doi:10.3969/j.issn.1002-0802.2018.04.033

基于 NVPA 算法的社交网络影响力最大化算法 *

徐 浩[1]，潘 理[2]

（1. 上海交通大学 电子信息与电气工程学院，上海 200240；2. 上海市信息安全综合管理技术研究重点实验室，上海 200240）

摘 要： 衡量与评估用户影响力是在线社交网络分析中的一个经典问题。现有的相关研究主要从个体角度出发，利用贪婪算法进行影响力分析，很少考虑网络中用户一般都会形成社区这样一个客观事实，而一般个体角度的影响力最大化算法都存在运行效率低的问题。因此，提出了一种基于社区的影响力最大化算法 NVPA-IM（Neighborhood Vector Propagation Algorithm Influence Maximization）。通过与经典影响力最大化算法的对比分析，证明了所提算法在保证算法精度的同时，显著提高了算法效率。

关键词： 社交网络；社区发现；节点影响力评估；影响力最大化

中图分类号： TP393 **文献标志码：** A **文章编号：** 1002-0802(2018)-04-0924-06

Social Network Influence Maximization Algorithm based on NVPA Algorithm

XU Hao[1], PAN Li[2]

(1. School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai 200240, China; 2.Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai 200240, China)

Abstract: Measuring and evaluating user influence is a classic issue in online social network analysis. The existing related research mainly uses greedy algorithm to analyze the influence from individual perspective, and seldom considers the fact that users in the network generally form the community. However, in general, the algorithm of maximizing influence based on individual perspectives has the problem of low operational efficiency. Therefore, a community-based influence maximization algorithm NVPA-IM (Neighborhood Vector Propagation Algorithm Influence Maximization) is proposed. Through comparison with the classical maximization algorithm, it is proved that the proposed algorithm can improve the algorithm efficiency while ensuring the accuracy of the algorithm.

Key words: social network; community detection; node influence evaluation; influence maximization

0 引 言

随着在线社交网络分析在推荐系统、市场营销、信息检索等依据影响力对用户行为进行预测的领域的广泛应用，影响力分析问题在学术界和工业界引起了越来越多重视。基于社交网络的市场营销，如何准确找到最具影响力的个体集合、获取最大的影响范围，是极其关键的问题。对此，文献 [1] 给出了准确化定义，将影响力最大化问题转化为如何选择 K 个初始节点，通过激活这 K 个初始节点，在给定信息传播模型下，使网络中最终被激活的用户节点数最多。已有的相关研究主要从个体的角度分析问题[2-4]，Kempe 等人[2] 首先使用贪婪爬山算法进

* **收稿日期：** 2017-12-20；**修回日期：** 2018-03-17 Received date:2017-12-20;Revised date:2018-03-17
基金项目： 国家自然科学基金“面向网络舆情监管的目标群体行为特征检测与分析技术”（No.U1636105）
Foundation Item: National Natural Science Foundation of China “Targeted Group Behavior Characteristics Detection and Analysis Technology for Internet Public Opinion Supervision” (No.U1636105)
通讯联系人： panli@sjtu.edu.cn Corresponding author:panli@sjtu.edu.cn

行影响力分析，证明了这一问题是 NP 难问题，用贪婪算法求解可以达到 (1−1/e) 的精度。但是，贪婪算法的时间复杂度太高。后来研究人员主要解决的问题是提高算法的运行效率 [5, 6]，但是这些算法都是基于网络全局进行算法模型设计。当面对大规模社交网络时，过高的时间复杂度依旧是一个棘手的难题。因此，本文基于 NVPA[7] 社区划分算法，提出了一种从局部角度解决影响力最大化问题的算法 NVPA−IM。在传统的社区划分算法中，相似度衡量指标一般都是节点的一跳邻居节点数的重合度，而忽略了多跳邻居节点集对相似度的影响。相对于传统算法，NVPA 算法在计算节点之间的相似度时，考虑了节点的多跳邻居节点对相似度的贡献，显著提高了社区划分的精度。同时，传统的社区划分算法一般将模块度指标 [8] 作为衡量社区划分好坏的重要指标，以模块度最优作为社区划分过程迭代终止的判断条件。所以，最终网络中社区的个数并不能确定。相对于传统社区划分算法，NVPA 算法在保证模块度指标优异的情况下，可以将网络指定划分成 K 个社区。经实验验证，该算法能够精确挖掘网络中的社区结构。

1　NVPA−IM 算法

在社交网络研究中，一般将社交网络抽象成一张有向(或无向)图 $G=(V,E)$。其中，用户抽象成节点，$V=\{v_1,v_2,\cdots,v_n\}$ 为节点集合。用户之间的关系抽象成边，$E=\{(v_i,v_j)|v_i,v_j \in V\}$ 为边的集合。

定义 1　影响力最大化。给定一个社交网络 $G=(V,E)$ 和一个正整数 K，要求从社交网络中寻找 K 个初始活跃节点集合 $S(|S|=K,S\subseteq V)$，S 在社交网络中按照某种信息传播规律传播影响，要求最终活跃的节点数目最多。用影响力函数 $\sigma(\cdot)$ 表示社交网络中最终活跃节点的数目。影响力最大化问题即要求寻找合适的集合 S 使得 $\sigma(\cdot)$ 最大化，一般将集合 S 称为种子集，形式化描述如下：

$$\begin{aligned} &Max\sigma(S)\\ &s.t.S\subseteq V\\ &|S|=K \end{aligned} \tag{1}$$

1.1　NVPA 算法

NVPA 算法是一种基于社交网络结构而进行社区划分的算法。如上介绍，为了更精确地挖掘网络的社区结构，在计算节点之间相似度时，NVPA 算法考虑了与节点距离为 $d(d>1)$ 的邻居节点对结果的影响。该算法定义节点的 m 跳邻居节点集如下：

$$N_m(v)=\{u\in V\mid d(u,v)=m,v\in V\} \tag{2}$$

式中 $d(u,v)$ 是节点 u 到节点 v 的距离。同时，该算法定义邻域向量 $\boldsymbol{Z}_i$ 用于保存节点 i 的拓扑结构信息。$\boldsymbol{Z}_i$ 是归一化向量，能够有效表示节点在空间中所处的位置。在一个具有 n 个节点的网络图中，节点 v_i 的邻域向量定义为：

$$\hat{\boldsymbol{Z}}_i=(\hat{\boldsymbol{Z}}_{i1},\hat{\boldsymbol{Z}}_{i2},\cdots,\hat{\boldsymbol{Z}}_{in}),|\hat{\boldsymbol{Z}}_i|=1 \tag{3}$$

在 n 维欧式空间中，$\hat{\boldsymbol{Z}}_i$ 是一个单位向量，反映了节点和其邻居节点之间的关系。基于以上两个定义，NVPA 算法设计了一个邻域向量传播规则：

$$\hat{\boldsymbol{Z}}(i,0)=\boldsymbol{\alpha}_i,i=1,2,\cdots,n \tag{4}$$

$$\boldsymbol{Z}(i,t)=\hat{\boldsymbol{Z}}(i,t-1)+\frac{1}{d_i}\sum_{v_j\in N_1(v_i)}\hat{\boldsymbol{Z}}(j,t-1) \tag{5}$$

$$\hat{\boldsymbol{Z}}(i,t)=\frac{\boldsymbol{Z}(i,t)}{|\boldsymbol{Z}(i,t)|} \tag{6}$$

$$sim_m(v_i,v_j)=<\hat{\boldsymbol{Z}}_i,\hat{\boldsymbol{Z}}_j> \tag{7}$$

式（4）对节点的邻域向量进行初始化，$\{\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_n\}$ 是欧式空间 $\boldsymbol{R}^n$ 的正交基；式（5）中 d_i 表示节点的度。依据式（5）、式（6），循环更新节点的邻域向量值 m 次，则每个节点的邻域向量 $\boldsymbol{Z}_i$ 都包含了节点的 m 跳及 m 跳以内的邻居节点的网络拓扑结构信息。基于节点的邻域向量计算相邻节点的 cosine 相似度即式（7），将相似度最大的两个节点 v_i、v_j 组合成一个社区，用节点 v_i 表示该新社区，同时将节点 v_j 从网络中删除。最后，按照一定的降维规则将维度 $\boldsymbol{\alpha}_j$ 删除。循环以上过程，直至网络中只包含 K 个节点时，社区划分结束，这 K 个节点即是获取的 K 个社区。

1.2　种子节点选取

利用 NVPA 算法划分后的网络包含 K 个节点，每个节点代表一个社区。从每个社区中挖掘出一个影响力最大的种子节点，即组成种子集 S。通常，启发式算法在选取种子节点时效率较高，如传统的度中心算法。该算法认为，社区中度最大的节点即一跳邻居节点数最多的节点，在传播过程中具有重要作用。由 NVPA 算法可知，节点在网络结构中的重要性也受节点的 m 跳及 m 跳以内邻居节点数的影响。所以，本文设计了基于式（8）的种子节点选取算法 NVPA−IM：

$$\Gamma_m(v)=\sum_{i=0}^{m}|N_m(v)| \tag{8}$$

该算法基于 NVPA 算法的性质，对度中心算法进行扩展。在 NVPA 社区划分算法中，计算相邻节点之间的相似度时考虑到了距节点距离小于等于 m 的所有邻居节点对节点之间相似度的贡献，即相对于传统的相似度算法，NVPA 算法在计算相邻节点相似度时不仅仅只考虑一阶邻居节点数目。所以在选取社区中重要的节点时，本文将度中心算法的衡量指标一跳邻居节点数扩展到 m 跳及 m 跳以内的邻居节点数，m 的取值由社区划分阶段邻域向量的传播次数决定。在式（8）中，$N_i(v)$ 是节点 v 的 i 跳邻居节点集合包含的节点数。$\Gamma_m(v)$ 值越大，说明以节点 v 为中心，到节点 v 距离为 m 及 m 以内的节点集覆盖的范围越广，节点 v 在影响力传播方面的作用越大。在社区划分阶段，一般情况下设置 m 为 2 ~ 4 的整数。在本文实验阶段，设置 m 的值为 3。综上所述，NVPA-IM 算法的伪代码如下：

输入：网络 $G=V \in E$；邻域向量传播次数 m；种子节点数 K

输出：包含 K 个节点的种子集 S

1 S ←∅ ;// 将 S 初始化为空集

// 第一部分：社区划分

2 for each v_i in V, do:

3 根据式（4）初始化节点 v_i 的邻域向量

4 for t=1 to m,

5 根据式（5）和式（6）更新网络中每一个节点的邻域向量

6 for each e_{ij} in E, do:

7 $Sim(v_i,v_j) \leftarrow <\hat{\boldsymbol{Z}}_i,\hat{\boldsymbol{Z}}_j>$

8 for l=1 to $n-K$

9 选择相似度最大的两个节点 v_i,v_j，将它们组合成一个新社区，用 v_i 表示该社区

// 空间降维

10 $\boldsymbol{Z}_i=n_i\hat{\boldsymbol{Z}}_i+n_j\hat{\boldsymbol{Z}}_j$

11 for each $\hat{\boldsymbol{Z}}_s$,if($\boldsymbol{Z}_{sj} > 0$)

12 $\boldsymbol{Z}_{si}=\boldsymbol{Z}_{si}+\boldsymbol{Z}_{sj},\boldsymbol{Z}_{sj}=0$

13 删除维度 $\boldsymbol{\alpha}_j$ 并且归一化更新后的邻域向量

// 第二部分：种子节点选取

14 for each C_i

15 for each v in C_i

16 $\Gamma_m(v)=\sum_{i=0}^{m}|N_m(v)|$

17 在社区 C_i 中选择 $\Gamma_m(v)$ 值最大的节点 v

18 $S=S \cup v$

19 return S

可见，NVPA-IM 算法首先利用 NVPA 算法对社交网络进行社区划分（lines1 ~ lines13）。该阶段中，网络中每一个节点都代表一个社区。因此，line9 中是将社区 v_j 组合到社区 v_i 中，且 line10 中的 n_i、n_j 分别表示社区 v_i、v_j 中用户的个数。然后，该算法从每个社区中进行种子节点选取（lines14 ~ lines18），用社区 $C_1,C_2,\cdots,C_k$ 表示第一部分社区划分的结果 $v_1,v_2,\cdots,v_k$，并依据 $\Gamma_m(v)$ 指标挖掘种子节点集。

2 实验与结果分析

本文实验使用 Mcauley[9] 于 2012 年从 Facebook 上获取的数据集和文献 [10] 中提到的 LRF 基准网络数据集。LRF 数据集是人工数据集。在生成该数据集的过程中，同时考虑了社区内部的紧密连接特征和现实世界复杂网络的真实特征。Facebook 数据集包含 4 039 个节点和 88 234 条边；LRF 人工数据集包含 1 000 个节点和 9 935 条边。本文实验的实验环境是 Intel Corei7-6700 3.4G 处理器，16 GB 内存。

为了便于对比分析，在实验阶段主要用影响收益来评估算法的性能。影响收益是指在社交网络中选取 K 个初始节点并激活，在影响传播结束后，网络中被激活的节点数 w 与网络中节点总数 n 的比值 w/n。该值的取值范围是 0 ~ 1。算法的影响收益越大，说明该算法的种子节点集选取的越准确。通过激活该种子节点集合获取的信息，传播范围越广。

本文实验内容主要包括两部分：（1）从影响收益角度验证 NVPA 社区划分算法的性能；（2）通过与已有的种子挖掘算法对比，验证 NVPA-IM 算法的精度与效率。算法精度主要用影响收益衡量，而算法效率主要用实验运行时间衡量。

首先，为了评估 NVPA 社区划分算法在解决影响力最大化问题时的性能，本文将其和目前已有的几种典型的社区划分算法进行对比，即基于模块度最大化的贪心算法 FN[11]、基于图的半监督学习算法标签传播算法 LPA[12] 和基于相似度的聚合算法 H-Clustering[13]。在实验结果分析阶段，本文先利用以上 4 种社区划分算法将网络进行社区划分，然后依据社区规模大小选择前 K 个社区，随机从每个社区中选择一个节点组成种子节点集，最后在独立级联传播模型（IC）上对比分析种子节点集的影响收益。在该阶段，种子节点数目 K 的取值范围是 1 ~ 15。同时，下文分别用 NVPA-R、FN-R、

H-Clustering-R、LPA-R 表示以上 4 种种子节点选取算法。

在 IC 模型中，本文设置每条边的传播概率为 0.05。在 LRF 数据集上，4 种社区划分算法的影响收益如图 1 所示。

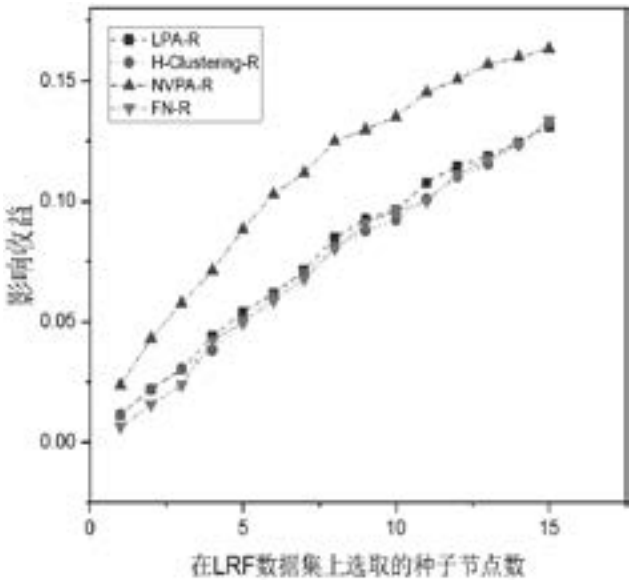

图 1　不同社区划分算法在 LRF 数据集上的影响收益对比

如图 1 所示，当种子节点数目 K 属于 1 ~ 15 时，FN-R、LPA-R、H-Clustering-R 这 3 种算法的影响收益近似，而 NVPA-R 算法的影响收益要明显优于其余 3 种算法。在 Facebook 数据集上，4 种社区划分算法的影响收益如图 2 所示。

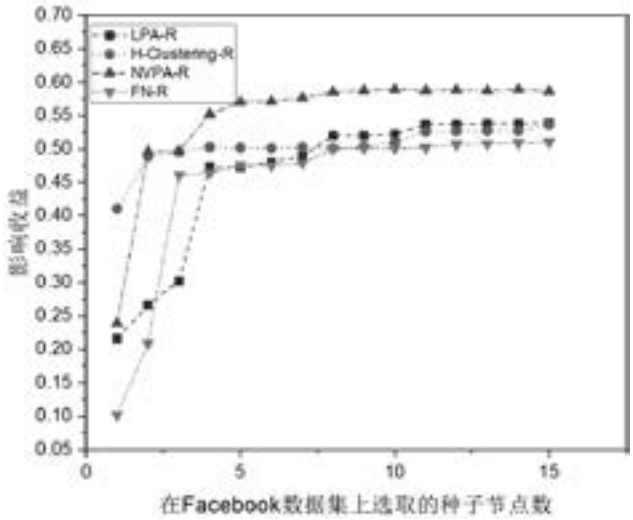

图 2　不同社区划分算法在 Facebook 数据集上的影响收益对比

如图 2 所示，随着 K 值的增加，FN-R、LPA-R、H-Clustering-R 及 NVPA-R 这 4 种算法的影响收益都趋于稳定，且前 3 种算法的影响收益要明显小于 NVPA-IM 算法。由以上实验结果可知，对于从社区角度解决影响力最大化问题，本文选取的 NVPA 社区划分算法的性能要明显优于已知的 3 种从不同角度进行社区划分的典型算法。

为了验证上文提出的 NVPA-IM 算法的性能，本文将其与从社区中选取度最大节点的度中心算法（NVPA-DEG）、从社区中随机选取种子节点的算法（NVPA-R）以及从个体角度解决影响力最大化问题的贪婪算法（CELF++）进行对比分析。因为文献 [2] 中已经证明贪婪算法能够保证精度达到最优解的 (1-1/e)，而 CELF++ 算法 [5] 是优化后的贪婪算法。相对于传统的贪婪算法，它将算法的运行效率提高了 35% ~ 55%。所以，在贪婪算法方面，本文选择 CELF++ 算法，并将其影响收益作为算法精度对比分析的基准。

在 LRF 数据集上，基于 IC 模型，4 种算法的影响收益对比结果如图 3 所示。

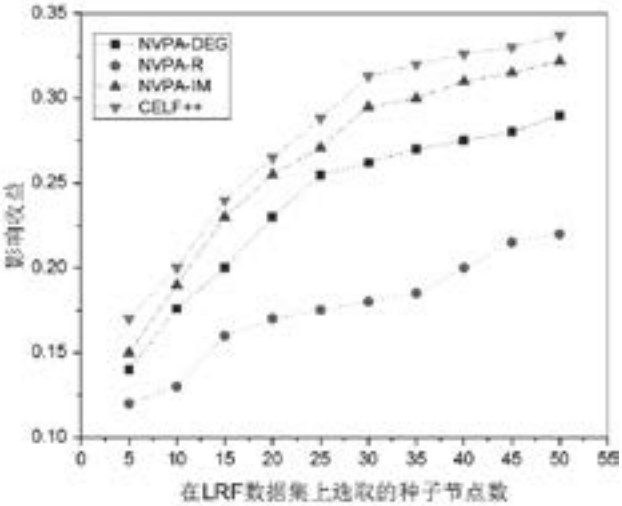

图 3　不同影响力最大化算法在 LRF 数据集上的影响收益对比

如图 3 所示，NVPA-R 算法获得的影响收益最差，CELF++ 算法获取的影响收益最好。相对于 NVPA-DEG 算法，NVPA-IM 算法获取的影响收益更接近于 CELF++ 算法，即 NVPA-IM 算法能够更准确地挖掘影响力最大的种子节点。在 Facebook 数据集上，4 种算法的影响收益对比结果如图 4 所示。

如图 4 所示，在 Facebook 数据集上，以 CELF++ 算法获取的影响收益为基准，NVPA-IM 算法的影响收益要明显优于 NVPA-DEG 算法和 NVPA-R 算法。该算法只有极小的精度损失，这与在 LRF 数据集上 NVPA-IM 算法的表现相同。由以上分析可知，本文提出的种子节点选取算法相对于已有算法显著提高了算法精度。

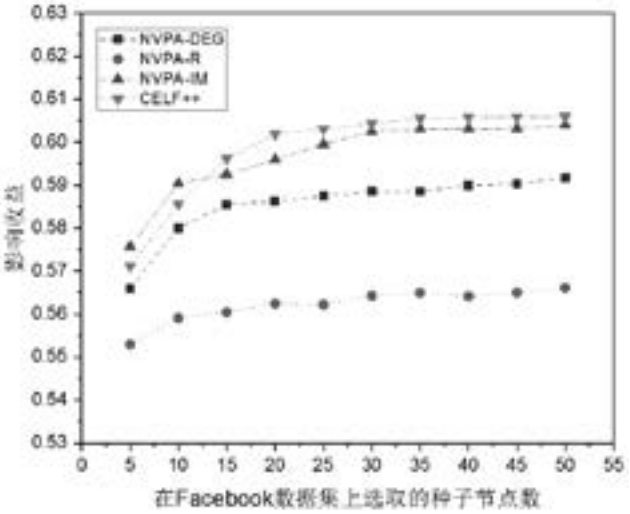

图 4 不同影响力最大化算法在 Facebook 数据集上的影响收益对比

在验证算法效率阶段，本文将其和 CELF++ 算法进行对比分析。为了保证算法的精度，CELF++ 算法为每个节点计算边际收益时，均需进行 R 次蒙特卡罗模拟，R 一般取值为 20 000。蒙特卡罗模拟的时间复杂度与网络中的边的数量成线性关系。所以，当网络的平均度较大、连边数量较多时，CELF++ 算法一般效率较低。而 NVPA-IM 算法主要的时间开销在社区划分阶段，由文献 [7] 实验证明，NVPA 社区划分算法的效率较高。在本文实验中，基于 IC 传播模型，在 LRF 数据集上，当挖掘 50 个种子节点时，NVPA-IM 算法只需要 55 s，而 CELF++ 算法需要 729 s。可见，前一种算法用时只占到 CELF++ 算法的 7.5%。

3 结 语

本文提出了一种从社区层面解决影响力最大化问题的算法。基于社交网络的结构，本文首先使用 NVPA 算法对网络进行社区划分。考虑到划分的过程，节点之间的相似度受节点的 m 跳邻居节点影响。所以，本文设计了一个基于节点 m 跳及 m 跳以内的种子节点选取算法，从划分后的每个社区中挖掘出一个种子节点组成种子节点集。通过实验验证分析，在解决影响力最大化问题方面，与传统贪心算法相比，NVPA-IM 算法以一定的算法精度为代价，显著提高了算法的运行效率。当然，该算法也存在一定不足，后续研究中，在影响力挖掘方面应该更加全面地结合网络和用户的各种属性特征进一步精确影响效果，贴近真实。

参考文献：

[1] Richardson M,Domingos P.Mining Knowledge-Sharing Sites for Viral Marketing[C].Proceedings of the Eighth ACM SIGKDD International Conference On Knowledge Discovery And Data Mining ACM,2002:61-70.

[2] 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.

[3] Kempe D,Kleinberg J M,Tardos É.Maximizing the Spread of Influence through a Social Network[J].Theory of Computing,2015,11(04):105-147.

[4] Morone F,Makse H A.Influence Maximization Incomplex Networks Through Optimal Percolation[J].Nature,2015,524(7563):65-68.

[5] 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 Companion On World Wide Web ACM,2011:47-48.

[6] Tang Y,Xiao X,Shi Y.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.

[7] Liang X,Tang J,Pan L.A Neighborhood Vector Propagation Algorithm for Community Detection[C].Global Communications Conference (GLOBECOM),2014:2923-2928.

[8] Newman M E J.Modularity and Community Structure in Networks[J].Proceedings of the National Academy of Sciences,2006,103(23):8577-8582.

[9] Leskovec J,Mcauley J J.Learning to Discover Social Circles in Ego Networks[C].Advances in Neural Information Processing Systems,2012:539-547.

[10] Lancichinetti A,Fortunato S,Radicchi F.Benchmark Graphs for Testing Community Detection Algorithms[J].Physical Review E,2008,78(04):046110.

[11] Newman M E J.Fast Algorithm for Detecting Community Structure in Networks[J].Physical Review E,2004,69(06):066133.

[12] Raghavan U N,Albert R,Kumara S.Near Linear Time Algorithm to Detect Community Structures in Large-Scale

Networks[J].Physical Review E,2007,76(03):036106.

[13] Chen Y C,Zhu W Y,Peng W C,et al.CIM:Community-based Influence Maximization in Social Networks[J]. ACM Transactions on Intelligent Systems and Technology(TIST),2014,5(02):25.

作者简介：

徐　浩（1992—），男，硕士，主要研究方向为社交网络分析；

潘　理（1974—），男，博士，研究员，主要研究方向为网络安全管理、社交网络分析、云计算与大数据安全等。