\geq\eta_{W'\to j}-\eta_{W\to j}$ 成立.

如上所述,在假设团体的邻居满足子模性后,该团体同样满足子模性,根据归纳法,有 $\eta_{x\to y}$ 满足子模性成立.又已知子模函数的和函数同样是子模的,所以$\sigma(S)$具有子模性. □

3.3 CGIM算法

根据推论 1 和推论 2,我们将给出一种贪心算法来快速求解团体影响最大化问题,具体描述如算法 2 所示.

算法 **2**. CGIM.

输入:$IG(M,I,W)$, k, $R=\{R_1,\ldots,R_n\}$.

输出: seed set S.

1: $S\leftarrow\varnothing$

2: **while** $|S|<k$ **do**

3: **for each** m_i **in** $M-S$

4: $m_j \leftarrow \arg\max_{m_i \in V-S} \sigma(S \cup m_i) - \sigma(S)$

5: $S \leftarrow S \cup m_j$

6: **return** S

初始时,CGIM 算法将种子集 S 初始化为空集.每一轮迭代中,我们以 $S \cup m_i$ 作为备选种子,通过 GIC 模型估计 $S \cup m_i$ 的影响范围 $\sigma(S \cup m_i)$,并选取边际影响收益 $\sigma(S \cup m_i)-\sigma(S)$ 最大的 m_i 加入 S.重复此过程,直到 S 的大小为 k,则返回.

CGIM 算法每次迭代需计算一次所有备选种子的边际收益,总共 k 轮迭代,一共需要 $O(k \times |M|)$ 次.每次边际收益的计算需要通过一次 GIC 传播模拟过程来确定收益大小.最坏情况下,GIC 传播模拟需要遍历 $IG(M,I,W)$ 中所有点和边,时间复杂度为 $O(|M|+|I|)$.由此,算法 2 的总的时间复杂度为 $O(k \times |M| \times (|M|+|I|))$.由于团体的规模 $|M|$ 远小于社会网络中点的规模 n,所以该方法时间复杂度相对于 n 仍可看作线性时间.

4 实验分析

- 人工数据集

采用 LFR(lancichinetti fortunato radicchi)算法[15]生成的人工网络.特别指出:LFR 算法在生成网络的同时,能够自动给出社区划分基准,而社区是团体一种自然表现形式,属于本文研究对象的范畴.LFR 算法的关键参数说明如下:n 表示模拟网络的点个数,m 表示模拟网络的边数,w 表示社区数量.在本文实验中,我们将通过调节算法参数生成多种规模网络来对 CGIM 算法进行测试,见表 3.

Table 3 Experimental network

表 **3** 实验网络

	n (K)	m (K)	w (K)	*Avg-s*
net_1	10	20	0.4	25
net_2	40	50	0.4	100
net_3	100	400	2	50
dblp	8.345	343.261	0.7	119

- 真实数据集

采用作者合作网络,其中,节点表示作者,边表示两个作者之间存在合作关系.我们从 DBLP 中 2012 年计算机领域的 700 个期刊或会议中抽取了 83 450 个作者、343 261 条合作关系.为获得该合作网络中的团体划分,我们视一个期刊或者会议为一个团体,并将作者划分至投稿次数最高的团体.如,作者 A 向会议(或期刊)L_1,L_2 的投稿次数分别为 3,5,那么 A 将被划分至 L_2.

- 历史观测数据集生成

为了获得历史观测数据,我们将通过以下方式生成.

假定实验网络中点的传播概率 p 相同.在每次“疾病”传播过程中,我们从测试网络中随机的选择 1%的点作为“感染”点,并根据 IC 模型进行影响传播模拟.在传播模拟结束后,我们记录各个团体的“感染”状态作为一条记录,并生成多条记录作为本文实验的观测数据集.

为了验证 CGIM 算法的性能,我们将在点的粒度上使用多种传统方法求解团体影响最大化问题,并以此作为对比基准.实验中,算法的效果好坏可通过各算法输出种子的影响范围(即 $\sigma(S)$,见公式(6))来评价,其中,种子的影响范围越大,说明算法效果越好.特别指出:为了避免不同传播模型对影响范围估计的差异,本文将统一在 IC 模型下对各算法输出种子的影响范围进行计算.由相关工作的分析可知:贪心算法的优势在于选点质量高,而启发算法的优势在于效率,所以本文将选取 CELF 算法、TIM 算法和 degreeDis 算法作为比对算法.

在本文实验中,我们在不同 k 值下对比了各算法的影响范围和运行时间.此外,我们还测试了团体大小、数量对算法的影响.本文程序采用 Java 编写,实验环境为 Quad-Core 2.0 GHz CPU,8GB 内存的个人电脑.

4.1 实验结果

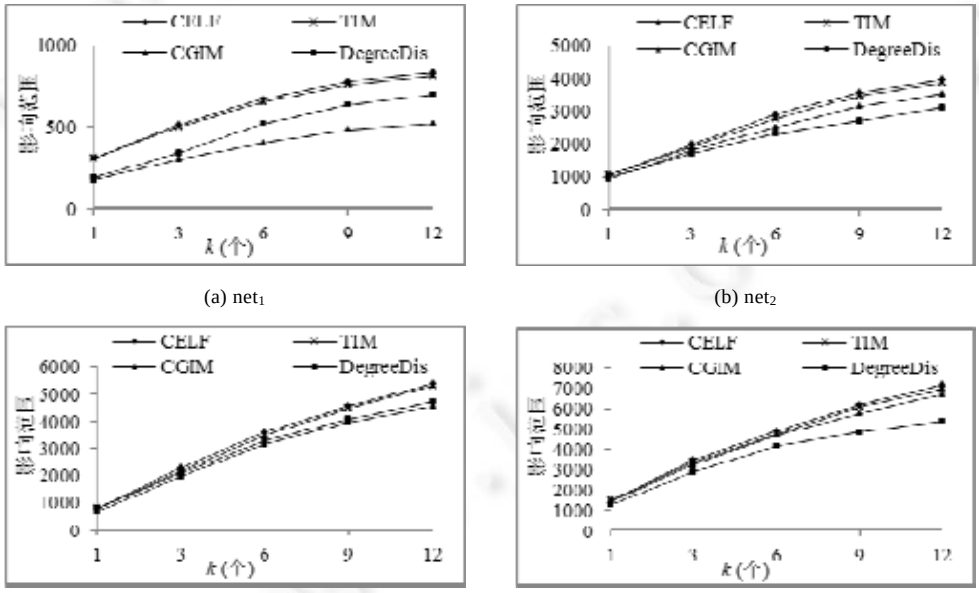
4.1.1 影响范围对比

为了验证网络特征变化对算法效果的影响,本文首先考虑团体规模的因素,并分别在团体平均大小不等的人工网络上验证了各算法输出种子的影响范围.

- 由图 4(a)所示,当团体平均大小 *Avg-s* 为 25 时,CELF,TIM 在 net_1 上的影响范围明显高于其他算法,其次是 DegreeDis,CGIM 的效果较差.
- 当 *Avg-s*=50 时(如图 4(c)所示),CELF,TIM 相对于其他算法的优势逐渐减弱,而 CGIM 与 DegreeDis 的差距被拉近.
- 当 *Avg-s*=100 时(如图 4(b)所示),CELF,TIM 的影响范围依然保持领先,但 CGIM 的表现已明显好于 DegreeDis.

由此可见,团体规模大小对算法效果的影响显著.DegreeDis 的影响范围随团体规模的增大出现了下降的原因是:在规模较大的团体中,度数较高的点之间有很大可能在多跳之后存在较严重的影响重叠,所以度启发规则的选种策略存在单个种子影响力大而总体影响力小的问题.CGIM 的影响范围随团体规模的增大明显增加的原因是:团体中点数越多,单个点的状态变化对整体的影响就越小,即在同质性假设下,点的概率值更贴近团体"感染"比例.所以,基于 GIC 更能反映团体的影响关系,从而间接提高了 CGIM 选点的质量.此外,我们还注意到:图 4(a)中,DegreeDis 的影响曲线在 k=3 和 k=6 之间存在明显跳动.该现象反映了 DegreeDis 算法存在不稳定的问题.CELF 和 CGIM 的曲线随着 k 的增大平缓增加,反映出基于传播模型的算法相对于度启发式算法更具稳定优势.

为了验证 CGIM 算法的实用性,我们在 DBLP 网络上进行对比测试,如图 4(d)所示.图中显示:CGIM 的效果随 k 变化始终较为贴近 CELF,而 DegreeDis 在 $k\geq 6$ 时效果较差.从各算法曲线的对比观察中可以发现:DegreeDis 随着 k 增大,其影响范围增长缓慢,这是该算法效果不佳的主要因素.这是由于 DBLP 网络中存在较为明显的领域性质,例如,同属于数据库领域的 Sigmod 会议和 VlDB 会议往往对其领域内的其他会议和期刊同时造成影响.因为属于同一个领域的团体(会议)间往往存在较大的影响重叠,该特点直接导致了度启发规则的效果不佳.



(a) net_1 (b) net_2

(c) net_3 (d) dblp

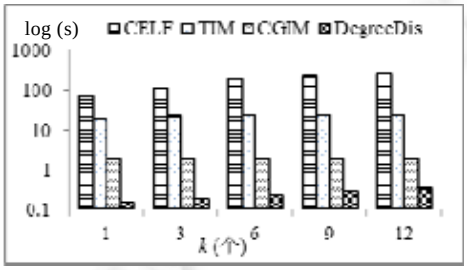
Fig.4 Comparison of influence

图 4 影响力对比

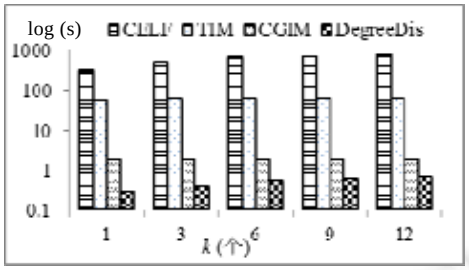
4.1.2 效率对比

图 5(a)~图 5(c)分别显示了在 net_1,net_2,net_3 上各算法的执行时间.特别指出:由于不同算法的执行时间相差较大,我们选择对数刻度描述时间轴.实验结果印证了 CELF 的低效性.如图 5(a)所示,CELF 在小规模网络(net_1)上选取 1 个团体种子的时间即为分钟级别,且随着 k 的增大,执行时间增加明显,效率远远低于 TIM,CGIM 和 DegreeDis;CGIM 的效率明显高于 CELF,TIM 而略低于 DegreeDis.这说明当团体数目远小于点的数量时(net_1:n=10k,w=0.2k),GIC 模型相对于 IC 模型在估计团体边际收益时的效率优势明显.此外,我们还注意到,当 k 增大时,CGIM 的执行时间基本保持不变.这是由于 net_1 中团体数仅为 200,以至于 CGIM 选择局部最优解所消耗的时间相对于扫描历史观测数据来说基本可以忽略引起的.

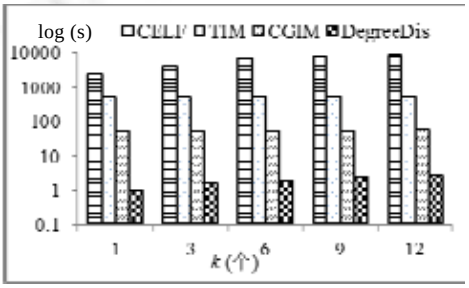
横向对比图 5(a)~图 5(c)我们发现,CGIM 执行时间对网络团体数量规模变化敏感.例如,当团体数同为 w=200 时,CGIM 的执行时间在 net_2 上(如图 5(b)所示)与在 net_1 上(如图 5(a)所示)相比基本无变化,而在 w=2000 的 net_3 上(如图 5(c)所示),CGIM 的执行时间相对于 net_1 上升了 51 倍.由此可见,团体数目对 CGIM 的效率影响较大.为此,我们保持点、边规模不变(n=20k,m=50k),在 w={1k,2k,3k,4k,5k}的网络上进一步验证团体数对 CGIM 效率的影响.由图 5(d)可知,CGIM 在 w={1k,2k,3k,4k,5k}时的执行时间分别为{21,73,179,913,2445}s.可以看出,CGIM 随着团体个数增加,执行时间成指数级上升.造成该问题的原因是 CGIM 依赖于团体关联图的构建,而当团体数量较多时,计算团体间条件概率独立需要大量的时间开销.由此可得出结论:CGIM 算法在团体数目较多的网络上,效率优势不明显,相比之下,更适合处理点数规模较大而团体规模相对较小的网络.



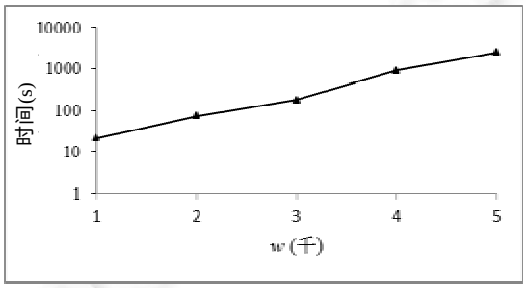
(a) net_1 (log scale)



(b) net_2 (log scale)



(c) net_3 (log scale)



(d) 团体数与 CGIM 效率

Fig.5 Comparison of execution time

图 5 执行时间对比

5 结 论

本文提出并研究了团体影响最大化问题,通过发现历史“感染”数据中团体的概率关联,建立了团体传播模型 GIC,由此给出了一种高效的团体最大化算法 CGIM.与传统的面向点的影响最大化方法不同的是,本文方法不依赖于点影响关系的获取,即可快速定位最有影响力的团体种子集.实验结果表明:当网络中团体数量远小于点数量时,CGIM 算法比 CELF,TIM 算法更高效,且比 degreeDis 算法更准确,适合于处理点数规模较大而团体规

模相对较小的网络.

未来可行的研究方向包括以下 4 个方面.

1) 针对本文团体关联图构建算法(算法 1)不适用于团体数量规模较大网络的问题,我们将研究如何利用概率性质对关联计算进行剪枝,从而提高该算法的效率.
2) 本文工作假设团体之间的点互不重叠,如何在考虑重叠情况下对团体最大化问题进行快速求解,则是我们试图研究的第 2 个问题.
3) 我们还将考虑社会网络的动态性,研究如何在点随时加入、退出团体的情况下求解团体最大化问题.
4) 最后,我们还将试图给出 CGIM 的并行版本,并基于 hadoop,spark 等平台进一步提高算法的可扩展性.

References:

[1] Guille A, Hacid H, Favre C, Zighed DA. Information diffusion in online social networks: A survey. ACM SIGMOD Record, 2013,42(2):17−28. [doi: 10.1145/2503792.2503797]

[2] Lü L, Zhang YC, Yeung CH, Zhou, T. Leaders in social networks, the delicious case. PloS One, 2011,6(6):e21202. [doi: 10.1371/journal.pone.0021202]

[3] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Lise G, Ted ES, Pedro MD, Christos F, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2003. 137−146. [doi: 10.1145/956750.956769]

[4] Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N. Cost-Effective outbreak detection in networks In: Pavel B, Rich C, Xindong W, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. San Jose: ACM Press, 2007. 420−429. [doi: 10.1145/1281192.12812 39]

[5] Wang Y, Cong G, Song G, Xie K. Community-Based greedy algorithm for mining top-*k* influential nodes in mobile social networks. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1039−1048. [doi: 10.1145/1835804.1835935]

[6] Fortunato S. Community detection in graphs. Physics Reports, 2009,486(3-5):75−174. [doi: 10.1016/j.physrep.2009.11.002]

[7] Borgs C, Brautbar M, Chayes J, Lucier B. Maximizing social influence in nearly optimal time. In: Chandra C, ed. Proc. of the Symp. on Discrete Algorithms. Portland: SIAM, 2014. 946−957.

[8] Tang Y, Xiao X, Shi Y. Influence maximization: Near-Optimal time complexity meets practical efficiency. In: Curtis ED, Feifei L, Özsu MT, eds. Proc. of the Int'l Conf. on Management of Data. Snowbird: ACM Press, 2014. 75−86. [doi: 10.1145/2588555.2593670]

[9] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: John FEI, Françoise F, Peter AF, Mohammed JZ, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Paris: ACM Press, 2009. 199−208. [doi: 10.1145/1557019.1557047]

[10] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1029−1038. [doi: 10.1145/1835804.1835934]

[11] Jung K, Heo W, Chen W. Irie: Scalable and robust influence maximization in social networks. In: Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI, Wu XD, eds. Proc. of 2012 IEEE the 12th Int'l Conf. on Data Mining. Brussels: IEEE Computer Society, 2012. 918−923. [doi: 10.1109/ICDM.2012.79]

[12] Cao JX, Dong D, Xu S, Zheng X, Liu B, Luo JZ. A *K*-core based algorithm for influence maximization in social networks. Chinese Journal of Computers, 2015,38(2):238−248 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2015.00238]

[13] Myers S, Leskovec J. On the convexity of latent social network inference. In: John DL, Christopher KIW, John S, Richard SZ, Aron C, eds. Proc. of the Neural Information Processing Systems. Vancouver: Curran Associates, Inc., 2010. 1741−1749.

[14] Gomez-Rodriguez M, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks. In: Lise G, Tobias S, eds. Proc. of the Int'l Conf. on Machine Learning. Washington: Omni Press, 2011. 561−568.

[15] Gomez Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence. In: Bharat R, Balaji K, Andrew T, Qiang Y, eds. Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. Washington: ACM Press, 2010. 1019−1028. [doi: 10.1145/ 2086737.2086741]

[16] Gomez Rodriguez M, Leskovec J, Schölkopf B. Structure and dynamics of information pathways in online media. In: Stefano L, Alessandro P, Paolo F, Aristides G, eds. Proc. of the 6th ACM Int'l Conf. on Web Search and Data Mining. Rome: ACM Press, 2013. 23−32. [doi: 10.1145/2433396.2433402]

[17] Mehmood Y, Barbieri N, Bonchi F, Ukkonen A. CSI: Community-Level Social Influence Analysis. Springer-Verlag, 2013. 48−63. [doi: 10.1007/978-3-642-40991-2_4]

[18] Hu Z, Yao J, Cui B, Xing E. Community level diffusion extraction. In: Timos KS, Susan BD, Zachary GI, eds. Proc. of the Int'l Conf. on Management of Data. Melbourne: ACM Press, 2015. 1555−1569. [doi: 10.1145/2723372.2723737]

[19] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Reasoning. Morgan Kaufmann Publishers, 1988. 84−86.

[20] Roughgarden T, Tardos E, Vazirani VV. Algorithmic Game Theory. Cambridge: Cambridge University Press, 2007. 648−651.

[21] Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 1978,14(1):265−294. [doi: 10.1007/BF01588971]

附中文参考文献:

[12] 曹玖新,董丹,徐顺,郑啸,刘波,罗军舟.一种基于 *k*-核的社会网络影响最大化算法.计算机学报,2015,38(2):238−248. [doi: 10.3724/SP.J.1016.2015.00238]

张平(1984－),男,湖北武汉人,博士生,主要研究领域为社会化网络,Web 数据管理.

岳昆(1980－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为不确定数据管理,不确定性知识发现与推理,数据密集型计算环境下的数据挖掘与知识发现.

王黎维(1981－),女,博士,副教授,CCF 专业会员,主要研究领域为数据质量,数据溯源,科学工作流.

黄浩(1986－),男,博士,副教授,CCF 专业会员,主要研究领域为数据挖掘.

彭智勇(1963－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为复杂数据管理,可信数据管理,Web 数据管理.