复杂网络的重叠社区及社区间的结构洞识别

刘世超,朱福喜,冯 曦
(武汉大学计算机学院,湖北武汉 430072)

摘 要: 大数据环境下如何有效地、准确地识别复杂网络的重叠社区是近年来学者关注的重点. 本文提出一种基于多标签传播方式 MLPS(Multiple Label Propagation Strategy)的重叠社区识别算法,该算法首先利用影响力最大化模型选取初始种子集合并赋予它们唯一的标签,然后采用结点间的相似性和影响传播特性共同作用于标签的传播迭代过程,迭代停止后将具有相同标签的结点划分为同一社区. 通过合成网络和真实网络的实验验证了 MLPS 算法具有较高的准确度和模块度,且具有接近线性的时间复杂度. 另外,在对 MLPS 算法输出的重叠结构进行分析的基础上,本文提出社区间的结构洞识别算法 SHCDA(Structural Holes Between Communities Detection Algorithm),该算法通过分析重叠结构和重叠结点的位置特征,计算重叠结点作为结构洞的得分,最后输出 top-k 结构洞. 本文在不同特性的数据集上进行实验,结果证明了 SHCDA 算法具有最好的准确度.

关键词: 复杂网络; 重叠社区; 多标签传播; 结构洞识别
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2016)11-2600-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.11.006

Identifying Overlapping Communities and Structural Holes Between Communities in Complex Networks

LIU Shi-chao, ZHU Fu-xi, FENG Xi
(*Computer School of Wuhan University, Wuhan, Hubei* 430072, *China*)

Abstract: Many researchers focus on how to detect overlapping communities effectively and accurately when coping with large-scale networks in recent years. This paper proposes a novel overlapping community detection algorithm based on a multiple label propagation strategy, called MLPS algorithm. Firstly, MLPS selects a set of nodes as initial seeds by using Influence Maximization Model, each of which is assigned a unique label; Inspired by strategy based on similarity and influence diffusion, MLPS incorporates with these two strategies to guide the process of label propagation; Finally, nodes with the same tag are divided into one community after propagation. Experimental results on synthetic datasets and real networks illustrate that MLPS has both high accuracy and modularity at the same time. In addition, another algorithm named Structural Holes between Communities Detection Algorithm (SHCDA) is presented on the basis of the output of MLPS. SHCDA computes the scores of overlapping nodes who serve as structural holes by analyzing the overlapping structure and position feature of overlapping nodes, and then selects top-k structural holes as the output. Experimental results on different datasets show that SHCDA gets the best accuracy.

Key words: complex networks; overlapping community; multiple label propagation; structural holes detection

1 引言

现实世界的许多网络普遍体现了社区结构特性,即同一社区内的结点关系紧密,而不同社区间的结点关系稀疏[1]. 近年来针对社区发现的算法层出不穷,在社区的层次性和重叠性等方面开展了大量的研究[2~4]. 本文着重研究社区的重叠性,尤其是大数据环境下,如何有效地解决重叠社区挖掘的问题. 社区重叠结点是社区间的桥梁即社区间的结构洞,社区间的结构洞是两个或多个社区之间的非重复关系,占据结构洞的个体拥有更多的机会获取信息或资源[5]. 识别社区间结构洞的关键在于准确地挖掘网络的重叠结构,从而找

收稿日期:2015-05-12;修回日期:2015-11-09;责任编辑:马兰英
基金项目:国家自然科学基金(No. 61272277);中央高校基本科研业务费专项基金(No. 274742)

出结构洞的候选集合,即社区重叠结点的集合.

一些学者针对结构洞的形成和作用进行建模研究[6,7],而对结构洞的识别算法还比较少. 文献[8]采用拓扑势的方式识别重叠社区和社区间的结构洞,但作者直接将重叠结点作为结构洞,没有定量地分析结构洞的重要性. 在实际情况中,我们往往更加关心那些比较重要的结构洞,于是 Tang J 等[9]提出 top-k 结构洞挖掘的问题,作者认为结构洞的邻居结点在一定程度上反映该结构洞的重要性,但忽略了对重叠结构的分析. 如图 1(a)中的结构洞 A 与图 1(b)的结构洞 C 拥有相同的结构,而图 1(b)中的两个社区只存在一个结构洞 C,定量计算时 C 的重要性应大于 A.

本文的主要贡献:提出一种基于多标签传播方式(MLPS)的重叠社区挖掘算法,算法的输出结果拥有较高的准确度和模块度;分析社区重叠结构,提出社区间的结构洞识别算法(SHCDA),能有效地挖掘 top-k 结构洞.

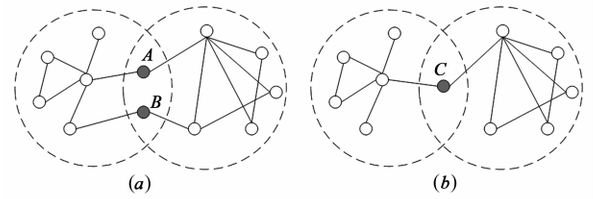


图1　重叠结构及结构洞示意图

2　标签传播的相关工作

Raghavan 等[10]提出的标签传播算法 LPA(Label Propagation Algorithm),首次将标签传播应用于社区发现,具有接近线性的时间复杂度. Barber 等人[11,12]将标签传播算法转化为优化问题,Xie J[13]结合了路径衰减的思想扩展了对结点邻居的计算,提高了社区划分结果的精度. COPRA(Community Overlap Propagation Algorithm)算法[14]是 LPA 算法的扩展,可以解决重叠社区挖掘的问题. 初始化时每个结点被赋予一组标签对(c,b),c 是标签,b 是结点拥有标签 c 的概率;在标签更新过程中设置阈值 v,删除 b 小于 v 的标签对. 文献[15]指出 COPRA 算法的阈值应根据结点所处的位置不同而变化,并通过平衡阈值的选择来提高生成社区的质量. SLPA(Speaker-listener Label Propagation Algorithm)[16]通过模拟人类交流的行为特性,定义了动态的交互规则来指导结点的标签更新. 文献[17]提出了 MLPA(Multi-label Propagation Algorithm)算法,该算法根据结点间的相似性来确定传播概率,具有较高的准确度,但是同 COPRA 算法一样,算法容易产生大量的社区,导致模块度较低.

3　多标签传播方式的重叠社区识别算法

3.1　算法主要思想

COPRA 是经典的标签传播算法,在该算法的标签传播过程中,结点直接接受邻居结点的标签,本文定义这种标签传播的方式为 DA(Directly Accept),即

$$P_{ij} = b_i \tag{1}$$

其中,P_{ij}表示标签 c_i 从结点 i 传播到邻居结点 j 的概率,b_i 是结点 i 拥有标签 c_i 的概率. 这种直接接受标签的方式并不能真实地反应社会网络中信息和资源的传播,可能会降低算法的准确度.

MLPA 算法假定两个结点拥有越多的共同邻居,标签在结点间传播的概率就越大,定义这种标签传播的方式为 SS(Structural Similarity),即

$$P_{ij}^{SS} = \sqrt{S_{ij} \cdot b_i} \tag{2}$$

$$S_{ij} = \frac{|\Gamma(i) \cap \Gamma(j)|}{\sqrt{|\Gamma(i)| \cdot |\Gamma(j)|}} \tag{3}$$

其中,S_{ij}表示结点 i 和 j 的相似性,$\Gamma(i)$是结点 i 的邻居结点集合. 在相关工作中[17]已经提到,MLPA 算法同 COPRA 算法一样,容易产生大量的社区,导致输出结果的模块度较低. 因此,如何在保证准确度的情况下提高算法的模块度和稳定性,是本文研究的重点.

分析现实的网络,结点影响(标签)的传播在一定程度上取决于结点所处的位置,具有较高影响力的结点往往处于社区的核心位置,能够影响周围影响力较低的结点[18]. 于是,本文定义了一种新的基于影响传播的标签传播方式 ID(Influence Diffusion),即

$$P_{ij}^{ID} = \sqrt{\mathscr{C}_{ij} \cdot b_i} \tag{4}$$

$$\mathscr{C}_{ij} = \frac{\varphi_i}{\varphi_i + \varphi_j} \tag{5}$$

其中,$\mathscr{C}_{ij}$为标签从结点 i 传播到 j 的度量,φ_i 表示结点 i 在其邻域结构的相对影响能力,即

$$\varphi_i = \frac{Inf_i - \min\limits_{k \in Nb_i} Inf_k + Z}{\max\limits_{k \in Nb_i} Inf_k - \min\limits_{k \in Nb_i} Inf_k + Z} \tag{6}$$

其中,Nb_i 为结点 i 及其邻居结点的集合;Inf_i 是结点 i 的影响力值,可用 LeaderRank 算法[19]计算获得;Z 是平滑因子,取结点邻域结构的影响力均值,其计算公式为:

$$Z = \frac{1}{|Nb_i|} \sum_{k \in Nb_i} Inf_k \tag{7}$$

由式(6)可知结点 i 的相对影响能力 φ_i 取值区间为(0,1],且 φ_i 值越大,结点 i 的标签传播能力越强. 结合式(5)和式(6)可得标签从结点 i 传播到 j 的度量 $\mathscr{C}_{ij}$取值区间为(0,1),且 φ_i 与 φ_j 的差值越大,标签越容易传播. 由于大部分网络结点的相对影响力有明显的差异,导致标签在网络中不均衡地流动,降低了原有算法的随机性,使

得算法能够很快收敛且输出的结果较为稳定.

于是本文提出一种基于多标签传播方式(MLPS)的重叠社区识别算法,综合 SS 和 ID 两种方式进行标签传播,那么结点 i 的标签传播到邻居结点 j 的概率 P_{ij} 表示为:

$$P_{ij}=P_{ij}^{ID}\cdot P_{ij}^{SS} \tag{8}$$

目前的标签传播算法对网络进行初始化时,为每个结点都赋予一个独立的标签,这会导致输出结果出现大量孤立的小社区,从而降低输出社区的质量. 为此,本文采用文献[20]提出的影响力最大化模型,选取网络中一组初始传播结点,使得这些结点能辐射到网络大部分结点,且尽可能的散落在每一个可能的社区,可以减少在传播时不需要的判断开销.

另外,COPRA 算法的参数 v 是过滤标签对的阈值,设为全局值是不合理的. 本文采用结点依赖的阈值选择方式[15,17],即任意结点 i 拥有独立阈值 v_i. 给定结点 i 的标签对集合为 $\{(c_1,b_1),(c_2,b_2),\cdots,(c_k,b_k)\}$,定义 $v_i=b_i^{\max}\cdot p$,其中 $b_i^{\max}=\max\{b_1,b_2,\cdots,b_k\}$,$p$ 作为算法的阈值选择参数. 这样,既平衡了所有结点的阈值,又能有效地挖掘结构洞.

3.2 重叠社区发现的 MLPS 算法

为了提升算法运行的准确性和稳定性,本节提出一种基于多标签传播方式(MLPS)的重叠社区识别算法,算法逻辑流程如图 2 所示.

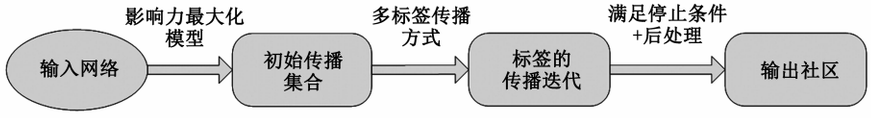


图2 MLPS算法逻辑流程图

MLPS 算法的迭代停止条件和后处理沿用 COPRA 算法的设定,因此本文只给出改进的标签传播方式和阈值选择策略的部分代码,即算法 1. 算法设置两个标签向量:***old*** 和 ***new***,***old. x***(***new. x***)表示结点 x 更新前(后)的标签,每个结点都拥有一组标签对(c,b),c 是标签,b 是结点拥有标签 c 的概率,$N(x)$是结点 x 的邻居集合,算法的唯一参数是阈值选择参数 p.

算法 1 Propagate_MLPS(x,*old*,*new*)

```
1.  old. x←{};
2.  FOR EACH vertex y in N(x)
3.      FOR EACH (c,b_y) in old. y
4.          Compute P_xy← formula(8);
5.          IF for some b_x,(c,b_x) is in new. x
6.              new. x ← new. x −{(c,b_x)}∪{(c,b_x + P_xy)};
7.          ELSE new. x ← new. x∪{(c,P_xy)};
8.  b_max← 0;
9.  FOR EACH(c,b) in new. x
10.     IF b > b_max
11.         b_max← b;
12. FOR EACH(c,b) in new. x
13.     IF b < b_max · p
14.         new. x ← new. x −{(c,b)};
15. Normalize(new. x);
```

3.3 算法的复杂度分析

定义 N 是结点个数,$\bar{d}$ 表示平均度,m 为结点拥有的平均标签个数,在 COPRA 算法的基础上,MLPS 算法:在初始化时增加了种子集合的选取 $O(N)$;增加了结点间的相似性计算 $O(N\cdot\bar{d}^2)$ 和影响传播度量的计算 $O(N\cdot\bar{d})+O(N\cdot\log N)$;平衡了阈值后,算法不限制结点拥有的标签数,每次迭代的复杂度 $O(N\cdot\bar{d}\cdot m)$. 算法的迭代次数一般很小,且大部分网络是稀疏的($\bar{d}<<N$),当数据量变大时($m<<N$),MLPS 算法拥有 $O(N\cdot\log N)$ 的复杂度. 在实际计算中,$O(N\cdot\log N)$ 的复杂度主要是影响力计算造成的. 对此,我们可以离线计算所有结点的影响力来提高算法效率.

4 社区间的结构洞识别算法

本节根据 MLPS 算法的输出结果,提出社区间的结构洞识别算法(SHCDA). 算法首先将所有重叠结点加入结构洞的候选集合,然后分析社区重叠结构和重叠结点的位置特征,输出重要性得分 top-k 的结构洞.

定义 ON 为重叠结点集合,C_i 是重叠结点 i($i\in ON$)归属的社区集合,$p_{i,c}$ 表示结点 i 属于社区 c 的概率,满足 $\sum_{c\in C_i}p_{i,c}=1$. 根据文献[9]的分析,结构洞的邻居结点能一定程度地反映其重要性,于是本文将结构洞的邻居结点的加权影响力均值作为该结构洞的重要性得分. 那么,结构洞 i 的重要性得分为:

$$score_i=\frac{1}{n}\sum_{j\in c,c\in C_i}Inf_j\cdot w_{ij} \tag{9}$$

其中,n 为结点 i 的邻居数,Inf_j 表示结点 j 的重要性;$w_{ij}=p_{i,c}$为结构洞 i 与其邻居结点 j 的关系权重,c 是结点 j 归属的社区.

本文观测真实的网络数据,发现部分结构洞连接的只是一些普通结点,因受到邻居结点的影响,导致结构洞的得分过低. 此外,结构洞所处的结构特性也会影响其重要性得分,令 C_i 表示重叠结点 i 归属的社区集

合,OC_i 为结点 i 所处的重叠结构,即 $OC_i = \{oc \mid oc = C_{i1} \cap C_{i2} \cap \cdots \cap C_{im}\}$ 其中,$C_i = \{C_{i1}, C_{i2}, \cdots, C_{im}\}$,$m$ 是集合 C_i 的大小,$m = |C_i|$ 且 $m \geqslant 2$. 然后加入惩罚系数 ε 来衡量重叠结构的影响,这样算法输出的结果将更加符合真实的情况. 根据实验结果,设置 ε 的经验值为 0.1. 最终结构洞 i 的重要性得分计算公式如下:

$$score_i = \max\left\{Inf_i, \frac{1}{n}\sum_{j \in c, c \in C_i} Inf_j \cdot w_{ij}\right\} \cdot e^{-\varepsilon \cdot |OC_i|} \quad (10)$$

SHCDA 算法描述如算法 2.

算法 2　SHCDA

输入:网络 $G(V,E)$,MLPS 算法输出的社区重叠结构 ON,任意结点 i 归属的社区 c 及归属概率 $p_{i,c}$,惩罚系数 ε,结点 i 的邻居结点集合 Nb_i;

输出:top-k 结构洞.

```
score ← { }; output ← { }; sh_candidate ← ON;
FOR EACH vertex i in ON
    IF ∀ vertex p,q in NB_i, ∃(p,q) in E
        OR ∀ vertex p in NB_i, ∃p in ON
            sh_candidate ← sh_candidate − i;
FOR EACH sh in sh_candidate
    Compute score_sh ← formaula(10);
    score ← score ∪ score_sh;
Count = 0;
WHILE Count < k
    output ← output ∪ max(score);
    score ← score − max(score);
    Count = Count + 1;
```

一般的,k 取值较小,因此时间复杂度近似为 $O(d \cdot r^2)$,其中 d 为重叠结点的平均度,r 是重叠结点的个数. 在大数据的稀疏网络环境下,d 和 r 都较小,所以算法能够有效地处理大规模网络.

5　实验分析

5.1　MLPS 实验

MLPS 算法的参数 p 取值根据数据集的不同而变化. 给定数据集 D,本文定义 p 在[0,1]范围内以步长 0.05 搜索,取使得算法输出结果最优的值作为算法在数据集 D 上的参数.

5.1.1　评价方法

NMI(Normalized Mutual Information)[21]广泛应用于重叠社区挖掘的实验中,NMI 值越接近 1 说明算法输出的结果越接近真实的情况. 对于合成网络数据,存在标准的社区划分结果,而大部分的真实网络并没有标准的结果,因此实验采用模块度 Q_{ov}[22]来分析算法在真实网络中运行的效果.

5.1.2　合成网络实验

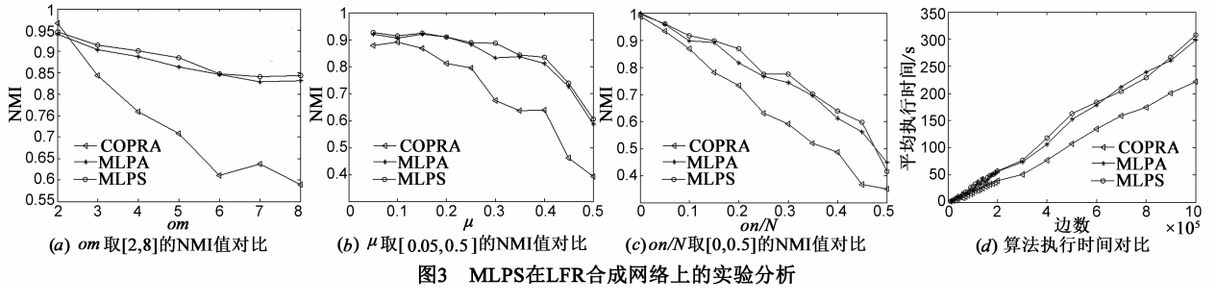
LFR 网络[21]通过模拟不同特性的真实网络,可从多个角度测试算法的性能. 实验选取 LFR 网络的三个参数(om、μ 和 on/N)来验证算法的 NMI 值,其中 om 为重叠结点拥有的标签数;μ 表示结点与社区外部结点链接的概率,随着 μ 值增大,结点的社区外部链接增多,导致合成网络结构变得更加复杂,社区发现的难度也随之增加;on/N 表示重叠结点占网络结点的比例. 当实验观测其中某参数变化时,其他参数不变. 实验结果如图 3(a)、3(b)、3(c)所示,MLPS 算法的 NMI 值最优,表明 MLPS 算法在处理不同特性的网络时能够取得较好的准确度. 图 3(d)分析了三种算法在网络边数从 10^4 增长到 10^6 时的运行效率,结果显示了这三种算法都具有接近线性的时间复杂度.

5.1.3　真实网络实验

本文选取 7 个不同的真实网络,如表 1 所示,其中 C-DBLP1 和 C-DBLP2 是从 WAMDM 实验室提供的数据库中抽取得到的两个不同规模的学者合著网络. 表 1 比较了 MLPS、MLPA 和 COPRA 算法取最优参数时的平均模块度 Q_{ov} 和方差 Std.,结果显示本文提出的 MLPS 算法能够取得较好的效果.

表 1　三种算法在真实网络上运行的模块度比较

Network Reference	N,E	MLPS			MLPA			COPRA		
		p	Q_{ov}	Std.	p	Q_{ov}	Std.	v	Q_{ov}	Std.
Karate[23]	34,78	0.95	**0.754**	0.000	0.65	0.744	0.002	3	0.459	0.140
Dolphins[24]	62,159	0.95	0.774	0.000	0.7	**0.775**	0.006	4	0.677	0.058
Football[25]	115,613	0.8	**0.769**	0.000	0.45	0.701	0.000	3	0.699	0.025
Netscience[26]	379,914	0.5	**0.853**	0.000	0.45	0.840	0.004	6	0.816	0.016
C-DBLP1	2207,3899	0.5	**0.870**	0.006	0.45	0.852	0.002	6	0.797	0.012
C-DBLP2	8257,19683	0.65	**0.755**	0.000	0.5	0.720	0.002	7	0.740	0.006
Cond-mat[27]	16264,47594	0.35	**0.69**	0.000	0.678	0.68	0.000	6	0.656	0.045



(a) om 取[2,8]的NMI值对比　(b) μ 取[0.05,0.5]的NMI值对比　(c) on/N 取[0,0.5]的NMI值对比　(d) 算法执行时间对比

图3　MLPS在LFR合成网络上的实验分析

5.2 SHCDA 实验

本节实验包含 Coauthor 实验和新浪微博实验. Coauthor 数据集[9]存在可对比的实验结果,能够验证 SHCDA 算法的准确度;新浪微博数据集通过爬虫程序抓取,从一个特定用户开始,抓取最近发表的微博中转发数较高的微博及转发该微博的用户,并以广度优先的策略循环抓取数据,最后整理出 63,641 个用户及 1,391,718 条用户关系,其中包含 27,759 条微博转发关系.

5.2.1 Coauthor 实验

Coauthor 数据集包含 28 个计算机相关的会议,涉及六个研究领域,如表 2 所示. 每个领域都拥有一组程序委员会成员,我们认为不同领域间的共同委员会成员即为这些领域间的结构洞. 该数据集包含 1718 个委员会成员,其中拥有 107 个结构洞.

根据上述六个领域的相关性,本文选取三对领域(AI-DM,DB-DM 和 DP-NC)来验证算法的准确度. 同时实验选取 PageRank[28]、COPRA 和 MLPA 算法进行对比分析,如图 4、图 5 和图 6 所示. 本文提出的 SHCDA 算法在 top-30 之前都能取得最好的效果,而且从三个图中可以发现,PageRank 选取的重要结点大部分都不是结构洞,因此结构洞在网络中可能只是一些普通结点,却扮演着极其重要的角色.

表 2 Coauthor 数据集

领域	全称	包含会议
AI	Artificial Intelligence	IJCAI, AAAI, ICML, UAI, UMAP, NIPS, AAMAS
DB	Databases	VLDB, SIGMOD, PODS, ICDE, ICDT, EDBT
DM	Data Mining	SIGKDD, ICDM
DP	Distributed Parallel Computing	PPoPP, PACT, IPDPS, ICPP, Euro-Par
GV	Graphics, Vision and HCI	SIGGRAPH, CVPR, ICCV
NC	Networks, Communications and Performance	SIGCOMM, PERFORMANCE, SIGMETRICS, INFOCOM, MOBICOM

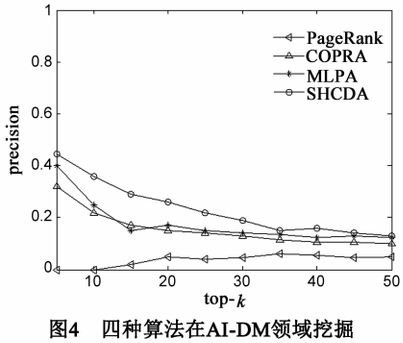


图4 四种算法在AI-DM领域挖掘top-k结构洞的准确度

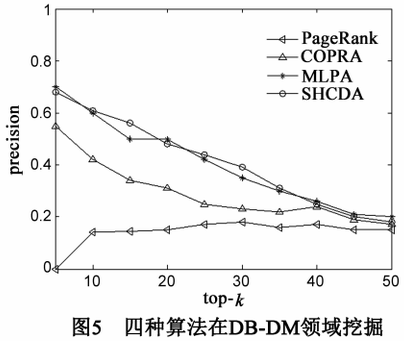


图5 四种算法在DB-DM领域挖掘top-k结构洞的准确度

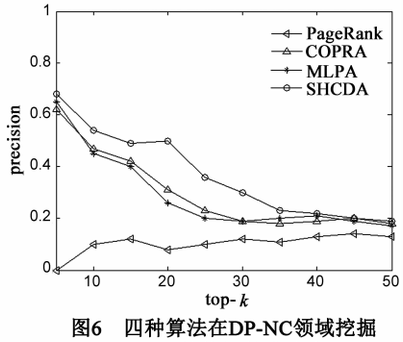


图6 四种算法在DP-NC领域挖掘top-k结构洞的准确度

5.2.2 新浪微博实验

在微博、Twitter 等社交网络中,社区间的结构洞是不同群体进行信息传播的桥梁,一般存在于微博的转发路径中[9]. 因此,本文定义 Precision 来衡量 SHCDA 算法的准确性:

$$\text{Precision} = \frac{|N_f(k)|}{|N_{sh}(k)|} \tag{11}$$

其中,$N_{sh}(k)$ 表示 SHCDA 算法输出的 top-k 结构洞,$N_f(k)$ 是 $N_{sh}(k)$ 中存在转发微博行为的结点集合. 实验结果如图 7 所示,显示出本文提出的 SHCDA 算法取得了最优的准确性,能够有效地挖掘出社区间的结构洞.

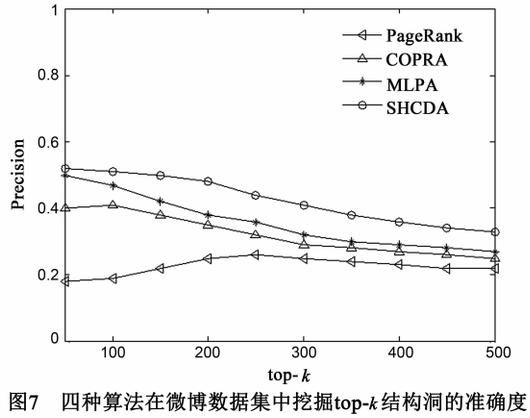


图7 四种算法在微博数据集中挖掘top-k结构洞的准确度

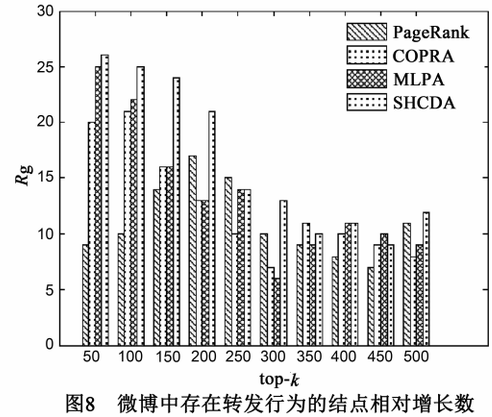


图8 微博中存在转发行为的结点相对增长数

本文还考察了结构洞的重要性对社区间的信息传播的影响. 针对算法输出的 top-k 结构洞,定义 Rg 表示在固定步长下存在转发行为的结点相对增长数:

$$Rg = |N_f(k+C)| - |N_f(k)| \quad (12)$$

其中,C 为增长步长,本节实验中 C 取 50. 从图 8 中可以看出 SHCDA 算法在 top-200 前有较快的增幅,而后趋于稳定,说明重要度较高的结构洞更容易进行社区间的信息传播,这对社会网络的信息传播机制研究和舆情监控具有重要的指导意义.

6　结论

为提高大数据环境下重叠社区发现的效果,本文结合结点间的相似性传播和影响传播两种方式,提出了基于多标签传播方式的重叠社区发现算法(MLPS),实验验证了该算法取得了较高的准确度和模块度,且具有接近线性的时间复杂度;同时,基于 MLPS 算法的输出结果,本文针对重叠结构进行分析,提出了社区间的结构洞识别算法(SHCDA),在不同特性数据集的实验显示了该算法拥有较高的准确度.

参考文献

[1] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821 – 7826.

[2] Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814 – 818.

[3] Shi C, Cai Y, Fu D, et al. A link clustering based overlapping community detection algorithm[J]. Data & Knowledge Engineering, 2013, 87: 394 – 404.

[4] Whang J J, Gleich D F, Dhillon I S. Overlapping community detection using seed set expansion[A]. CIKM 2013[C]. San Francisco, CA, USA, ACM, 2013. 2099 – 2108.

[5] Burt R S. Structural Holes: The Social Structure of Competition[M]. MA, Harvard University Press, 1992.

[6] Kleinberg J, Suri S, Tardos é, et al. Strategic network formation with structural holes[A]. Proceedings of the 9th ACM Conference on Electronic Commerce[C]. Chicago, Illinois, USA, ACM, 2008. 284 – 293.

[7] Buskens V, Van de Rijt A. Dynamics of networks if everyone strives for structural holes[J]. American Journal of Sociology, 2008, 114(2): 371 – 407.

[8] 李泓波,等. 基于拓扑势的重叠社区及社区间结构洞识别. 电子学报, 2014, 42(1): 62 – 69.
Li H B, et al. Identification of overlapping communities and structural holes between communities based on topological potential[J]. Acta Electronica Sinica, 2014, 42(1): 62 – 69. (in Chinese)

[9] Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks[A]. WWW 2013[C]. Rio de Janeiro, Brazil, International World Wide Web Conferences Steering Committee, 2013. 825 – 836.

[10] 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(3): 036106.

[11] Barber M J, Clark J W. Detecting network communities by propagating labels under constraints[J]. Physical Review E, 2009, 80(2): 026129.

[12] Liu X, Murata T. Advanced modularity-specialized label propagation algorithm for detecting communities in networks[J]. Physica A: Statistical Mechanics and its Applications, 2010, 389(7): 1493 – 1500.

[13] Xie J, Szymanski B K. Community detection using a neighborhood strength driven label propagation algorithm[A]. NSW 2011[C]. New York, USA, IEEE, 2011. 188 – 195.

[14] Gregory S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 103018.

[15] Wu Z H, Lin Y F, Gregory S, et al. Balanced multi-label propagation for overlapping community detection in social networks[J]. Journal of Computer Science and Technology, 2012, 27(3): 468 – 479.

[16] Xie J, Szymanski B K, Liu X. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[A]. ICDMW 2011[C]. Vancouver, BC, Canada, IEEE, 2011. 344 – 349.

[17] Dai Q, Guo M, Liu Y, et al. MLPA: Detecting overlapping communities by multi-label propagation approach[A]. CEC 2013[C]. Cancun, Mexico, IEEE, 2013. 681 – 688.

[18] Aral S, Walker D. Identifying influential and susceptible members of social networks[J]. Science, 2012, 337(6092): 337 – 341.

[19] Li Q, et al. Identifying influential spreaders by weighted LeaderRank[J]. Physica A: Statistical Mechanics and its Applications, 2014, 404: 47 – 55.

[20] Wang Y, Feng X. A Potential-Based Node Selection Strategy for Influence Maximization in a Social Network[M]. Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2009. 350 – 361.

[21] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms[J]. Physical Review E, 2008, 78(4): 046110.

[22] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009(03): P03024.

[23] Zachary W. An information flow model for conflict and fission in small groups1[J]. Journal of anthropological research,1977,33(4):452 -473.
[24] Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations[J]. Behavioral Ecology and Sociobiology,2003,54(4):396 -405.
[25] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences,2002,99(12):7821 -7826.
[26] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Physical review E,2006,74(3):036104.
[27] Newman M E J. The structure of scientific collaboration networks[J]. Proceedings of the National Academy of Sciences,2001,98(2):404 -409.
[28] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999 -66.

作者简介

刘世超 男,1989 年 2 月出生,河南周口人. 武汉大学计算机学院博士生,研究方向为社会网络分析和 Web 数据挖掘.
E-mail:nani@ whu. edu. cn

朱福喜(通讯作者) 男,1957 年 7 月出生,湖北武汉人. 武汉大学计算机学院教授、博士生导师,主要从事人工智能和 Web 数据挖掘.
E-mail:fxzhu@ whu. edu. cn

冯 曦 女,1991 年 11 月出生,陕西西安人. 武汉大学计算机学院硕士生,研究方向为社会网络分析和数据挖掘.
E-mail:fengxiwhu@ outlook. com