软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(12):3829−3845 [doi: 10.13328/j.cnki.jos.005593]

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

节点不对称转移概率的网络社区发现算法*

许平华, 胡文斌, 邱振宇, 聂 聪, 唐传慧, 高 旷, 刘中舟

(武汉大学 计算机学院,湖北 武汉 430072)

通讯作者: 胡文斌, E-mail: hwb@whu.edu.cn

摘 要: 社区发现是当前社会网络研究领域的一个热点和难点,现有的研究方法包括:(1) 优化以网络拓扑结构为基础的社区质量指标;(2) 评估节点间的相似性并进行聚类;(3) 根据特定网络设计相应的社区模型等.这些方法存在如下问题:(1) 通用性不高,难以同时在无向网络和有向网络上发挥出好的效果;(2) 无法充分利用网络的结构信息,在真实数据集上表现不佳.针对上述问题,提出一种基于节点不对称转移概率的网络社区发现算法 CDATP.该算法通过分析网络拓扑结构来设计节点转移概率,并使用 random walk 方法评估节点对网络社区的重要性.最后,以重要性较高的节点作为核心构造网络社区.与现有的基于 random walk 的方法不同,CDATP 为网络中节点设计的转移概率具有不对称性,并只通过节点局部转移来评估节点对社区的重要程度.通过大量仿真实验表明,CDATP 在人工模拟数据集和真实数据集上均比其他最新算法有更好的表现.

关键词: 复杂网络;社区结构;社区发现;随机游走;核心系数

中图法分类号: TP311

中文引用格式: 许平华,胡文斌,邱振宇,聂聪,唐传慧,高旷,刘中舟.节点不对称转移概率的网络社区发现算法.软件学报,2019,30(12):3829−3845. http://www.jos.org.cn/1000-9825/5593.htm

英文引用格式: Xu PH, Hu WB, Qiu ZY, Nie C, Tang CH, Gao K, Liu ZZ. A community detection algorithm based on asymmetric transition probability of nodes. Ruan Jian Xue Bao/Journal of Software, 2019,30(12):3829−3845 (in Chinese). http://www.jos.org.cn/1000-9825/5593.htm

Community Detection Algorithm Based on Asymmetric Transition Probability of Nodes

XU Ping-Hua, HU Wen-Bin, QIU Zhen-Yu, NIE Cong, TANG Chuan-Hui, GAO Kuang, LIU Zhong-Zhou

(School of Computer Science, Wuhan University, Wuhan 430072, China)

Abstract: Community detection is a popular and difficult problem in the field of social network analysis. Most of the current researches mainly focus on optimizing the modularity index, evaluating the similarity of nodes, and designing different models to fit particular networks. These approaches usually suffer from following problems: (1) just a few of them can deal with directed networks as well as undirected networks; and (2) real-world networks being more complex than synthetic networks, many community detection strategies cannot perform well in real-world networks. To solve these problems, this paper presents an algorithm for community detection in complex networks based on random walk method. Different from existing methods based on random walk method, the asymmetric transition probability is designed for the nodes according to network topology and other information. The event propagation law is also applied to the evaluation of nodes importance. The algorithm CDATP performs well on both real-world networks and synthetic networks.

Key words: complex networks; community structure; community detection; random walk; core index

社会网络中的社区由网络中一定数量的节点组成,其内部有着较为紧密的结构.研究网络中的社区结构可以帮助分析复杂网络、预测社会网络的发展趋势,而且在广告投放和作弊用户检测等实际场景中得到应用.

* 基金项目: 国家自然科学基金(61711530238, 61572369); 国家重点基础研究发展计划(973)(2012CB719905)

Foundation item: National Natural Science Foundation of China (61711530238, 61572369); National Program on Key Basic Research Project of China (973) (2012CB719905)

收稿时间: 2017-09-04; 修改时间: 2017-11-14, 2018-02-03; 采用时间: 2018-04-24

相关文献中已经有很多种社区发现方法被提出,其中一类是优化与图的拓扑结构相关联的社区质量指标,例如由 Newman 等人[1]提出的模块度.基于优化指标数值来获得更加可靠的社区结构的这一思路,有很多学者提出了相关的社区发现算法,其中较为典型的有优化变体模块度的 BiLPA 算法[2]、优化结构密度的 IsoFdp 算法[3]和对混合指标进行优化的 EFA 算法[4]、MOCD-PSO 算法[5]等.这些算法一般是通过相应的迭代步骤来更新需要优化的指标,并在最后输出最优指标对应的社区结构.这类算法的优点是实现简单,在人工构造的网络上可以发挥出很好的效果.然而真实世界的网络要比人工构造的网络复杂许多,很多时候真实社区结构对应的质量指标并不是最优的,导致了上述基于指标优化的算法难以正确地检测到社区.同时,由于上述部分算法是基于全局拓扑结构来进行优化,因此会受到分辨率极限[6]的限制.并且,某些并不具有明显社区结构的网络同样会具有很高的质量指标,例如某些树或类树结构[7].因此,这些缺陷都在一定程度上限制了上述方法的应用场景.

另外有一些学者从节点相似性的角度提出了基于 random walk 的社区发现方法[8−16],这类方法以马尔可夫模型为理论基础,通常是通过节点的随机转移来评估节点的相似度,并将相似度较高的节点划分到同一社区中.在真实网络中,同一社区质量指标与不同类型的社区结构的匹配程度变化较大,由此可能会导致基于社区质量指标的社区发现算法适应性较弱,而基于 random walk 的算法受社区类型的影响较小,具有更好的适应性.但是,基于“利用 random walk 来评价节点相似度”这一思路的社区发现算法对游走过程的迭代次数非常敏感,往往需要先验知识来辅助决策.

将现实世界中的复杂系统抽象为图论中的网络虽然便于研究工作的开展,但也不可避免地会遗漏掉一些重要的信息.例如,Reddit 中属于同一兴趣组的用户往往有着相同的兴趣标签,若能将属性信息转化为网络的一部分,可能会使得社区内部的结构更加紧密,也能使处于社区边缘位置上的节点有更大概率被划分到正确的社区中.然而,现有的仅从网络结构层面划分社区的算法无法利用节点的属性信息.

基于以上分析,本文提出了一种可用于无向和有向网络的社区发现算法 CDATP(community detection algorithm based on asymmetric transition probability of nodes),此算法可以将节点的属性转化为拓扑结构的一部分,并且受到事件在网络中的传播规律的启发,根据网络的拓扑结构计算每一节点向邻接节点转移的概率,以带有限制的 random walk 来模拟逆向的事件传播过程,并以此为基础,评估节点在社区中的重要程度(核心系数).在聚类时,无需预先指定社区的数目,节点会根据转移概率等参数向所属社区转移.本文的主要工作可以总结如下:

(1) 充分利用了网络拓扑结构信息,为节点设计了不对称的转移概率,能够反映节点间的不对等关系;
(2) 参考了事件在网络中传播的规律,提出一种基于 random walk 且具有固定转移步长的方法来评估节点对于社区的重要程度,基于该重要程度指标的聚类不需要预先指定社区数目.

本文第 1 节介绍相关研究工作.第 2 节详细介绍 CDATP 算法.第 3 节为实验和分析.第 4 节是总结与展望.

1 相关工作

近年来,普遍存在于网络中的社区结构已经受到了国内外学者的广泛关注.关于社区发现的研究也已经被应用到了许多领域中,并取得了不错的成果.

基于社区质量指标来划分社区是一类很经典的方法,这类方法通常先定义一个基于网络拓扑结构的社区质量指标,若一种社区划分与预先定义的质量指标的“含义”越接近(如社区内部的边数远多于社区边缘上的边数),那么该社区划分的得分越高.这类方法的优点是思路简洁明了,在确定了质量指标后只需通过一系列迭代运算来搜索最优社区划分.但同时,如何确定社区质量指标也成了最大的问题.因为若一个社区质量指标不能较好地体现真实社区的“含义”,或者说一种社区划分得分很高但却与真实社区结构相差甚远,那么基于该质量指标的后续工作都将变得没有意义.经过长时间的研究,学者们提出了一些表现较好的经典社区质量指标,包括结构密度和模块度等.结构密度的大小和社区内部边的数量与社区内部节点的数量比值相关,一般来说,结构较为紧密的社区对应的结构密度较大.You 等人[3]提出的 IsoFdp 算法将网络数据映射到低维空间并自动找出社区的中心节点,然后再以中心节点为基础建立社区,并通过对结构密度的优化来搜索更好的社区划分.相较于传统的

基于结构密度的社区发现算法,IsoFdp 的最大优势是可以自动识别出社区的中心节点,有较好的可行性.模块度的内涵是评价人工社区划分与随机社区划分的差异性.Newman 等人[17]提出的 FastQ 算法是最为经典的基于模块度的社区发现算法,该算法以贪心策略进行分层聚类,其优点是收敛速度快,可以在较短时间内找到模块度最大的社区划分.虽然很多基于社区质量指标的社区发现算法在人工网络上有很好的表现,但在处理真实网络时,容易产生时好时坏的结果.这是因为真实网络的度分布相较于人工网络随机性较大,社区质量指标有时与真实社区结构不匹配[18].为了减少这种不匹配带来的问题,一些学者开始研究更小粒度的质量指标.Bai 等人[19]提出的 ISCD+算法中定义了针对节点的质量指标,包含节点的局部重要性和全局重要性.与传统的社区质量指标不同,该指标并非是基于社区划分,而是基于原始网络中的每一个节点.实验结果表明:在一些真实网络上,该算法的表现优于部分基于社区质量指标的算法.受到现有的各类质量指标的启发,本文尝试定义一种新的基于节点的质量指标,并希望该指标能够较好地适应不同类型的真实网络.

将基于马尔可夫模型的 random walk 用于社区发现也是较为主流的研究方法之一,其主要思想是以一个初始分布释放大量的无规则行走者,在扩散过程之后,可以得到行走者的分布函数.通过一系列研究,国内外学者提出了若干种基于 random walk 的社区发现算法.Pons 等人[8]提出的 Walktrap 算法是最早的基于 random walk 的方法之一,该算法的主要思想类似于“在社区结构中,节点间有着更多的联系,而不同社区间的联系则相对较少.因此,一个随机选择方向的行走者将会被更长时间地困在社区内部[20]”.Walktrap 算法采用了分层聚类的方式发现社区,得到的社区有很清晰的层级结构.虽然 Walktrap 算法在准确度上有所欠缺,但该算法的思路对于后来的同类型算法有很强的启发作用.Lai 等人[9]通过将边的方向信息转化为边的权重实现了将有向网络转化为无向网络,并进一步通过基于无向网络的算法来发现原来的有向网络中的社区.该算法处理边的方向信息的方式很新颖,且在基于有向网络的社区发现问题中取得了不错的效果,但转化过程计算量较大,且不适用于基于无权值网络算法的拓展.Jiao 等人[10]综合考虑了全局拓扑结构和局部拓扑结构,并在此基础上提出了新的节点相似度计算方法,与以往算法相比,对不同类型社区的适应性更强.Huang 等人[11]提出的 SCMAG 算法通过计算节点属性相似度来从节点属性的角度构建社区,并证明了节点属性与社区结构之间存在密切的联系,属于同一社区的节点往往具有相似的属性.本文受其启发,尝试以将节点属性转化为拓扑结构信息的方式来处理节点属性.此外,文献[12−16]各自提出了将 random walk 与其他经典方法进行融合得到的社区发现算法,在一些特定的网络模型上有较好的表现.然而,上述算法在节点转移的步骤中多采用的是无差别转移概率,不能完全反映真实网络中节点关系的不对称性,且社区划分的准确性受转移迭代次数影响较大,需要较多的先验知识来辅助决策;同时,本文认为,由于 random walk 在模拟随机过程等方面具备一些优良的特性,可以将其作相应改进后用于评价节点的质量指标,而在目前的工作中,尚未发现有人进行过这样的尝试.

在社区发现领域的研究初期,大部分学者都是以无向网络作为研究对象.但随着社区发现技术在实际生产情景中的应用推广,有越来越多的学者开始关注如何在有向网络中发现社区.早期的一种处理方法是忽略掉边的方向,将其直接作为无向网络来处理.例如,将经典的基于无向网络的 LP 算法[21]直接用于忽略了边的方向的有向网络,在某些情况下仍有不错的准确率,可用作对有向网络算法测试的基线.但文献[22]中指出:边的方向应该被考虑,否则会使得网络的重要特征丢失.其中一个重要原因就是:当忽略了边的方向后,节点间的相互关系将变得不完整.例如,Twitter 中的某一用户单方面关注了另一用户,那他们之间的位置是不平等的,但无向边无法描述这种关系.一部分学者根据有向网络的特征重新设计了质量指标,例如,Newman 等人[23]提出了有向版本的模块度,在原来的模块度定义的基础上考虑了边的方向信息,是最早的基于有向网络的社区质量指标之一.Rosvall 等人[24]基于信息论提出了 Infomap 算法,该算法根据网络的拓扑结构来预测数据的流动,然后对其进行信息编码,而平均长度最短的编码方式就对应了最优的社区划分.得益于其简洁而优美的设计思路,Infomap 算法可被用于处理各种不同类型的网络,且均有不错的表现.Lancichinetti 等人[25]提出的 OSLOM 算法是另一个非常经典的可用于有向网络的社区发现算法,它使用了一个基于簇的质量指标来评价人工划分得到的簇与随机生成的簇之间的差异,并通过局部优化来找到得分较高的簇,最后基于这些簇来生成社区.Lancichinetti 等人对网络中的度分布有较深入的研究,并且开发了基于幂分布的基准网络[26]用于检验社区发现算法的准确性,而

OSLOM 算法也在人工网络上有非常好的表现.Santos 等人[27]基于改进后的模块度提出了 ConClus 算法,该算法规避了传统的模块度优化算法常常会碰到的分辨率极限问题.实验结果表明,ConClus 在人工有向网络上有与 OSLOM 十分相近的表现,且在真实网络上也有不错的表现.上述的可用于有向网络的社区算法一般在人工网络上有较高的准确率,但在一些真实网络上,其准确率仍有较大的提升空间.因此,提高算法在真实网络上的准确率也是本文尝试去实现的目标之一.

受到现有的社区发现算法的优势及缺陷的启发,本文参考了网络中的事件传播规律,提出一种基于 random walk 的方法来评价节点对社区的重要程度.与现有的基于 random walk 的节点相似度评价方法不同,本文设计了基于拓扑结构的不对称节点转移概率,并尝试从模拟事件传播的角度来评价节点对社区的重要程度而非节点相似度,且转移过程具有固定的步长,不需要额外的先验知识的辅助.最后,本文在该评价方法的基础上提出了一种不需要预先指定社区数量,且在真实数据集和人工数据集上都能有较好表现的网络社区发现算法.

2 社区发现算法 CDATP

CDATP 算法基于 random walk 方法来计算节点属性相似性、节点间影响力,并以此为基准评估节点对社区重要程度,最后使节点向社区核心靠拢来划分社区.

为使下文的描述简洁,本文将研究对象的相关重要概念进行符号化定义,见表 1.

Table 1 Symbols and their remarks

表 1 相关符号及其注释

符号	注释	符号	注释
G	有向或无向网络	*Affect*	属性的结构影响度
W	邻接矩阵	*Force*	影响力矩阵
N	G 包含的节点总数	P	转移概率矩阵
$Attr^G$	G 的属性空间	*Core*	核心系数矩阵
Y	属性信息矩阵	*Dir*	转移方向矩阵
W^{Attr}	增量邻接矩阵		

2.1 整体框架

图 1 描述了 CDATP 进行社区检测的整体框架,输入数据集包括社交网络等复杂社会网络,输出结果为社区序列.

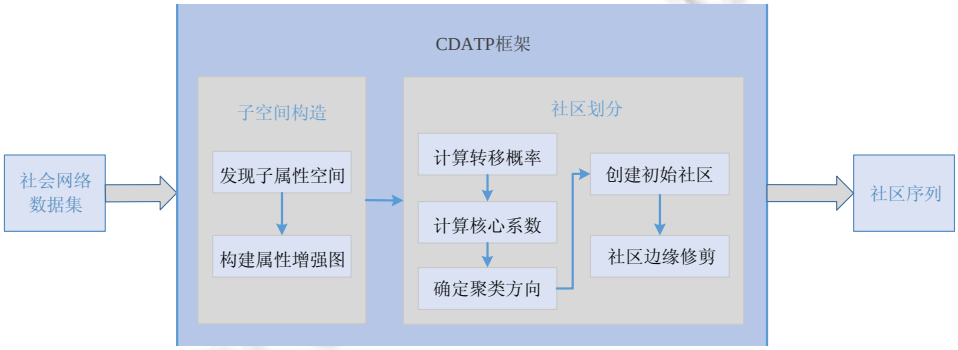


Fig.1 Framework of CDATP

图 1 CDATP 框架

框架包含以下两个部分.

(1) 在子空间构造阶段中找到表现最好的子属性空间,并将相应的属性转化为网络中的虚拟节点,构造属性增强网络;

(2) 在社区划分阶段中,以属性增强网络为对象计算节点转移概率,使用 random walk 方法评估节点核心

系数,在此基础上确定每个节点的聚类方向,创建初始社区,再进行边缘修剪,最终输出社区序列 *Comms*.

CDATP 的描述见算法 1.

算法 1. CDATP.

输入:网络 G 的邻接矩阵 W 和属性空间 $Attr^G$;

输出:社区序列 *CommS*.

1. 子空间构造(见第 2.2 节)
   1.1. 发现子属性空间
   1.2. 构造属性增强网络
2. 社区划分(见第 2.3 节)
   2.1. 计算节点转移概率
   2.2. 计算节点核心系数
   2.3. 确定节点的聚类方向
   2.4. 创建初始社区
   2.5. 社区边缘修剪
   2.6. 输出社区序列 *CommS*

2.2 子空间构造和属性增强图

“物以类聚,人以群分”.人们通常会依其兴趣爱好、工作内容来发展社交圈,而各种物件也能被按照其特性、功能划分类别,属性是对象间建立起联系的重要“桥梁”.与研究高度抽象的网络不同,在研究现实世界中更为复杂的情景时,若忽略了节点本身具备的属性,很可能就会错过一些重要的信息.属于同一社区的节点往往拥有某些相近甚至相同的属性值,在考虑这些属性后,能更科学地度量节点间联系的强弱,使得社区的边界变得更加清晰,同时还能够从节点属性的角度帮助挖掘社区结构形成的原因.

为了度量属性对节点间联系的影响,本文采用了将属性转化为网络中的虚拟节点来构造属性增强网络的方法.对于属性 A_i,若离散化后有 $Dom(A_i)=\{a_1,a_2,\ldots,a_k\}$,则在图中加入 k 个虚拟节点,与 A_i 的取值一一对应,并在原网络节点与其对应属性值的虚拟节点间建立一条双向边,如图 2 所示,拥有相同属性值的节点以虚拟节点为中介(虚拟节点可视为边的一部分而非独立的节点)产生了新的联系.

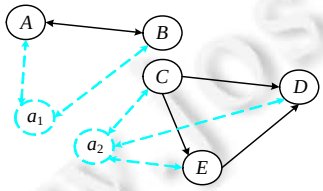
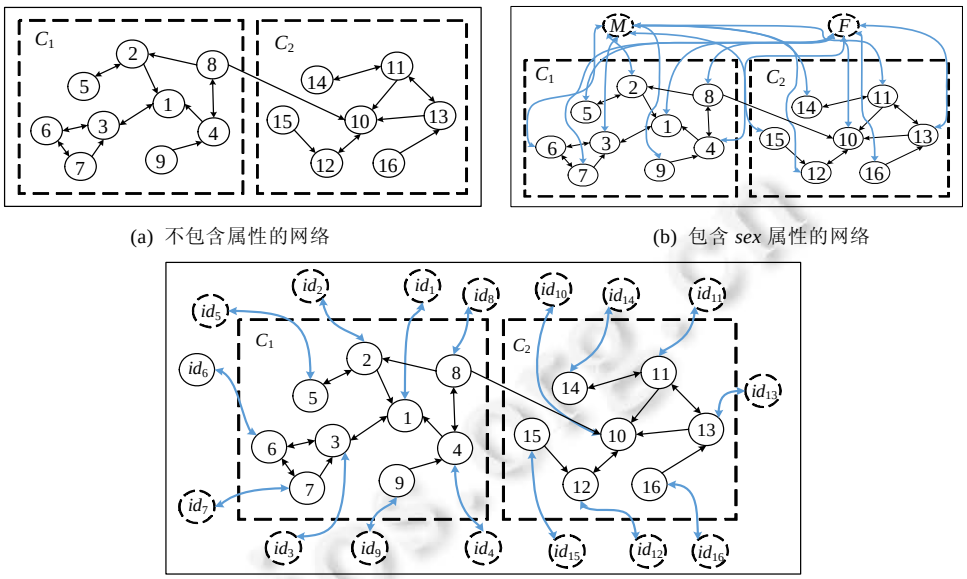


Fig.2 Virtual nodes

图 2 虚拟节点

但是,并非所有属性都是有价值的.将节点的所有属性直接加入计算不仅会降低计算效率,甚至可能因为在不同社区间产生了过多联系而导致社区检测准确率下降.现在先考虑使用单个属性.如图 3(a)所示,网络中存在两个社区 C_1 和 C_2,C_1 和 C_2 之间联系非常少.若考虑描述对象性别的二元属性 *sex*,则 C_1 和 C_2 之间 *sex* 值相同的节点(假定有这样的节点对存在)间就会产生联系,如图 3(b)所示,这种联系的数量是很多的,破坏了 C_1 和 C_2 较为独立的状态,对社区划分的结果产生负面影响;若考虑描述对象身份证号码的属性 *ID*,则由于每个节点的 *ID* 的值都不一样,如图 3(c)所示,节点间不会产生新的联系,因此对社区的划分同样没有帮助.



(a) 不包含属性的网络 (b) 包含 *sex* 属性的网络

(c) 包含 ID 属性的网络

Fig.3 An example of an attribute enhancement network

图 3 属性增强网络的示例

因此,需要选择合适的属性,这种属性应当满足下面两个条件.

(1) 在结构上联系紧密的节点在该属性上有相同属性值的概率较大;

(2) 在考虑该属性之后,不同社区间应尽可能少地产生新的联系.

而且,只有一个属性相同往往不能说明节点间就有很强的联系,考虑的属性数目越多,则属性值相同的偶然性越小,考虑包含多个属性的子属性空间相较于考虑单个属性更加可靠.将子属性空间中的所有属性看作一个复合属性,它同样应该满足前面提到的两个条件.

信息熵可以理解为特定信息的出现概率.对于属性 A_i,若 $Dom(A_i)=\{a_1,a_2,\ldots,a_k\}$,且对应属性值为 a_i 的节点个数为 n_i,则其信息熵 $H(A_i)$可用公式(1)计算:

$$H(A_i)=-\sum_{i=1}^{k}\frac{n_i}{N}\log\frac{n_i}{N} \tag{1}$$

由于不存在重复的取值,所以前文中提到 ID 属性的信息熵非常大,而这样的属性是没有意义的,因此不应将信息熵超过阈值 *ht* 的属性加入子属性空间.

本文构造了属性信息矩阵 $Y=(y_{ij})_{N\times N}$ 来描述虚拟节点对原节点间关系的影响,若节点 v_i 与节点 v_j 具有相同的属性值,则 $y_{ij}=1$;否则 $y_{ij}=0$.另外,本文构造了增量邻接矩阵 $W^{Attr}=(w_{ij}^{Attr})_{N\times N}$ 来描述加入了虚拟节点后新的网络拓扑结构,若 y_{ij} 与 w_{ij} 之和不为 0,则 $w_{ij}^{Attr}=1$;否则, $w_{ij}^{Attr}=0$.

在此基础上,本文定义了属性的结构影响度 $Affect(A_i)$.属性的结构影响度应能反映在将属性信息转化为拓扑结构信息后,原节点间的联系受到了多大的影响.$Affect(A_i)$可由公式(2)计算得到:

$$Affect(A_i)=\frac{\sum_i\sum_j(W^{attr^2})_{ij}-\sum_i\sum_j(W^2)_{ij}}{N^{\alpha_A}} \tag{2}$$

其中,α_A 是矩阵缩放因子,旨在更加明显地区分属性的结构影响度.

子属性空间应同时满足信息熵较小和结构影响度较小的条件,其构造步骤如下.

(1) 计算每个属性的信息熵,并筛除信息熵大于阈值 *ht* 的属性;

(2) 计算剩余属性的结构影响度,并选择结构影响度小于阈值 *at* 且信息熵最小的属性加入子属性空间;

(3) 将剩余属性按信息熵从小到大排序,若其中的属性加入子属性空间后,使得子属性空间的信息熵和结构影响度均小于阈值,则将其加入子属性空间.

2.3 聚类方向和社区划分

由于缺少先验知识,需要预先指定社区数目的聚类方法往往难以在社区发现中取得良好效果.为了能够得到更加准确的社区结构,本文提出了一种自动确定社区的核心,并使核心以外的节点按照各自的聚类方向向核心靠拢的聚类方法,由此得到的社区不仅内部聚合度高,并且有着很清晰的层次结构,便于对社区中的事件传播过程做进一步的研究.

文献[28]指出,一个节点的状态有一定概率会因其邻接节点的行为而发生改变.若一个光标可从一个节点向它的任意邻接节点转移,且倾向于向有更大概率会对其状态产生影响的节点转移,那么在进行一定次数的转移后,光标会有较大的概率落到事件传播流中原节点的上游位置.而由于事件在传播过程中,其影响力会呈现衰退的趋势,本文参考了文献[28]中的实验结果,设定光标在寻找上游的过程中,将转移 2 次.

本文构造了节点影响力 *force* 的概念来描述任意节点的邻接节点会对其产生的影响,并假定该影响是由拓扑结构中的出链与入链以及节点间相同的属性产生的.

记 $force_{ij}$ 为节点 v_j 对节点 v_i 的影响力, $force_{ij} = f_{ij}^{out} + f_{ij}^{in} + f_{ij}^{attr}$,其中, f_{ij}^{out} 是由节点 v_i 指向节点 v_j 的出链产生的影响力, f_{ij}^{in} 是由节点 v_j 指向节点 v_i 的入链产生的影响力, f_{ij}^{attr} 是由属性关系产生的影响力.

以微博为例,如图 4 所示,边的宽度代表影响力的大小,关注关系和粉丝关系可由网络中的有向边表示,属性关系可由原节点与虚拟节点的联系表示.若用户 *A* 关注了用户 *B*,说明 *B* 产生的内容对 *A* 有一定的的影响力.*A* 接收到的内容信息量与其关注的总人数有关,人数越多,信息量越大,*B* 产生的内容的信息量占比就小,对 *A* 产生的吸引力也会变弱;相反地,当 *A* 关注人数较少时,*B* 对 *A* 的影响力会更强.同时,由于 *A* 成为了 *B* 的粉丝,*B* 因为这种关系,也会在一定程度上被 *A* 产生的内容吸引,同样的,影响力的强弱与 *B* 的粉丝数有一定关系,不过这种影响力的大小远小于由关注关系产生的影响力大小.此外,由属性产生的新关系同样会作用于影响力,本文假定其大小介于前两者之间.子属性空间包含的属性数目越多,则 *A* 与 *B* 拥有相同属性值就越不可能是偶然发生的,由属性产生的影响力就越大.类似的,在子属性空间下,拥有相同属性值的节点越多,则该子属性空间越有可能与社区特征相关,由此产生的影响力也越大.本文进行了大量的预实验,试图找到最能体现节点间关系的影响力模型.由于篇幅的限制,本文不对建立影响力模型过程中的相关预实验作展开说明.

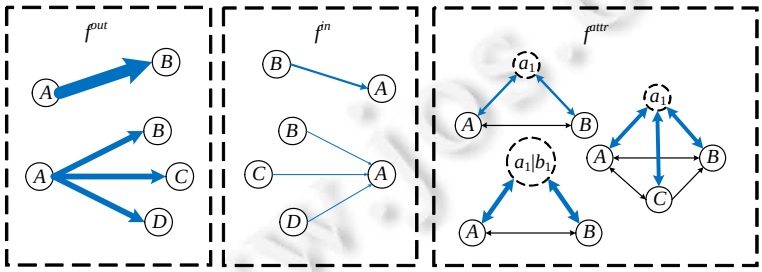


Fig.4 Influence of outbound link, inbound link and attribute

图 4 出链、入链、属性与影响力的关系

根据预实验的结果,本文使用了以自然常数 *e* 为底的指数函数来描述 f_{ij}^{out} 和 f_{ij}^{in} 的变化,它们的计算公式分别为公式(3)和公式(4),其中, α_{out} 和 α_{in} 分别为出链和入链系数,用于控制函数的收敛速度:

$$f_{ij}^{out} = \begin{cases} 0, & w_{ij} = 0 \\ 1 + e^{(1-\alpha_{out}\sum_k w_{ik})}, & w_{ij} = 1 \end{cases} \tag{3}$$

$$f_{ij}^{in}=\begin{cases}0, & w_{ji}=0\\ (e/2)^{1-\alpha_{in}\sum_k w_{ki}}, & w_{ji}=1\end{cases} \tag{4}$$

f_{ij}^{attr} 可由公式(5)计算得到,其大小和子空间包含的属性数量及属性对应的节点数量呈正相关关系.其中,M'为子属性空间包含的属性的数目,α_{Attr} 为属性系数,用于控制函数的收敛速度:

$$f_{ij}^{attr}=\begin{cases}0, & y_{ij}=0\\ \dfrac{(1-1.1^{1-\alpha_{attr}\sum_k y_{ik}})}{2/(1+\log M')}, & y_{ij}=1\end{cases} \tag{5}$$

得到影响力矩阵后,按公式(6)将矩阵每一行归一化,即得到转移概率矩阵 $P=(p_{ij})_{N\times N}$,p_{ij} 为光标从节点 v_i 向节点 v_j 转移的概率:

$$p_{ij}=\frac{f_{ij}}{\sum_k f_{ik}} \tag{6}$$

在现实世界中,无论是蚁群还是 Reddit 的兴趣组,社区中的成员都并非完全平等,而是存在金字塔式的成员结构.受到社区中不平等现象的启发,为了评价节点在社区中是否处于金字塔顶端位置,本文引入了节点核心系数 $Core=(core_1,core_2,\ldots,core_N)^T$,节点 v_i 的核心系数 $core_i$ 越大,就越有可能成为社区的核心.现在介绍节点核心系数的计算方法.假设一个光标依次从网络中各个节点出发,按照 P 中的转移概率随机选择下一次转移的目标节点,每次出发后共转移 2 次,取最后所在节点为终点.每个节点的核心系数 $core_i$ 即是该节点为终点节点的次数期望.同时,与一些经典的基于 random walk 的方法一样,本文设定光标在每次转移后有 *back* 几率退回转移前的节点.那些经典的基于 random walk 的方法加入参数 *back* 通常是为了避免光标的转移在进入某些特殊路径后陷入死循环,但本文提出方法的转移次数固定且较小,几乎不受这些特殊路径的影响,加入参数 *back* 是为了能更好地模拟数据的传播.设想某人在浏览网络论坛时常常会因为对当前页面的内容不感兴趣而回退到上一级页面,而类似的回退操作在其他场景中也时有发生,本文加入的参数 *back* 即蕴含了这类回退操作的含义.

表 2 为当 *back*=20%时,图 3(a)中节点的核心系数.可以看出,处于社区中心位置的节点往往具有非常高的核心系数,这与社区中的金字塔结构也是相对应的.

在核心系数的基础上,本文设计了一种不需要预先指定社区数目的聚类方法,在网络中,每个节点都会按照该方法确定其聚类方向并和对应的节点聚合到一起.

节点的聚类方向 $Dir=(dir_1,dir_2,\ldots,dir_N)^T$ 由转移概率和核心系数共同决定,若 p_{ij} 为转移概率矩阵 P 第 i 行的唯一最大值,则节点 v_i 的聚类方向为 $dir_i=j$;若等于最大值的元素有多个,则取其中核心系数最大的作为聚类方向.

Table 2 Sample of core index

表 2 节点核心系数示例

节点编号	1	2	3	4	5	6	7	8
核心系数	**1.937**	0.907	1.632	0.617	1.247	1.047	0.604	0.894
节点编号	9	10	11	12	13	14	15	16
核心系数	0.159	**2.148**	1.443	1.367	1.100	0.512	0.317	0.130

构建一个将原网络中边的信息全部去除后的新网络,再按照下面的步骤添加边.

(1) 从节点序列中取出核心系数最大的节点 v_i,若 v_i 未与任何边相连,则将该节点作为新社区的中心,并将其聚类方向记为空;

(2) 扫描 Dir,若有节点 v_j 未与任何边相连且聚类方向 $dir_j=i$,则建立一条由节点 v_j 指向节点 v_i 的有向边;

(3) 重复上述步骤,直到没有度为 0 的节点.

这样就得到了初始社区,下面再继续进行边缘修剪工作.

(1) 对于网络中的每一个节点,计算其所属社区中它的邻接节点的核心系数之和,再计算其他社区中它的

邻接节点的核心系数之和,并将最大的核心系数之和对应的社区标记为节点新的所属社区,但暂时不将节点划入到新的社区当中;

(2) 在得到所有节点对应的新社区后,将所有节点划入新社区当中,若有节点的所属社区发生了改变,则重复(1)的工作,否则停止.

图 5 描述了图 3(a)网络的社区发现过程.在新的网络中,任意条边两端的节点都属于同一社区.

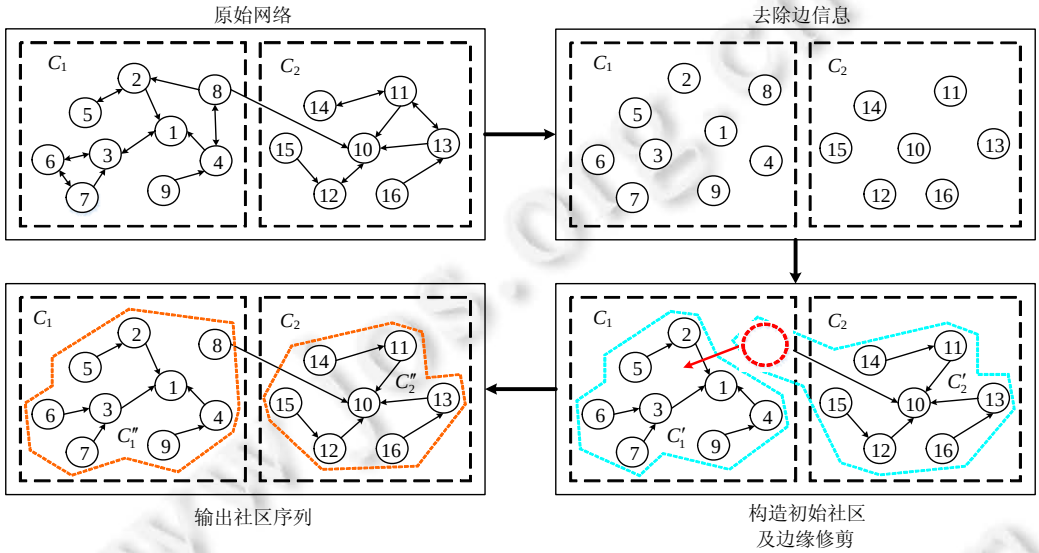


Fig.5 Example of community partitioning process

图 5 社区划分过程示例

3 实验分析

本节通过在多个社会网络数据集上的实验来验证 CDATP 算法的有效性.第 3.1 节为实验准备部分,介绍了实验中使用的各类型的数据集、对比算法、CDATP 的参数设置和实验结果评价标准.本文将实验按照使用到的数据集类型分为两部分,其中,基于无向网络的实验将在第 3.2 节中详细介绍,基于有向网络的实验将在第 3.3 节中详细介绍.

3.1 实验准备

在第 3.2 节基于无向网络的实验中,本文共使用了 Karate[29],Dolphins[30],PolBooks[31],PolBlogs[32]这 4 个带基准的真实网络数据集(http://networkdata.ics.uci.edu/),表 3 介绍了这些数据集的相关特征信息.实验选择 ISCD+,ROCONA[33],InfoMap 和 FastQ 算法作为对比算法.其中,ROCONA 算法是基于信息粒度观点的最新社区发现算法,通过节点之间的关联度来构建社区.ROCONA 算法通过与其他对比算法不同的视角来发现网络中的社区,且有较好的表现,因此本文也将其作为对比算法.

Table 3 Datasets of undirected network

表 3 无向网络数据集

数据集	节点数	边数	社区数
Karate	34	78	2
Dolphins	62	159	2
PolBooks	105	441	3
PolBlogs	1 490	19 090	2

在第 3.3 节基于有向网络的实验中,本文使用了有向版本的 PolBlogs 数据集和 LFR 基准网络[26].构造 LFR

基准网络使用的参数见表 4,其中:N 表示节点总数;μ表示社区间边数与内部边数的比值,μ的值越大,则社区结构越模糊;k 表示社区内平均节点度;max k 表示社区内最大节点度;t_1 表示度序列的负指数;t_2 表示社区规模分布的负指数;min s 表示社区节点数下限;max s 表示社区节点数上限.对于μ和 N 以外的参数,本文使用了文献[26]中的推荐取值.另外,本文参考了现有工作中对参数μ和 N 的设置[27],对于每一种μ和 N 的组合(对应不同环境下的网络)都生成 5 个网络,累计起来一共是 120 个人工网络.实验选择 ConClus,OSLOM 和 LP 作为对比算法.

Table 4 Parameters of LFR

表 4 LFR 参数

参数名	N	μ	k	max k	t_1	t_2	min s	max s
数值	{1000,2000,3000,4000}	{0.1,0.2,0.3,0.4,0.5,0.6}	20	50	2	1	20	100

在第 3.2 节和第 3.3 节中所使用的数据集均是带有基准信息的,因此本文使用 NMI[34]评价实验结果.*NMI*=1 时,说明实验结果与基准信息完全一样.*NMI* 的值越大,说明社区划分的准确度越高.公式(7)是 *NMI* 的计算公式:

$$NMI(X,Y)=\frac{2\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}}{-\sum_{x\in X}x\log p(x)-\sum_{y\in Y}y\log p(y)} \tag{7}$$

表 5 中介绍了 CDATP 的参数设置.在预实验中,本文发现:在较小规模(包含的节点数小于 5 000)的网络中,与 *force* 相关的 3 个收敛因子取表 5 给出的预设值时准确度较高,且在预设值附近的小范围变动对社区划分的准确度影响非常小.受篇幅限制,本文在后面的实验中对收敛因子的取值不作展开讨论.参数 *back* 对实验的准确度有一定影响,在第 3.2 节和第 3.3 节中,本文针对不同的 *back* 取值进行了实验并对实验结果进行了比较.

Table 5 Parameters of CDATP

表 5 CDATP 参数

参数名	*back*	α_{out}	α_{in}	α_{attr}	α_A	*ht*	*at*
数值	{0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}	0.02	0.02	0.05	1.2	1.5	5.0

3.2 基于无向网络的CDATP有效性验证

Karate 数据集是 Zachary 对一个美国大学空手道俱乐部进行了 2 年观察而构建出的一个社会网络,它被广泛应用于社区检测方法的测试.网络中的节点表示俱乐部中的成员,而边表示成员之间的朋友关系.由于俱乐部中一名管理人员和一名教练的矛盾,导致俱乐部分裂成了两个派系.图 6 描述了该网络的拓扑结构,节点的颜色对应基准信息中的两个社区.紫色虚线表示 CDATP 初始社区划分,红色虚线表示边缘修剪引发的节点转移.

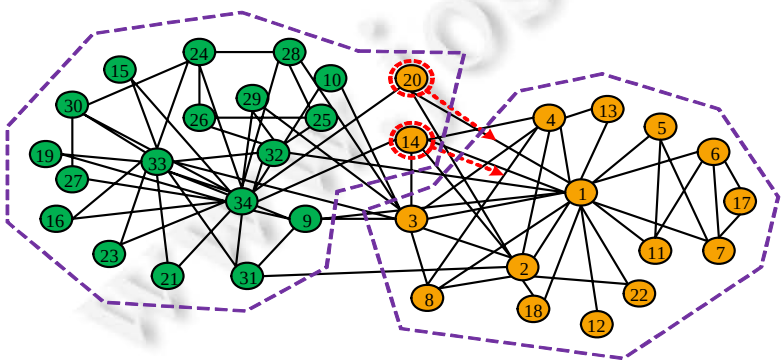


Fig.6 Karate network and community division result of CDATP

图 6 Karate 网络及 CDATP 的社区划分

图 7 是不同 *back* 取值下各节点 *Core* 的平均值,可以观察到,节点 1 和节点 34 的 *Core* 总是最大或第二大的.而在现实世界中,节点 1 和节点 34 分别对应两个派系的领导[29],因此他们处于社区的核心位置,所以对应的 *Core*

才会较大.CDATP 的节点对社区重要程度的评价方法能很好地适应真实网络条件.

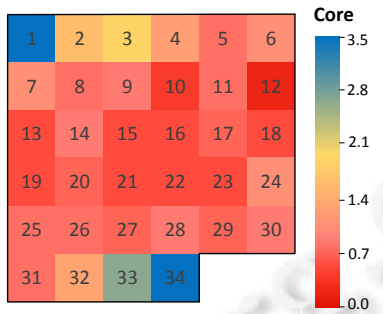


Fig.7 *Core* index of nodes in Karate network

图 7 Karate 中各节点的 *Core*

Dolphins 数据集共包含 62 个节点和 159 条无向边.网络中的每个节点代表一只宽吻海豚,被每条边连接起来的两节点对应的海豚间有频繁的互动.如图 8 所示,网络中原本有两个社区,以节点的颜色作为区分,绿色虚线是 CDATP 的初始社区划分,紫色虚线表示边缘修剪引发的节点转移.

当网络按照标记被划分为两个社区时,模块度 Q=0.396,并未达到最大值,因此,以模块度为基础的算法会继续让社区分裂,难以找到正确的社区数目.如图 8 所示:当 *back*=0.1 时,当初始聚类结束后,只有节点“DN63”和“Oscar”被划分到了错误的社区,此时的 NMI 值为 0.780 3.以“DN63”为例,它同时和“Upbang”“SN9”这两个核心系数较大的节点相连,而“SN9”的核心系数为 1.356 8,略大于“Upbang”的 1.308 7,所以“DN63”在聚类初始化时选择了错误的聚类方向,但是在边缘修剪过程中,由于左边社区对“DN63”有更大的吸引力,所以“DN63”转移到了左边的也即正确的社区中.

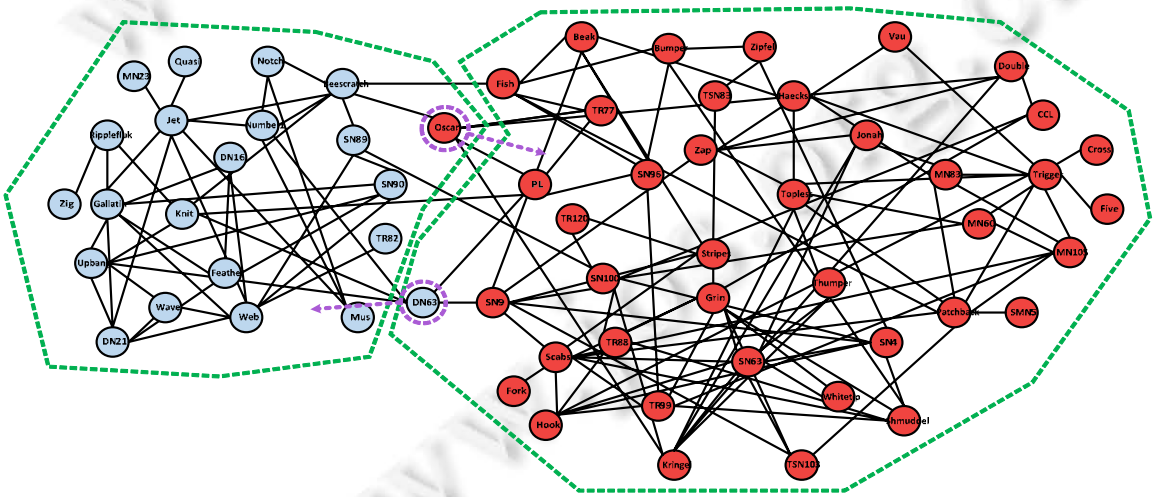


Fig.8 Dolphins network and community division result of CDATP

图 8 Dolphins 网络及 CDATP 的社区划分

PolBooks 数据集中的每个节点都代表一本美国的政治类书籍,如果有两本书被同时从 Amazon.com 买走,则对应的两个节点间会有边相连.书的类型按政治倾向分为“左派”“右派”和“中立派”这 3 类,其中,“中立派”数量最少.

CDATP 的社区划分结果以紫色虚线形式在图 9 中标出,橙色代表“右派”,绿色代表“左派”,浅蓝色代表“中立派”,红色虚线表示边缘修剪引发的节点转移.CDATP 识别出了两个最大的社区,并且只有很少量“保守派”节

点被划分错误,但 CDATP 没有识别出包含节点数量最少的“中立派”社区,而是将其分散到了另外两个大的社区当中,这是由于“中立派”书籍总是与其他类型的书籍被一起购买,而不同“中立派”书籍间联系又比较少导致的.

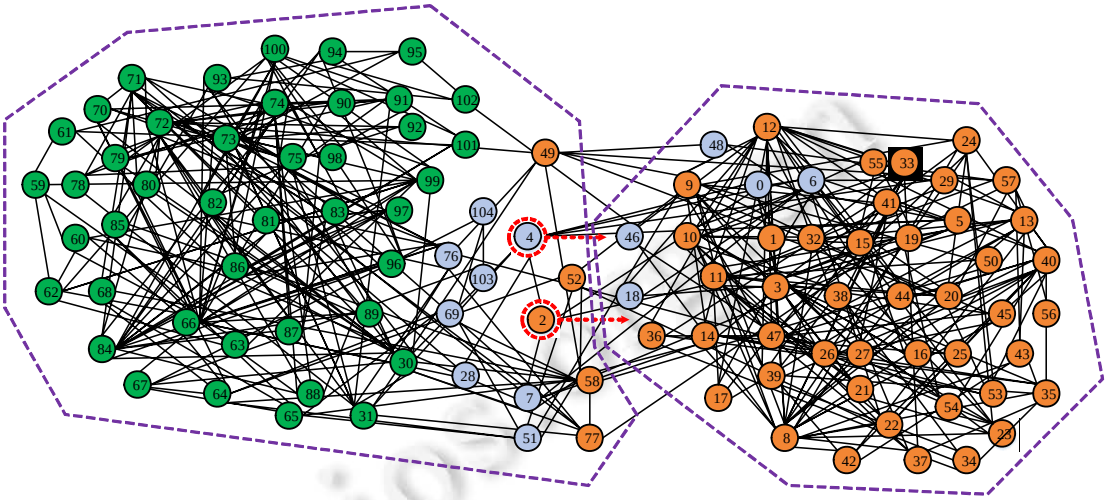


Fig.9 PolBooks network and community division result of CDATP

图 9 PolBooks 网络及 CDATP 的社区划分

PolBlogs 网络是围绕 2004 年美国总统大选时的政治类型博客建立的,其中的每个节点代表一个博客,按政治倾向分为“左派”和“右派”,节点间的边代表博客间的超链接,与对比算法一样,CDATP 将其作为无向边处理.由于节点数目太多,在图 10 中使用了超节点来表示社区,其中,浅蓝色代表“左派”,深蓝色代表“右派”,其大小和包含节点的数目成正比,虚线上标的数字为边缘修剪时的转移节点数,而孤立节点未在图中标出.从图中可以看到,使用 CDATP 算法划分社区后,只有少量的节点被错误划分.

如图 11 所示,本文在实验中测试了不同 *back* 值的条件下,CDATP 聚类初始化和边缘修剪完成后的聚类效果.由于 PolBlogs 数据集中有 266 个节点未与其他节点产生联系,所以理论上最大 NMI 为 0.666 3.可以看出,除了在 PolBooks 数据集中部分情况下,边缘修剪后导致 NMI 有小幅度下降,大多数情景中,边缘修剪都能使 NMI 有一个不错的提升.在 Dolphins 网络中,当 *back* 超过 0.1 时,NMI 出现了下降.这是因为此时检测到的社区数由 2 变成了 3,原来的两个社区中较大的一个出现了分裂,而在其他情况下,算法准确度对 *back* 的变化并不敏感,边缘修剪步骤很好地提升了初始社区的质量.

图 12 是各种算法的实验结果对比.Karate 和 Dolphins 是两个典型的基准社区结构对应较低质量指标的例子.Karate 网络和 Dolphins 网络的基准社区数量均为 2,且当它们的模块度达到最大值时,都会被划分为 4 个社区,对应的 NMI 较低,因此,基于优化社区质量指标的社区发现算法在这样的真实网络上表现不佳.并且在 Karate 数据集中还存在着“双峰结构”[35],当模块度第 1 次到达极大值时,对应的社区划分质量非常低,这也导致了一些基于社区质量指标的社区发现算法无法适应这类网络.ISCD+在 Karate 网络上的 NMI 虽然也达到了 1,但需要通过大量的准备实验和专家知识来设定社区数量,实用性较弱.而 CDATP 是参考事件传播规律设计的,对真实网络的适应性更强,因此表现较好.

在 PolBooks 数据集中,由于“中立派”书籍总是和其他类型书籍被一起购买,所以“中立派”社区内部边的数目明显小于与其他社区之间边的数目,社区结构非常松散.CDATP 在该数据集上划分正确的节点数目最多,但由于没有将“中立派”单独划分为一个社区,所以 NMI 略低于 ROCONA.PolBlogs 数据集描述的网络是具有方向性的,但基于无向网络的社区发现算法直接将其作为无向网络处理,所以各算法在该数据集上的表现均不是太好.这说明对于有向网络,当忽略掉边的方向性后,将丢失掉网络中一些重要的信息.

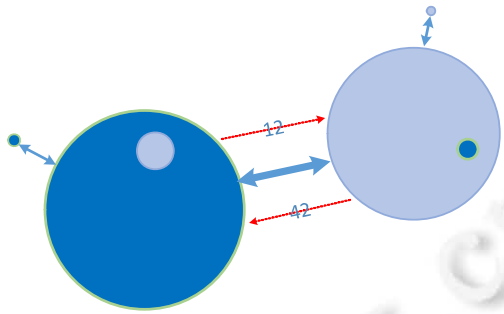


Fig.10 PolBlogs network and community division result of CDATP

图 10 PolBlogs 网络及 CDATP 的社区划分

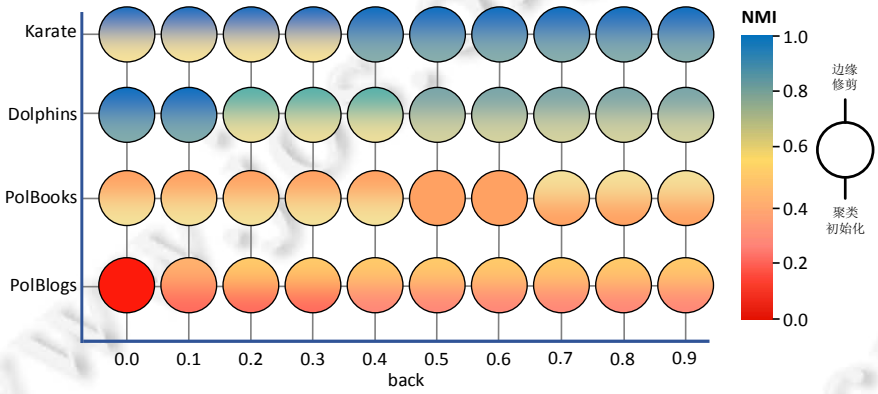


Fig.11 Relation of NMI and *back* index

图 11 NMI 与 *back* 的关系

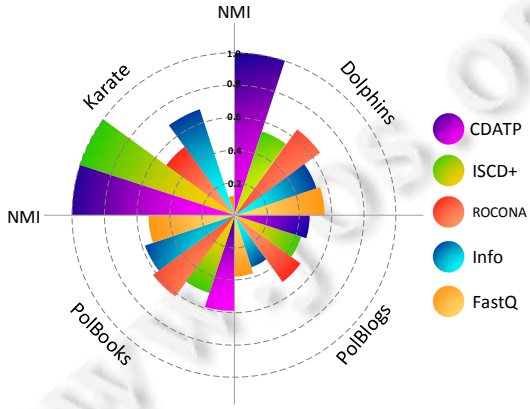


Fig.12 NMI comparison on real world undirected datasets

图 12 社区划分结果 NMI 对比

3.3 基于有向网络的CDATP有效性验证

有向版本的 PolBlogs 描述了边的方向.与文献[27]一样,首先将网络中的 266 个孤立节点除去.图 13 展示了各算法实验结果对比.算法 ConClus,OSLOM 和 LP 的 NMI 分别为 0.678 9,0.572 1 和 0.385 3,均低于 CDATP 的 NMI.

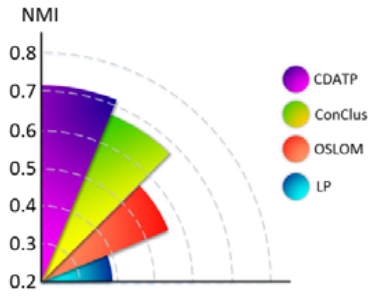


Fig.13 Comparison of NMI on PolBlogs network

图 13 PolBlogs 网络社区划分 NMI 对比

在源数据中,每个博客都是从博客检索网站得到的.这些博客检索网站包括 Blogarama,LeftyDirectory 等,一共有 6 个.如图 14 所示:若将每个博客的检索源作为属性构造属性增强图,其聚类结果的 *NMI* 比未使用属性时提高了 9.8%,说明了属性增强网络在提高社区划分准确度上的有效性,但同时也应注意到,准确度的提升并不是非常大.

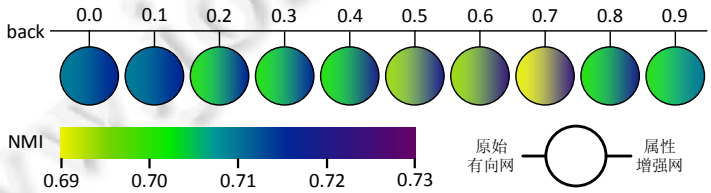


Fig.14 Influence of attribute enhancement network on results

图 14 属性增强网络对结果的影响

这种现象的发生有两个原因.

- 一是大多数博客的检索源都是 Blogarama,而 Blogarama 并没有明显的政治倾向,其中包含的两种政治倾向的博客数量基本持平,所以不会对结果造成什么影响;
- 二是因为某一些检索源虽然有非常明确的政治倾向,如 LeftyDirectory 中基本全是“左派”博客,但对应这些检索源的博客数量又相对较少,所以只能使结果的 NMI 有小幅度的上升.

图 15 介绍了在不同参数的人工网络中各算法的表现.如图所示,ConClus,OSLOM 和 LP 的准确度对数据集规模比较敏感.当数据集规模增大时,上述算法划分社区的准确度会有小幅提升.而 CDATP 的准确度几乎不受数据集规模的影响,在各种条件下都有较好的表现.由此可以看出,CDATP 在人工构造的网络中同样有很好的表现且适应性较强.

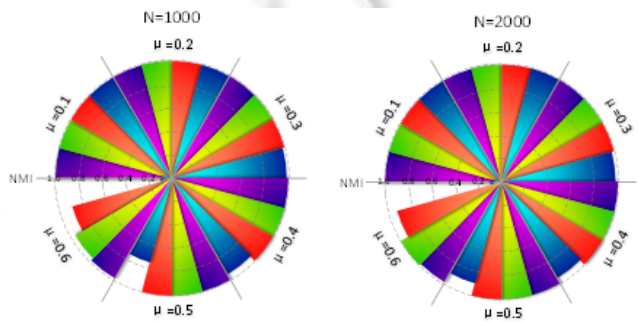


Fig.15 NMI of different algorithms on LFR benchmark network

图 15 LFR 基准网络上的算法 NMI 对比

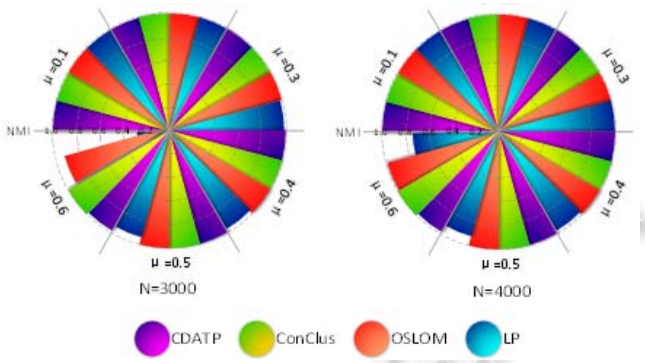


Fig.15 NMI of different algorithms on LFR benchmark network (Continued)

图 15 LFR 基准网络上的算法 NMI 对比(续)

4 结论及展望

为了能够在各种类型的社会网络中准确地划分社区,本文提出了一种新的基于节点不对称转移概率的社区发现算法 CDATP.CDATP 为每个节点设计了不对称的转移概率,并结合事件传播规律来对节点在社区中的重要程度进行评价.在聚类过程中,节点会根据转移概率等信息自发地确定转移方向,不需要预先设定社区数目.

为了检验 CDATP 的表现,本文做了大量实验并得出了以下结论.

(1) *Core* 指标正确地描述了节点在社区中的重要性,这说明基于网络拓扑结构设计的节点不对称转移概率充分体现了网络中节点的不对等关系;

(2) 在真实数据集上,CDATP 有着非常好的表现,无需通过额外的实验和专家知识指定转移迭代次数.这说明基于事件传播规律的聚类方法能够很好地适应较为复杂的真实社会网络.

进一步的研究需要在以下 3 个方面展开.

(1) 对于有权重网络,如何利用边的权重来构建节点转移概率;

(2) 节点的聚类方向可以不限于 1 个,特别是对社区边缘的节点.可以对确定聚类方向的流程加以改进,以发现重叠社区;

(3) 研究如何根据不同网络的特征来对本文中公式的参数进行相应的调整,以提高社区发现的准确率.

致谢 在此,我们向对本文的工作给予支持和建议的审稿人、主编、编辑、同行、同学和老师表示感谢.

References:

[1] Newman MEJ, Girvan M. Finding and evaluating community structure in networks, Physical Review E, 2004,69:026113.

[2] Li ZP, Wang RS, Zhang SH, Zhang XS. Quantitative function and algorithm for community detection in bipartite networks. Information Sciences, 2016,367(C):874–889.

[3] You T, Cheng HM, Ning YZ, Shia BC, Zhang ZY. Community detection in complex networks using density-based clustering algorithm and manifold learning. Physica A: Statistical Mechanics and its Applications, 2016,464:221–230.

[4] Amiri B, Hossain L, Crawford JW, Wigand RT. Community detection in complex networks: Multi–objective enhanced firefly algorithm. Knowledge-Based Systems, 2013,46(1):1–11.

[5] Huang FL, Zhang SC, Zhu XF. Discovering network community based on multi-objective optimization. Ruan Jian Xue Bao/Journal of Software, 2013,24(9):2062–2077 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4400.htm [doi: 10.3724/SP.J.1001.2013.04400]

[6] Fortunato S, Barthélemy M. Resolution limit in community detection. Proc. of the National Academy of Sciences, 2007,104:36–41.

[7] Bagrow JP. Communities and bottlenecks: Trees and treelike networks have high modularity. Physical Review E, 2012,85:066118.

[8] Pons P, Latapy M. Computing communities in large networks using random walks. In: Proc. of the Int'l Symp. on Computer and Information Sciences. Springer-Verlag, 2005. 284−293.

[9] Lai D, Lu H, Nardini C. Finding communities in directed networks by PageRank random walk induced network embedding. Physica A: Statistical Mechanics and Its Applications, 2010,389(12):2443−2454.

[10] Jiao QJ, Huang Y, Shen HB. Community mining with new node similarity by incorporating both global and local topological knowledge in a constrained random walk. Physica A: Statistical Mechanics and Its Applications, 2015,424:363−371.

[11] Huang X, Cheng H, Yu JX. Dense community detection in multi-valued attributed networks. Information Sciences, 2015,314(C): 77−99.

[12] Su C, Jia X, Xie X, Yu Y. A new random-walk based label propagation community detection algorithm. In: Proc. of the IEEE/WIC/ ACM Int'l Conf. on Web Intelligence and Intelligent Agent Technology. IEEE, 2015. 137−140.

[13] Jin D, Yang B, Liu J, Liu DY, He DX. Ant colony optimization based on random walk for community detection in complex networks. Ruan Jian Xue Bao/Journal of Software, 2012,23(3):451−464 (in Chinese with English abstract). http://www.jos.org.cn/ 1000-9825/3996.htm [doi: 10.3724/SP.J.1001.2012.03996]

[14] Wang W, Liu D, Liu X, Pan L. Fuzzy overlapping community detection based on local random walk and multidimensional scaling. Physica A: Statistical Mechanics and Its Applications, 2013,392(24):6578−6586.

[15] Xin Y, Xie ZQ, Yang J. The adaptive dynamic community detection algorithm based on the non-homogeneous random walking. Physica A: Statistical Mechanics and Its Applications, 2016,450:241−252.

[16] Meo PD, Ferrara E, Fiumara G, Provetti A. Enhancing community detection using a network weighting strategy. Information Sciences: An Int'l Journal, 2013,222(3):648−668.

[17] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70:66111.

[18] Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. In: Proc. of the IEEE Int'l Conf. on Data Mining. IEEE Computer Society, 2012. 745−754.

[19] Bai L, Cheng X, Liang J, Guo Y. Fast graph clustering with a new description model for community detection. Information Sciences, 2017,s 388-389(C):37−47.

[20] Fortunato S. Community detection in graphs. Physics Reports, 2010,486:75−174.

[21] Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 2007,76(3 Pt 2):036106.

[22] Malliaros FD, Vazirgiannis M. Clustering and community detection in directed networks: A survey. Physics Reports, 2013,533(4): 95−142.

[23] Leicht EA, Newman MEJ. Community structure in directed networks. Physical Review Letters, 2007,100(11):118703.

[24] Rosvall M, Bergstrom CT. An information-theoretic framework for resolving community structure in complex networks. Proc. of the National Academy of Sciences, 2007,104:7327.

[25] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S. Finding statistically significant communities in networks. PloS One, 2011, 6(4):e18961.

[26] Lancichinetti A, Fortunato S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E, 2009,80:016118.

[27] Santos CP, Carvalho DM, Nascimento MCV. A consensus graph clustering algorithm for directed networks. Expert Systems with Applications, 2016,54:121−135.

[28] Zhang J, Liu B, Tang J, Chen T, Li J. Social influence locality for modeling retweeting behaviors. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2013. 2761−2767.

[29] Zachary WW. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1997,33: 452−473.

[30] Lusseau D, Schneider K, Boisseau OJ, Haase P, Slooten E, Dawson SM. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. Behavioral Ecology and Sociobiology, 2003,54(4):396−405.

[31] Krebs B. Books about US Politics. 2004. http://www.orgnet.com/

[32] Adamic LA, Glance N. The political blogosphere and the 2004 US election. In: Proc. of the WWW 2005 Workshop on the Weblogging Ecosystem. 2005.

[33] Kumar P, Gupta S, Bhasker B. An upper approximation based community detection algorithm for complex networks. Decision Support Systems, 2017. 103−118.

[34] Tesmer M, Perez CA, Zurada JM. Normalized mutual information feature selection. IEEE Trans. on Neural Networks, 2009,20(2): 189.

[35] Yang Y, Sun PG, Hu X, Li ZJ. Closed walks for community detection. Physica A: Statistical Mechanics and Its Applications, 2014, 397(3):129−143.

附中文参考文献:

[5] 黄发良,张师超,朱晓峰.基于多目标优化的网络社区发现方法.软件学报,2013,24(9):2062−2077. http://www.jos.org.cn/1000-9825/4400.htm [doi: 10.3724/SP.J.1001.2013.04400]

[13] 金弟,杨博,刘杰,等.复杂网络簇结构探测——基于随机游走的蚁群算法.软件学报,2012,23(3):451−464. http://www.jos.org.cn/1000-9825/3996.htm [doi: 10.3724/SP.J.1001.2012.03996]

许平华(1995－),男,湖北荆州人,学士,主要研究领域为复杂网络.

胡文斌(1976－),男,博士,教授,博士生导师,主要研究领域为人工智能,智能仿真优化,大数据与数据挖掘.

邱振宇(1992－),男,博士生,主要研究领域为社会网络.

聂聪(1993－),男,硕士,主要研究领域为智能仿真与优化.

唐传慧(1995－),男,学士,CCF 学生会员,主要研究领域为智能交通.

高旷(1994－),男,学士,主要研究领域为复杂网络,车载自组织网络.

刘中舟(1993－),男,学士,主要研究领域为生物信息学,复杂网络.