# 一种新的基于社区结构的影响最大化方法

冀进朝，黄 岚，王 喆，李红明，李三义
(吉林大学 计算机科学与技术学院，长春 130012)

**摘要**：基于传播网络的结构性，提出一种新的基于社区结构的影响最大化方法AMICS. 该方法先利用已有社区挖掘算法识别出隐藏在网络中的社区结构，然后迭代选择跨越社区数最多的$k$个节点作为影响的初始传播者最大化影响的社区覆盖. 在小型网络和中等规模网络数据集上的实验表明，该算法比传统的影响最大化方法更具优势.

**关键词**：社区结构；影响最大化；社区覆盖
**中图分类号**：TP391 **文献标志码**：A **文章编号**：1671-5489(2011)01-0093-05

# A New Approach to Maximizing the Spread of Influence Based on Community Structure

JI Jin-chao, HUANG Lan, WANG Zhe, LI Hong-ming, LI San-yi
(*College of Computer Science and Technology, Jilin University, Changchun* 130012, *China*)

**Abstract**: Considering the structure of diffusion network, we proposed a new approach to maximizing the spread of influence based on community structure (AMICS). Our approach employs the community algorithm such as Radicchi's algorithm/ICS algorithm to detect the community structure hidden in the network firstly, then iteratively chooses $k$ important nodes which span the maximum communities to maximize the influence's community coverage. The experiments on the small network and medium network show that AMICS is feasible and effective.

**Key words**: community structure; maximizing influence; community covered

## 0 引 言

在社会网络中，通常个体用节点表示，个体间的联系用边表示. 传统的影响最大化问题的核心是在网络中发现最有影响力的$k$个节点作为影响的初始传播者，目的是使最终受到影响(如采纳某项建议、购买某种商品等)的个体数最多. 节点间的边被分配一个0～1间的权值，该值表示当其中一个节点被激活时影响另一个节点的概率. 文献[1-3]提出了一些识别最有影响力节点集的算法. 这些算法的主要目的为在影响传播的最终阶段，使受到影响的个体数最多. 但在实际生活中，社会网络具有社区结构，即网络中的节点因为某些因素(如朋友、亲戚、业余爱好等)而形成了社区[4-5]，这些社区中的节点具有某些共同特征. 同一社区内的节点联系密切，社区间的节点联系松散. 因此，在一定时期内使最终受到影响的社区数最多，可能更有意义. 如商品销售者为了让尽可能多的人了解他的商品，自然希望让商品影响到尽可能多的社区，这样受到影响的每个社区中的每个人都有可能了解他的商品.

收稿日期：2010-01-17.

作者简介：冀进朝(1982—)，男，汉族，硕士研究生，从事数据挖掘的研究，E-mail：jinchao0374@163.com. 通讯作者：王 喆(1974—)，男，汉族，博士，副教授，从事数据挖掘和商务智能的研究，E-mail：wz2000@jlu.edu.cn.

基金项目：国家自然科学基金(批准号：60673099；60873146)和国家高技术研究发展计划863项目基金(批准号：2007AA04Z114).

图1表示一个具有12个节点、2个社区的小型社会网络，A～G属于社区1，H～L属于社区2. 假设希望在该网络中找到一个最好的节点最大化影响传播. 目前的算法可能会选择激活点A(主要因为它的度高). 根据影响传播模型，激活点A后很可能会使激活的节点数最多，但影响却可能不会传播到社区2的任何一个节点上. 而如果选择节点G，则能提高两个社区都受影响的机会. 基于此，本文在分析社区挖掘算法的基础上，提出一种新的基于社区结构的影响最大化方法.

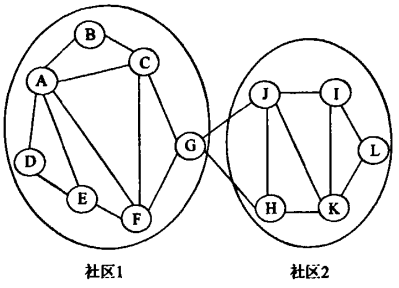


**图1 具有2个社区的社会网络**

**Fig. 1 Social network with two communities**

## 1 影响传播的模型及社区挖掘算法

通常影响最大化的关键是在网络中发现最有影响力的节点. 假设网络中的节点能够采纳某种思想，购买某种商品或类似的事，该过程称为激活. 再假设被激活的节点具有影响其邻居的能力，这些邻居可能自己选择活跃. 影响最大化的问题变为选择最好的 $k$ 个节点初始激活，目的是在影响最大化过程的最终阶段使受到影响的节点数最多. 文献[3]给出了描述节点激活行为的几种模型，其中线性阈值模型和独立级联模型是目前人们广泛研究的两种最基本的传播模型.

在考虑社会网络中某种观念或创新的传播模型时，通常把一个节点表示为活跃和不活跃两种状态. 一个节点只能处于这两种状态中的一种，并且在现有的模型中一个节点只能从不活跃状态变为活跃状态而不能从活跃状态转变为不活跃状态. 如果一个节点有越来越多的邻居节点变活跃，则它也越来越趋于活跃. 因此，传播过程可以从一个初始不活跃的节点 $v$ 的角度视为如下过程：即随着时间的流逝，$v$ 的邻居节点中有越来越多的节点变活跃；在某个时间点上，可能使 $v$ 变活跃，并且 $v$ 的决策可能依次触发与 $v$ 相连节点的决策. Granovetter[6]和Schelling[7]最早提出了表达该过程的模型，他们的方法基于具体节点的阈值. 之后人们提出了许多这种类型的模型[8-10]，其中一个重要的模型就是线性阈值模型. 在该模型中，一个节点 $v$ 受其每个邻居节点的影响，影响概率是权值 $b_{v,w}$，定义为

$$b_{vw} = \frac{f_{vw}}{\sum_{w=1}^{n} f_{vw}}, \tag{1}$$

其中：$f_{vw}$ 表示节点 $v$ 和 $w$ 之间的交互频率；权值 $b_{vw}$ 满足如下约束条件：

$$\sum_{v\text{的邻居节点}w} b_{v,w} \leqslant 1. \tag{2}$$

该过程按如下所述动态进行：每个节点 $v$ 随机地从0～1间选择一个阈值 $\theta_v$，此阈值表示为了让节点 $v$ 变活跃，$v$ 的邻居必须变活跃的权值部分. 给定一个阈值的随机选择，一个初始的活跃节点集 $A_0$(其他节点则不活跃)，传播过程在离散步骤确定性的展开：所有在步骤 $t-1$ 活跃的节点在步骤 $t$ 仍然活跃，如果 $v$ 活跃邻居的权值和不小于 $\theta_v$，即

$$\sum_{v\text{的邻居节点}w} b_{v,w} \geqslant \theta_v, \tag{3}$$

则在步骤 $t$ 激活任何满足上述条件的每个节点 $v$.

当节点 $v$ 的邻居采用某种新技术或接受某种新观念时，阈值 $\theta_v$ 表示节点 $v$ 也采用这种新技术或新观念的潜在趋势. 因为对节点的阈值缺乏相关知识，因此随机选择的阈值就是对这种状况的一种模拟，即对所有节点可能的阈值进行了有效平均；或将所有的阈值强制分配为一个固定值，如1/2.

另一种模型是独立级联模型. 在该模型下，节点到节点的影响传播按照相关规则以一个初始的活跃节点集开始，在离散的时间步内逐步完成. 在时间步 $t$ 变活跃的节点 $v$ 有一个机会使它当前不活跃的每个邻居节点 $w$ 在第 $t+1$ 步成为活跃节点. 节点 $v$ 成功激活节点 $w$ 的概率称为边的权 $p_{v,w}$($p_{v,w}$是系统参数，与其他值无关). 如果 $w$ 具有多个最近活跃的邻居节点，则它们以任意顺序尝试激活 $w$. 如果

邻居节点 $v$ 成功激活 $w$，则 $w$ 将在 $t+1$ 步变活跃. 无论 $v$ 是否成功激活 $w$，其不能在随后的步骤中再试探激活 $w$. 这个过程一直进行，直到不再发生激活事件为止. 目前该领域中的研究主要关心最大化最终激活的节点总数，本文将此问题扩展到关注最大化影响覆盖的社区数上. 如果一个社区中有一个节点被激活即称该社区被影响覆盖了.

目前，已有许多社区检测算法，主要包括二分法和层次聚类[11]. Radicchi 算法属于层次聚类算法，是对 GN 算法的改进，其主要思想是通过迭代移除具有较低边聚集系数 $C_{ij}$ 值的边找到一个理想的社区划分[12]；ICS 算法属于二分法[13]，根据聚类中心性从网络中得到节点的局部信息进而发现自然的网络社区. 最近，Scripps 等[14-15]探讨了在社会网络分析中节点角色的概念，对一个节点可能连接的社区数给出了估计，但前提是每个社区都是完全连通的，即社区内的任意两个节点都有联系. 而在实际社会网络中存在同一社区中两个个体之间没有联系的情况. 且 Scripps 等也未对一个节点连接几个社区给出精确估计. 基于此，本文提出一种新的算法 AMICS.

## 2　AMICS 算法

把个体数为 $n$ 的社会网络图及其拓扑结构用图 $G=(V,E)$ 表示. 其中：$V=\{v_1,v_2,\cdots,v_n\}$ 表示所要研究的社会网络中所有个体的集合；$E$ 为边集，表示个体间的联系. 用 $\boldsymbol{A}_{n\times n}$ 表示 $G$ 的邻接矩阵. 当个体 $i$ 和 $j$ 间有联系时，$\boldsymbol{A}_{i\times j}=1$，否则，$\boldsymbol{A}_{i\times j}=0$. 先用 Radicchi 算法或 ICS 算法在社会网络 $G$ 上发现社区，然后在已发现社区的基础上研究影响传播最大化，假设已发现的社区数为 $k$ 个，用 $\mathrm{Com}=\{\mathrm{Com}_1,\mathrm{Com}_2,\cdots,\mathrm{Com}_k\}$ 表示这些社区；$|\mathrm{Com}_i|$ 表示第 $i(i=1,2,\cdots,k)$ 个社区的个体数；$\mathrm{group}_i$ 表示第 $i$ 个社区的个体集；Com $N(v_i)$ 表示第 $i$ 个个体连接的社区数；$N(v_i)$ 表示第 $i$ 个节点的邻居节点集；$|N(v_i)|$ 表示节点 $v_i$ 的邻居节点数；$|V'|$ 表示个体集 $V'$ 中的元素数.

输入：邻接矩阵 $\boldsymbol{A}_{n\times n}$，$k$ 个社区的个体集 $\mathrm{group}=\{\mathrm{group}_1\sim\mathrm{group}_k\}$，$L$(限定选择的节点数)，个体集 $V'$，$m$.

输出：$L$ 个最好的节点.

步骤：

1) 扫描邻接矩阵 $\boldsymbol{A}_{n\times n}$，得到每个节点的邻居节点集 $N(v_1)\sim N(v_n)$；

2) 对于 $V$ 中的所有个体调用 CalculateCom $N(v_i)$ 计算每个节点的 Com $N$ 值；

3) 选择 Com $N$ 值最高的节点 $v_i$；如果具有最高 Com $N$ 值的节点多于一个，则选择度最大的节点；如果具有最高 Com $N$ 值并且度最大的节点数多于一个，则选择下标最小的节点；

4) $m=m+1$；

5) 删除与 $v_i$ 连接的所有个体集，更新 group；

6) 删除节点 $v_i$，更新 $V'$；

7) $t\leftarrow|V'|-1$；

8) 若 $m=L$，则算法结束；

9) 若 $m<L$：如果 $t\neq0$，且剩下的 group 数不为零，则转 2)；否则，在 $V'$ 中余下的其他节点中选择 $L-m$ 个 Com $N$ 值最高的节点，算法结束.

在步骤 3)中，计算节点的 Com $N$ 值时，需要调用 CalculateCom $N(v_i)$ 过程，该过程的主要思想是：对 $v_i$ 的所有邻居节点，统计计算它们属于社区的个数，然后把该值赋给 Com $N(v_i)$，此值即为 $v_i$ 连接的社区数. 下面简述该过程：

CalculateCom $N(v_i)$

1) 初始化 Com $N(v_i)=1$，$k=0$；

2) while $(k<|N(v_i)|)\{$

3) 对 $N(v_i)$ 中的每个节点 $v_j$；

① 若 $v_j$ 与 $v_i$ 在同一个 group 中，则 Com $N(v_i)$ 值不变；

② 若 $v_j$ 与 $v_i$ 在不同的 group 中，并且在该组中首次出现，则 Com $N(v_i)\leftarrow$ Com $N(v_i)+1$；

③ 若 $v_j$ 与 $v_i$ 在不同的 group 中，但之前已有 $N(v_i)$ 中的节点在该 group 中，则 Com $N(v_i)$ 值不变；
4) $k=k+1$；转 2)；}
5) return Com $N(v_i)$.

## 3 实验结果分析

假设每个社区至少有 2 个节点，一个网络至少有 2 个社区. 实验在 2 个数据集上进行，即某个小规模网络和合著网络数据集. 小规模网络为一个人造网络，合著网络数据集为来自 Pajek Datasets (http://vlado.fmf.uni-lj.si/pub/networks/data) 中的作者网络，是 86 个作者在引用 Graph products 的 158 篇文章中的合著关系，其中有 86 个节点，124 条边. 基于评估的目的，本文对 3 种不同算法得到的结果进行对比，比较了激活节点总数的同时也比较了覆盖的社区数. 基本随机方法随机选择 $k$ 个节点，度方法选择度最高的 $k$ 个节点，最后使用 AMICS算法，选择跨越社区数最多的 $k$ 个节点. 在影响最大化实验中，使用线性阈值传播模型. 由于阈值选取的随机性，本文对所有算法都运行了 1 000次，每次随机地从 0 ~ 1 之间重新选取节点的阈值，然后取这 1 000 次传播结果的平均值. 对于如图 2 所示的小规模网络初始目标集选为 2；对于合著网络数据初始目标集选为 4 个节点. 小规模网络的实验结果列于表 1，合著网络数据的实验结果列于表 2.

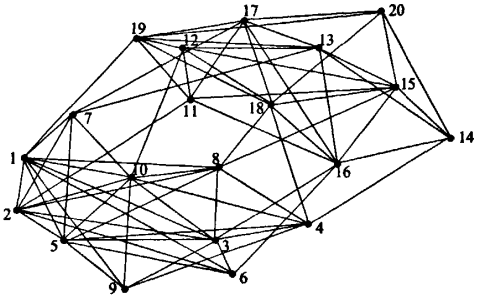


图 2 小规模网络

Fig. 2 Small scale network

表 1 影响最大化算法在小规模网络上的比较

Table 1 Comparison of the influence maximization's algorithm with others on a small scale network

| 算法 | 受影响的节点数/个 | 社区覆盖数/个 |
|---|---|---|
| 随机法 | 5.759 | 2.900 |
| 度方法 | 7.264 | 2.767 |
| AMICS | 5.588 | 2.989 |

表 2 影响最大化算法在合著网络上的比较

Table 2 Comparison of the influence maximization's algorithm with others on coauthors network

| 算法 | 受影响的节点数/个 | 社区覆盖数/个 |
|---|---|---|
| 随机法 | 9.220 | 4.540 |
| 度方法 | 29.623 | 7.414 |
| AMICS | 29.791 | 8.371 |

受影响的节点数列列出了被最初目标节点集激活的节点数平均值. 社区覆盖列表示至少有一个节点被激活的社区数平均值. 由表 1 可见，AMICS 算法的社区覆盖率最大，其次是随机方法和度方法. 由表 2 可见，AMICS 算法的社区覆盖率依然最大，其次是度方法，随机算法的效果最差. 因此，AMICS 算法在社区影响最大化方面优于其他两种方法，这也显示了本文所提算法的有效性. 此外，本文算法在具有多个社区结构的大型社会网络中更具优势.

综上所述，本文在现有最大化影响传播研究的基础上，根据社区挖掘的最新发展，把社区结构的知识应用到影响最大化领域，提出了一种新的基于社区结构的影响最大化算法，并将该算法在一个小型网络和一个中型的社会网络数据集上进行实验. 实验结果表明了本文算法的有效性.

## 参 考 文 献

[1] Domingos P, Richardon M. Mining the Network Value of Customers [C]//Conference on Knowledge Discovery in Data Mining. New York: ACM Press, 2001: 57-66.

[2] Richardson M, Domingos P. Mining Knowledge-Sharing Sites for Viral Marketing [C]//Proceedings of the Eighth Intl Conf on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 61-67.

[3] Kempe D, Kleinberg J, Tardos E. Maxmizing the Spread of Influence through a Social Network [C]//Proceedins of the 9th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 137-146.

[4] CAI Hua, ZHOU Chun-guang, LU Ting-yu, et al. OCSMA: An Alogorithm to Mine Overlapping Community Structure in Networks [J]. Journal of Jilin University: Engineering and Technology Edition, 2009, 39(4): 1035-1040. (才华，周春光，卢廷玉，等. 重叠社区结构的挖掘算法 [J]. 吉林大学学报：工学版，2009, 39(4): 1035-1040.)

[5] Girvan M, Newman M E J. Community Structure in Social and Biological Networks [J]. Proc Natl Acad Sci USA, 2002, 99(12): 7821-7826.

[6] Granovetter M. Threshold Models of Collective Behavior [J]. The American Journal of Sociology, 1978, 83(6): 1420-1443.

[7] Schelling T C. Micromotives and Macrobehavior [M]. New York: Norton, 1978.

[8] Berger E. Dynamic Monopolies of Constant Size [J]. Journal of Combinatorial Theory: Series B, 2001, 83(2): 191-200.

[9] Young H P. The Diffusion of Innovations in Social Networks [R]. MD, Baltimore: Department of Economics, The Johns Hopkins University, 2002: 1-19.

[10] Richardson M, Domingos P. Mining Knowledge-Sharing Sites for Viral Marketing [C]//Proceedings of the Eighth Intl Conf on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 61-70.

[11] 汪小帆，李翔，陈关荣. 复杂网络理论及其应用 [M]. 北京：清华大学出版社，2006: 162-188.

[12] Radicchi F, Cstellano C, Cecconi F, et al. Defining and Identifying Communities in Networks [J]. Proc Natl Acad Sci USA, 2004, 101(9): 2658-2662.

[13] YANG Bo, LIU Ji-ming. Discovering Global Network Communities Based on Local Centralities [J]. ACM Transactions on the Web, 2008, 2(1): Article9.

[14] Scripps J, TAN Pang-ning, Esfahanian A H. Node Roles and Community Structure in Network [C]//Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. New York: ACM Press, 2010: 26-35.

[15] Scripps J, TAN Pang-ning, Esfahanian A H. Exploration of Link Structure and Community-Based Node Roles in Network Analysis [C]//Proceedings of the 2007 7th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2007: 649-654.