

基于信息扩散概率的二分网络社区划分算法^①

张 弛¹, 周永刚¹, 范纯龙², 唐桂林¹

¹(安徽邮电职业技术学院 计算机系, 合肥 230031)

²(沈阳航空航天大学 计算机学院, 沈阳 110136)

摘 要: 二分网络是复杂网络的表现形式之一, 二分网络单侧节点的社区划分对研究复杂网络具有重要的实际意义. 基于信息在网络中的扩散概率和模块度思想, 本文提出了一个针对二分网络的社区划分聚类算法(IPS 算法). 该算法通过模拟信息在网络中扩散的过程, 利用各个节点的信息量在网络中扩散后, 每个节点收到其他节点的信息量作为社区之间合并的依据, 并引入二分网络模块度作为社区划分优劣判断的依据. 最后算法在典型网络上测试结果表明, 该算法不仅能够精确的识别二分网络社区个数, 而且可以获得高质量的社区划分结果.

关键词: 模块度; 二分网络; 支持度; 社区划分

Community Division of Bipartite Network Based on Information Transfer Probability

ZHANG Chi¹, ZHOU Yong-Gang¹, FAN Chun-Long², TANG Gui-Lin¹

¹(School of Computer Science, Anhui Posts And Telecommunications Technical College, Hefei 230031, China)

²(School of Computer Science, Shenyang Aerospace University, Shenyang 110136, China)

Abstract: Bipartite network is a performance of complex networks, the divided of unilateral node of bipartite network has important practical significance for the study of complex networks of community division. Based on the idea of information transfer probability, this paper presents a community of bipartite network divided clustering algorithm (IPS algorithm). The algorithm simulates the probability of information transfer in the network, through mutual support value between the nodes in network. select the max value as the basis for merger different communities and introduction of bipartite network module as a basis for judging the merits of community division. Finally, we using actual network test the performance of the algorithm. Experimental results show that the algorithm can not only accurate divided the unilateral node of bipartite network, but also can get high quality community division.

Key words: modularity; bipartite network; support value; community division

复杂网络是近年来兴起的一门涉及物理学、生物学、数学和计算机科学等领域的交叉学科, 针对社区结构的研究大致分为 3 个类型: 社区发现、社区演化分析、及社区结构与网络动力学和网络压缩表示等功能性特征特征之间的关系. 其中社区发现的研究大都集中在单部网络中, 二分网络社区划分的研究相对不足. 二分网络中的单侧节点社区划分情况的研究, 仅集中在利用二分网络构建单部网络后, 使用已经成熟的算法进行社区划分. 或者使用二分网络社区划分算法将两类节点当作一类进行处理, 进行社区划分后再进行单侧社区节点的提取. 如现实二分网络科学家-论

文合作网络^[1-3]中的科学家的合作研究、电影-演员合作网络^[4]中演员的演员合作研究、疾病-基因网络^[5]中疾病关联性、听众-歌曲网络^[6]的听众社区和电脑终端-数据的 P2P 网络终端群体^[7]等.

1 国内外研究现状介绍

传统对二分网络的社区划分常用方式是把二分网络映射成单部网络, 利用已有的、相对成熟的单部二分网络社区划分算法进行处理. Guimera 等人证明不论是无权投影还是加权投影, 每次都只能对单侧节点进行划分, 而且投影结果都会因二分网络中信息的缺失

① 收稿时间:2015-04-09;收到修改稿时间:2015-06-11

导致结果不准确. 而直接在原始二分网络上的划分, 常常将二分网络中的两类节点当作一类进行处理, 到目前为止也没有一种公认的二分网络社区划分评价标准, 因此, 有必要对二分网络社区划分算法做进一步的研究.

目前关于二分网络社区划分算法的研究分为两类, 一类是基于 Newman 对单分网络上模块度的定义, 设计出了二分网络上 Q 值的定义, 如文献[8][9][11]的研究, 文献[8][14]提出了一种将二分网络社区挖掘转化成一个可供蚂蚁搜索的图模型优化问题, 通过蚁群搜索过程网络中在两个顶点间的路径上留下的信息量作为节点间的连接强度. 文献[9]通过对二分网络图所对应的矩阵进行递归拆分, 应用矩阵分解的办法进行社区划分, 但是要将每一次的矩阵拆分结果都保持完整的社区信息非常困难, 因此划分的结果会随着矩阵分解过程不同而不同. 文献[10]借鉴单分网络模块度的定义, 设计了一种基于网络结构的 F 值, 利用该值作为聚类算法合并结果优劣的依据. 文献[11]基于快速 Newman 算法的思想对 BRIM 算法进行了改进, 提出了模块度凝聚算法(MAB), 虽相比 BRIM 算法不需要额外给定社团个数, 但算法复杂度要比 BRIM 算法高很多. 文献[15]结合 Adaptive BRIM 算法最大化二分模块度来获得来进行社区划分, 但是要得到最优解需要经过社区中节点的交换尝试, 提前定义二分网络的社区个数. 因此算法复杂度较高, 应用具有局限性, 不适用于大型网络.

另一类算法是在文献[12]提出的标号传播方法(LPA 算法)的基础上做的研究, 该算法首先给各顶点分配一个唯一的标签, 之后每个顶点尽可能的选择它们相邻的标签, 最后通过迭代具有相同标签的顶点组成一个单独的社区. 后期 Murata 等人在文献[13]中以牺牲精度为代价使得 LPA 算法适合大的二分网络和大型二分网络的并行实时社区分析.

本文基于信息扩散概率思想, 结合 Barber 等人对于模块度的定义^[15], 提出了一种二分网络社区划分算法, 首先初始化两侧节点分配信息量. 依照小世界六度分离理论进行 6 次信息的概率扩散交换后得到一个信息资源矩阵. 将矩阵中的信息资源数值大小作为进行社区的合并判断依据. 结合 Barber 对于二分模块度的定义, 作为社区划分的标准. 在 Q 值最大时停着社区合并, 得到最优的社区划分结果.

2 信息扩散概率(IPS)模型

二分网络可以表示为一个二部图 $G=(U, V, E)$, G 中的节点分为 U 和 V 两部分, E 为图 G 中的边. 集合 U (或 V) 中节点之间没有边相连, 集合 E 中的任意边 (u_i, v_j) , 必有 $u_i \in U, v_j \in V$. 设计和 U 中节点数为 m , 集合 V 中的节点数为 n , 二部图的邻接矩阵可表述为:

$$A = \begin{bmatrix} 0_{m \times m} & A_{m \times n} \\ A_{n \times m}^T & 0_{n \times n} \end{bmatrix}$$

其中矩阵 $A_{m \times n}$ 为二部图邻接矩阵的子矩阵, 称矩阵 $A_{m \times n}$ 为二部图 G 的关系矩阵, 关系矩阵 $A_{m \times n}$ 有:

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ connect } j \quad i \in U, j \in V \\ 0 & \text{else} \end{cases}$$

关系矩阵 $A_{m \times n}$ 的每一行为集合 U 中一个节点的链接情况, 每一列为集合 V 中一个节点的链接情况. 初始化赋予集合 U 中的每个节点单位为 1 的信息量, 赋予集合 V 中各个节点单位为 0 的信息量. 得到集合 U 的初始单位信息量矩阵 $I_{m \times m}^U$:

$$I_{ij}^U = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

得到集合 U 的单位信息量矩阵如如上公式 3 所示, 单位信息矩阵的每一行或者每一列表示集合 U 中的一个节点拥有 U 集合中其他节点的信息量多少, 初始时节点只拥有自己单位为 1 的信息量, 此时可以用向量 $\bar{x}_i$ 表示节点拥有的集合 U 中其他节点的信息量, 向量 $\bar{x}_i$ 的值为信息矩阵第 i 行或 i 列的元素值. 集合 U 中每个节点所拥有的信息量 $\bar{x}_i$, 按照如式 4 所示的规则进行扩散.

$$R_i' = \begin{cases} \frac{1}{k_i} \bar{x}_i & \text{if } i \text{ connect } j \\ 0 & \text{else} \end{cases}$$

其中 k_i 为集合 U 中节点 u_i 的度数, 如果节点 u_i, v_j 有边相连, 此时节点 u_i 的信息就扩散到节点 v_j 中, 节点 v_j 将得到的各个信息量向量进行相加, 得到来自集合 U 中各个节点扩散过来的信息量 $R_i = \sum_{i=1}^m R_i'$, R_i 的组合可以得到扩散左矩阵 $R_{m \times n}$, 集合 U 中节点对集合 V 中节点的信息扩散矩阵表示为:

$$S_{m \times n} = I_{m \times m}^U \bullet R_{m \times n}$$

其中 $R_{m \times n}$ 每一行表示集合 U 中的一个节点扩散到集合 V 中每个节点的信息量, 每一列表示集合 V 中的一个

节点接受集合 U 中每个节点的信息量. 此时 V 集合中的每个节点就得到了来自 U 集合中每个节点的信息量.

最后将集合 V 中每个节点上的信息量向量, 按照公式 6 再扩散到集合 U 中每个节点上.

$$T_j = \begin{cases} \frac{1}{k_j} y_j & \text{if } i \text{ connect } j \\ 0 & \text{else} \end{cases}$$

其中 y_j 为节点 v_j 拥有的来自集合 U 中每个节点的信息量向量, k_j 为节点 v_j 的度数. 信息量向量 T_j 的组合就得到一个 $n \times m$ 信息扩散右矩阵 $T_{n \times m}$, 此时, 集合 U 中每个节点将自身的信息量扩散到集合 U 中其他节点的信息量矩阵可以表示为:

$$S_{m \times m}^1 = I_{m \times m}^U \bullet R_{m \times n} \bullet T_{n \times m} \quad (7)$$

其中矩阵 $R_{m \times n}$ 元素 R_{ij} 可由对应的关系矩阵 $A_{m \times n}$ 中元素 A_{ij} 除以对节点 i 的度数得到, $T_{n \times m}$ 为可由关系矩阵 A 的转置矩阵 A^T 每个元素 A_{ij}^T 除以对应节点 i 的度数得到, 因此, 循环上述步骤进行 n 次信息的扩散后得到:

$$S_{m \times m}^n = (I_{m \times m}^U \bullet R_{m \times n} \bullet T_{n \times m})^n \quad (8)$$

其中 S_{ij}^n 表示信息在循环扩散 n 次之后, 节点 u_i 收集到节点 u_j 的信息量, 根据六度空间理论, 此处 n 取值为 6.

3 模块度与算法流程分析

3.1 模块度定义

假设 M 是二分网络中总边数, V 是二分网络中的顶点集合. V^X 、 V^Y 是二分网络中两类不同的顶点集合, V_l^X, V_m^Y 分别对应 $V^X = (V_1^X, V_2^X, V_3^X, \dots)$ 、 $V^Y = (V_1^Y, V_2^Y, V_3^Y, \dots)$ 中的社区个体. 定义:

$$e_{lm} = \frac{1}{2m} \sum_{i \in V_l^X} \sum_{j \in V_m^Y} A_{ij}$$

从 e_{lm} 的定义看出, 其含义为网络社区 V_l^X 连接 V_m^Y 社区边的集合. 定义一个 $K \times K$ 的矩阵, a_l 是该矩阵的元素, 定义:

$$a_l = \frac{1}{2m} \sum_{i \in V_l^X} \sum_{j \in V^Y} A_{ij}$$

因此, 对于二部图 A 的邻接矩阵 $A_{m \times n}$, Murata 的模块度定义为:

$$Q = \sum_l (e_{lm} - a_l a_m), m = \arg \max(m)$$

其中 Q 值越大, 代表网络具有很强的社区结构.

3.2 算法过程

输入: 二分网络

输出: 二分网络单侧节点社区划分结果

Begin

- ① 初始化二分网络
- ② 进行 6 次信息传递
- ③ 选择矩阵中信息值最大的节点对进行合并, 更新合并后的矩阵(合并后对外的信息量为两个社区对外信息量中的最大值)
- ④ 按照合并顺序计算 Q 值
- ⑤ 选择 Q 值最大处进行社区划分

End

3.3 算法复杂度分析

本算法由: 能量扩散计算、单侧节点映射和 Q 值计算三部分. 在能量扩散过程中进行矩阵的乘法运算, 这一步的算法复杂度为 $O(n^2 + m^2)$, 其中 m 为二部图节点集合 U 中的节点个数, n 表示集合 V 中节点的个数. 在进行 6 次矩阵乘法计算之后算法复杂度为 $O(6(m^2 + n^2))$, 要遍历矩阵进行单侧节点的投影, 这一步的算法复杂度为 $O(mn)$, Q 值的计算复杂度约为 $O(m^2)$ 因此最终的算法复杂度为 $O(k_1 n^2 + k_2 m^2)$, 其中 k_1 和 k_2 为常数.

4 实验结果与分析

4.1 在 Southern Women 数据集上的测试

南非妇女网络(southern women)数据集^[18]是美国南部妇女参加活动行成的一个实际网络, 由 Stephen 提出, 该网络由 18 个妇女和 14 个活动组成, 编号 1-18 为妇女节点, 19-32 为活动节点. 如下图所示:

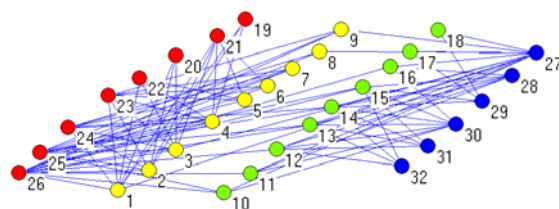


图 5 南非妇女-事件网络

实验结果显示该网络被划分成四个社区, 其中妇女 1-9 和 10-18 各被划分成一个社区, 19-26 和 27-32 被划分为另外一个社区, 这与 Davis1 原始的观察结果一致. 对该网络进行妇女网络和事件网络分别投影,

并且将其他的几个算法的划分结果与本算法的划分结果进行对比,用模块度作为衡量的标准,得到如表 1 所示的结果:

表 3 不同算法对妇女-事件网络划分结果与 Q 值比较

算法	社区划分情况	Q 值
IPS	{1-9, 19-26} {10-18, 27-32}	0.5860
BRIM	{1-6, 19-26} {7, 9, 10, 25-26} {11-15, 28, 30-32} {8, 16-18, 27, 29}	0.4750
LP&BRIM	{1-7, 9, 19-26} {8, 10-18, 27-32}	0.5777
ACODC	{1-9, 19-26} {10-18, 27-32}	0.5860

由图可见,比较各种算法的划分结果,IPS 算法在单侧网络中得到的划分结果模块度值最大.划分结果与基于蚁群优化的二分网络社区划分算法(ACODC 算法)相同;并且与 Davis 在原始文献中做的 Davis1 划分结果相同.因此,IPS 算法能够有效的识别二分网络的社区结构,并能得到准备的社区划分和较高质量的模块度.

5 结语

为了对二分网络进行社区划分,本文提出了一种基于信息扩散的二分网络社区划分算法.该算法可以在不需要额外参数的情况下,能够分别对二分网络进行精确的社区划分.但是该算法也存在着不足;主要是算法的复杂度与单部网络社区划分算法相比也有待进一步提高.因此,二分网络社区划分的研究在解决实际问题时还需要进一步探讨.

参考文献

- 汪小帆,李翔,陈关荣.复杂网络理论及其应用.北京:清华大学出版社,2006.
- Newman MEJ. Scientific collaboration networks.I. network construction and fundamental results. *Physical Review E*, 2001,64:016131
- Newman MEJ. Scientific collaboration networks. II. Shortest paths,weighted networks, and centrality. *Physical Review E*,2001,64:016132
- 刘爱芬,付春花,张增平等.中国大陆电影网络的实证统计研究. *复杂系统与复杂性科学*,2007,4(3):10-16.
- 陈文琴,陆君安,梁佳.疾病基因网络的二分图投影分析. *复杂系统与复杂科学*,2009,6(1): 13-19.
- Lambiotte R, Ausloos M. Uncovering collective listening habits and music genres in bipartite network. *Physical Review E*, 2005,72:066107.
- Le Blond S, Guillaume JL, Latapy M. Clustering in P2Pexchanges and consequences on performances. In: Castro M, Renesse R, eds. *Peer-to-Peer Systems IV*. Berlin: Heidelberg, 2005:193-204.
- 徐永成,陈峻.基于蚁群优化的二分网络社区挖掘. *计算机科学与探索*,2014,8(3):286-304.
- 陈伯伦,陈峻,邹盛荣等.基于矩阵分解的二分网络社区挖掘算法. *计算机科学*,2014,41(2):55-101.
- 吴亚晶,狄增如,樊瑛.基于资源分布矩阵的二分网络网聚类方法. *北京师范大学学报*,2010,46(5):643-646.
- 高明.二分网络社团发现算法研究.合肥:合肥工业大学, 2012.
- Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007,76:036106
- Liu X, Murata T. Community detection in large-scale bipartite newworks. *Proc. of the 2009IEEE/WIC/ACM International Joint conference on Web Intelligence and Intelligent Agent Technology(WI-IAT'09)*. Washington, DC, IEEE Computer Society, 2009:5-8.
- Dorigo M, Maniezzo V, Colomi A. Ant system:optimization by a colony of cooperating agents. *IEEE Trans. on Systems, Man, and Cybernetics: Part B*, 1996,26(1):29-41.
- Barber MJ. Modularity and community detection in bipartite networks. *Physical Review E*, 2007,76:066102.
- Newman M E J.Modularity and community structure in networks. *Proc. of the National Academy of Sciences of the United States of America*. 2006, 103(23): 8577-8582.
- Milgram S. The small world problem. *Psychology Today*, May 1967: 60-67.
- Davis A, Gardner BB, Gardner MR. Deep south. Chicago: University of Chicago Press, 1941.
- Scott J, Hughes M. The anatomy of Scottish capital: Scottish companies and Scottish capital, 1990-1979. London: Croom Helm, 1980.
- Fan CL,Yu L. 2nd International Conference on Mechatronic Sciences, Electric Engineering and Computer, (MEC2013) ShenYang, China, 2013