中国科学: 信息科学 2016 年 第 46 卷 第 4 期: 431–444
SCIENTIA SINICA Informationis

《中国科学》杂志社
SCIENCE CHINA PRESS

论 文

NCSS: 一种快速有效的复杂网络社团划分算法

韩忠明*, 谭旭升, 陈炎, 段大高

北京工商大学计算机与信息工程学院, 北京 100048
* 通信作者. E-mail: hanzm@th.btbu.edu.cn

收稿日期: 2015–11–01; 接受日期: 2015–12–25; 网络出版日期: 2016–04–13
国家自然科学基金项目 (批准号: 61170112)、中央财政支持地方高校发展专项资金人才培养和创新团队建设项目 (批准号: 19005323132) 和教育部人文社会科学研究基金项目 (批准号: 13YJC860006) 资助

摘要 划分复杂网络中的社团对研究复杂系统在理论和实践上都非常重要. 全局社团划分算法划分结果相对准确, 但计算量大; 局部社团划分算法多数精度低、稳定性差. 本文借鉴全局社团划分算法与局部社团划分算法的特点, 提出了基于节点中心度和节点与社团间结构关系强度的社团划分算法 (NCSS). NCSS 算法基于全局信息找到核心节点, 然后构建初始社团, 从而有效摆脱了传统局部社团算法中初始节点随机选择的稳定性差的弊端. 在社团扩张过程中, NCSS 每次只需计算社团边缘节点和与其有联系的社团之间的结构关系强度, 从而降低了算法时间复杂度. 本文在真实网络和模拟基准网络上进行了大量实验, 与主流的全局社团划分算法和局部社团划分算法作了比较. 在真实网络上 NCSS 算法的精度比其他算法平均提高 6%; 在模拟基准网络上, 精度则可以提高 18%.

关键词 复杂网络 社团结构 社团划分 核心节点 结构关系强度

1 引言

社团结构划分是复杂网络研究的重要内容, Newman 把社团结构定义为复杂网络的若干群体, 每个群体内部节点间的连接相对紧密, 而群体之间的连接相对稀疏 [1,2]. 很多研究表明, 复杂网络在社团结构上反映的结构特性与网络整体所反映的结构特性差异很大, 如果忽略社团结构, 很多有意义的结构特性难以有效表征 [3,4]. 因此, 复杂网络社团划分成为众多研究者关注的热点领域.

复杂网络社团划分对深入理解真实网络的结构具有非常重要的价值. 社团是由复杂网络中具有相同性质的个体组成, 如万维网中的社团可以是讨论相关主题的若干网站 [5], 生物化学网络或者电子电路网络中的社团可以是某一类功能相同的单元 [6]. 发现复杂网络中的社团有助于了解复杂系统的拓扑性质和组织结构. 复杂网络社团划分算法, 可分为基于全局信息的划分算法、基于局部信息的划分算法以及其他划分算法.

Kernighan 和 Lin [7] 提出了一种基于贪婪思想的试探优化算法, 即 Kernighan-Lin 算法. 谱方法 [8,9] 利用网络邻接矩阵的特征值和特征向量来进行社团划分. GN (Girvan-Newman) 算法 [10] 极大地推动了社团划分算法的发展, 提出了重要的边介数概念, 该算法通过逐步删除边介数较高的边, 完

引用格式: 韩忠明, 谭旭升, 陈炎, 等. NCSS: 一种快速有效的复杂网络社团划分算法. 中国科学: 信息科学, 2016, 46: 431–444, doi: 10.1360/N112015-00026

© 2016《中国科学》杂志社

成社团的划分过程. GN 算法在划分的精确程度上有了较明显地提高, 但时间复杂度非常高. 鉴于此, Newman [11] 提出快速算法, 该算法旨在快速有效地发现大型网络中的社团, 但需要不停地向已有社团中添加边. Blondel 等 [12] 提出了一种基于模块度优化的启发式 Fast Unfoliding 算法, 主要包括两个阶段: 首先将每个节点指定到一个社区, 然后按顺序将节点在这些社区间进行移动; 最后将第一个阶段得到的社区视为新的 "节点", 重新构造子图. Raghavan 等 [13] 提出了标签传播算法 (LPA), LPA 为所有节点指定一个唯一标签, 然后逐轮刷新所有节点的标签, 直到收敛为止. Ying 等 [14] 通过迭代方式把每个节点划分到和它拥有最大相似度的社团来发现社团. Zanjani 等 [15] 从网络中的一些特殊节点入手, 通过判断节点标准确定节点间的关系紧密与否, 依赖程度强弱, 然后将关系紧密的节点归于一个社团, 进而划分社团. 王兴元等 [16] 提出了一种基于节点间依赖度的划分社团算法. Hour 和 Kan [17] 提出了一个基于网络分析块的算法, 该算法基于结构对等性和规模对等性刻画一般网络结构. Riolo 和 Newman [18] 通过优化最小分割矩阵, 将网络对应 Laplace 矩阵的主特征向量映射到谱中. Darst 等 [19] 把非回溯算法和流量矩阵引入谱聚类中来探测社交网络的社团结构. Tsourakakis 等 [20] 提出一个框架, 这个框架合并了两个正交的启发式搜索算法, 并且能够通过权值来量化它们, 框架能够适应传统的社团发现算法, 提高它们的性能. Mokross 等 [21] 认为通过节点的连通性能够有效地划分社团结构. Albert 等 [22] 用双曲率来刻画有限的网络结构, 它能够较强地反映网络中高阶连通性节点和其余节点的拓扑关系, 将高阶节点作为社团划分的关键节点. 付立东等 [23] 提出了一种新的基于模块密度中心性算法, 该算法以模块密度为基础来度量一个节点对社团的贡献度, 通过节点对不同社团贡献度的多少来进行社团划分. 除此之外, 还有基于随机游走的 walktrap 算法 [24]、基于物理学中自旋玻璃模型的 spinglass 算法 [25]. 全局算法从整个网络的角度出发, 有助于提高划分结果的准确率, 但常需要计算整个网络的信息, 因此时间复杂度比较高.

Clauset [26] 提出了一个基于 R 值的局部社团挖掘算法, 该算法首先定义了一个局部模块度 R, 采用贪婪算法的思想, 迭代地添加使 R 值增加最大的邻居节点. Lancichinetti 等 [27] 提出基于适应度的重叠社团发现算法 LFM, 依据社团的局部结构定义一个适应度, 通过最优化适应度发现社团. 针对 LFM 算法以随机选择初始种子节点, 可能导致算法存在无法终止等问题, Lee 等 [28] 通过修改初始节点, 选择极大连通子图作为初始社团, 取代随机选择, 对算法进行改进, 提出 GCE 算法. 刘旭等 [29] 提出一种新的局部相似性度量, 并结合层次聚类算法用于社区发现. Chen 等 [30] 提出了局部度中心节点, 通过源节点的最近和次近局部度中心节点来探测局部社团. 袁超等 [31] 首先推导出一个适合局部社团挖掘的节点相似度模型, 然后基于该模型使用内外夹逼的思想, 实现了一个有效的局部社团挖掘 SICE 算法. Perozzi 等 [32] 提出了一个以用户需求为导向的社团划分算法 FocusCO, 从用户提供的关键节点开始进行局部社团划分. Moradi 等 [33] 提出一个不需要初始参数的种子节点扩张算法, 仅需要网络的局部结构来获得种子节点. Zalik 等 [34] 提出了一个从源顶点出发的多分辨率社区在线检测局部社区发现算法. 局部社团划分算法大体存在以下几个缺陷: (1) 算法稳定性不好, 对起始节点的位置敏感. 如果初始节点处于社团核心, 那么效果较好. 反之, 则效果较差. (2) 某些算法的收敛条件缺乏自适应性. 局部算法从网络的种子节点出发, 每次更新社团内节点只需用到社团内部节点和边缘节点的邻节点的信息, 算法的时间复杂度较低, 但因没考虑全局信息, 容易受种子节点影响, 算法精度不高.

将其他学科中的一些概念引入, 也可以用于社团划分, 淦文燕等 [35] 提出了一种基于拓扑势的社区发现算法, 该算法将每个社区视为拓扑势场的局部高势区, 通过寻找被低势区域所分割的连通高势区域实现网络的社区划分. Wu 和 Huberman [36] 提出了快速启发式算法 WH, 该算法将复杂网络建模为电路系统, 网络连接看成是具有电阻的线路, 不同位置的网络节点具有不同的电位势. 何东晓等 [37] 提出了一种基于聚类融合的遗传算法用于社区挖掘, 该算法将聚类融合引入到交叉算子中, 利用父个

体的聚类信息, 辅以网络拓扑结构的局部信息产生新个体, 避免了传统交叉算子单纯交换字符块而忽略了聚类内容所带来的问题. Okamoto [38] 借鉴神经网络中细胞集合假设, 提出了把复杂网络中的社团类比为神经网络中的细胞集合的社团划分算法. 张忠元 [39] 将字典学习算法应用到网络社团结构探测问题, 提出了一种简化的字典学习算法, 该算法结合了非负矩阵分解模型下的字典学习算法和最小二乘方回归, 这种字典学习算法更简单, 收敛速度更快, 并能取得较好的社团划分结果.

本文借鉴全局社团划分算法与局部社团划分算法的特点, 提出了基于节点中心度和节点与社团间结构关系强度的社团划分算法 (node centrality-structural strength based community detection algorithm, NCSS). NCSS 首先选择核心节点, 然后基于节点与社团间结构关系强度进行社团扩张. NCSS 算法核心节点的选择借助了网络全局信息, 有效摆脱了传统局部社团发现中初始节点随机选择的稳定性差的弊端; 在社团扩张过程中, NCSS 每次只需计算每个社团的边缘节点和与其有联系的社团之间的结构关系强度, 从而避免每次计算都考虑网络全局信息.

2 NCSS 算法

社团划分算法中, 初始种子节点的选择和社团结构的扩张策略对社团划分的精度至关重要. NCSS 算法首先采用网络核心节点作为社区的种子节点, 通过节点与社团间的关系强度进行扩张. 算法可以分为两个阶段, 第一阶段种子节点选择; 第二阶段进行社团扩张.

2.1 种子节点选择

在真实社会中, 每个社团都有对应的关键人物, 如某个领域的顶尖学者会有很多相关研究者关注其学术动态, 微博上的意见领袖会有很多关注者. 网络中节点的重要程度不同: 核心节点具有较高重要性, 而普通节点则是网络中的参与者. 社团是由核心节点和追随节点组成的一个团体, 划分一个社团, 首先需要找出社团的核心节点. 由复杂网络的无标度性 [40] 可知, 网络中大部分节点只和很少的节点连接 (节点的 “度” 很小), 而极少的节点与非常多的节点连接 (节点的 “度” 很大). 假设网络中每个节点都属于一个社区 (独立的节点单独成为一个社区), 则可以借鉴 PageRank 模型 [41] 定义节点中心度. 节点中心度对无向无权图中节点进行排序, 以找出网络中潜在的核心节点.

定义 1 (节点间共同邻居) 节点 i 与 j 的共同邻居定义为

$$C_{i,j} = Neighbor(i) \cap Neighbor(j), \tag{1}$$

其中, $Neighbor(i)$ 表示节点 i 的邻居节点集.

定义 2 (节点中心度) $G = \langle V, E\rangle$ 是一个无向无权网络, 节点 i 的中心度 ($PRcen$) 为

$$PRcen(i) = \mathrm{c}\sum_{j\neq i} PRcen(j)\frac{|C_{i,j}|}{\sum_{k\in Neighbor(i)}|C_{i,k}|+1} + \frac{1-\mathrm{c}}{|V|}, \tag{2}$$

其中, c 表示阻尼系数, 是一个常数, $\mathrm{c}\in(0,1)$; $|C_{i,j}|$ 表示节点 i 和 j 的共同邻居; $|V|$ 表示 G 的节点数. 由于存在度为 0 的孤立节点, 因此需要加上阻尼系数 c 以便加快收敛, 一般地 c 为 0.8. 用递归算法描述节点中心度的计算过程, 节点中心度值取决于所有邻接节点的中心度值, 而后者依据的是它和该邻接节点的共同邻居占它和它所有邻接节点的共同邻居总和的比例. 若某个节点具有较高中心度值, 其邻居节点的中心度较高. 它们之间共同邻居比例占该节点与所有邻节点共同邻居的比例表示这两个节点之间的强度, 强度越高, 中心度值越高. 计算每个节点的中心度的算法如算法 1 所示. 算法 1

采用阈值 ε 作为判别节点中心度的精度控制误差, ε 一般根据实际情况设定. 算法初始为每个节点的中心度赋值 $\frac{1}{n}$, 然后第 2 行到第 12 行进行循环迭代. 第 4 行到第 8 行计算一个节点的邻居节点的中心度和, 然后第 9 行到第 11 行根据 (2) 式计算一个节点的中心度.

将初始节点按照中心度从大到小排序, 选择中心度大的节点作为备选核心节点, 但单纯依赖排序结果选择初始核心节点会带来潜在风险. 如排序靠前的多个节点可能属于同一个社团, 按中心度从大到小选择, 可能会把一个大的社区分解为若干个小社区; 此外, 与最大中心度相连接的节点中心度较大, 若其中某个节点位于整个社团的边缘且与外界节点联系较少, 以它作为中心节点进行扩张会影响社团划分结果. 为了避免这些风险, 我们在核心节点选择上附加如下约束:

(1) 为使核心节点更能代表每个社团结构, 选择的核心节点中心度值必须大于网络中所有节点的中心度的平均值;

(2) 为避免核心节点与外界节点联系较少, 选择的核心节点的度必须大于网络中所有节点的度的平均值;

(3) 为避免先后选出的几个核心节点在同一个社区中, 当前选择的核心节点必须和已有核心节点都非直接相连.

算法 1 Node centrality algorithm

Require: $\varepsilon, \mathrm{c}, G =< V, E >$;
Ensure: $PRcen$;
1: $\forall i, PRcen[i] \Leftarrow \frac{1}{num(V)}$;
2: **while** $res > \varepsilon$ **do**
3: $i, PRcen_t[i] \Leftarrow 0$;
4: **for** all i **do**
5: **for** all j **do**
6: $PRcen_t[j] \leftarrow PRcen_t[j] + PRcent[i] * \frac{|C_{j,i}|}{\sum_{k \in Neightbor(i)} |C_{i,k}|+1}$;
7: **end for**
8: **end for**
9: $\forall i, PRcen_t[i] = \mathrm{c} * PRcen_t[i] + \frac{1-\mathrm{c}}{num(V)}$;
10: $res \leftarrow |PRcent_t - PRcent|$;
11: $PRcen \leftarrow PRcen_t$;
12: **end while**

2.2 节点与社团之间的结构关系强度和社团扩张策略

本小节提出节点与社团之间的依赖度和节点与社团之间的结构关系强度, 然后基于节点对社团依赖度和结构关系强度进行社团扩展.

定义 3 (节点对社团的依赖度) 节点 a 对社团 C 的依赖度 $D_{a,C}$ 定义为

$$D_{a,C} = \frac{n_{a,C}}{k_a}, \tag{3}$$

其中, $n_{a,C}$ 表示社团 C 中节点 a 的邻居数目, k_a 表示节点 a 的邻居数目.

定义 4 (节点对社团的结构关系强度) 节点 i 对社团 C 的结构关系强度 $S_{i,C}$ 定义为

$$S_{i,C} = \sum_{j \in C} 2 \cdot |C_{i,j}| - \frac{k_{i,j}^{out} \cdot s_C^{in}}{m} - \frac{k_{i,j}^{in} \cdot s_C^{out}}{m}, \tag{4}$$

其中, j 表示社团 C 内的节点; m 表示网络中所有节点度的总和; $k_i^{in/out}$ 表示节点 i 与 j 的共同 (非共同) 邻居节点数; $s_C^{in/out}$ 表示社团内部 (外部) 边的数目, $s_C^{in/out} = \sum_{j \in C} k_j^{in/out}$. 基于节点对社团依赖度和结构关系强度, 构建社团扩张算法, 如算法 2 所示. 首先依据 2.1 小节的选择策略, 选出核心节点, 并将其初始化为每个社团的初始种子节点, 把每个社团的节点的邻接节点作为社团扩展的候选邻居节点集. 然后计算候选邻居节点集中节点与对应社团的结构关系强度, 选择拥有最大结构关系强度且最大关系强度大于 0 的节点作为备选节点集. 最后, 依次检验备选集中节点是否已经划分到其他社团, 若该节点已划分到其他社团, 则分别计算该节点与对应社团和该节点与原社团的依赖强度, 若该节点对新社团的依赖强度大于对原社团的依赖强度且对新社团的依赖强度大于 0.5, 则把该节点加入到当前对应社团并将其从原社团中移除; 若该节点还未加入任何社团, 则将节点直接加入到对应社团中. 迭代进行上述过程, 算法采用布尔值 $flag$ 为 True 代表本次扩张已经没有节点可以被划分, 即每一个节点都找到了依赖度大于阈值且对应结构关系强度为正的社团; $flag$ 为 False 代表这次扩张还有节点被划分到相应的社区, 则继续迭代. 社团扩张算法基于节点对社团的依赖度和节点对社团的结构

算法 2 Community expansion

Require: Initial Communities C;
Ensure: Convergence Flag(boolean): $flag$;

```
Neighbors N_neigh
Alternative Set N_alter
flag ⇐ True
for each c_1 in C do
    add all c_1's members's neighbors to N_neigh
    for each i in N_neigh do
        calculate S_{i,c_1}
    end for
    find all N_alter with the maximum |S_{i,c_1}| in c if |S_{i,c_1}| > 0
    for each j in N_alter do
        flag ⇐ False
        add j to c_1
        for c_2 in C if c_2 ≠ c_1 do
            for k in c_2 do
                calculate D_{k,c_1} and D_{k,c_2}
                if D_{k,c_2} ⩽ 0.5 and D_{k,c_1} > 0.5 then
                    remove k from c_2
                end if
            end for
        end for
    end for
end for
return flag
```

关系强度进行扩张, 算法考虑了社团划分结果会受节点的吸收顺序的影响. 在已划分社团中每加入一个新的节点后, 算法计算已划分的节点对社团的依赖度, 当某个节点对其当前所属社团的依赖度低于阈值, 则把它从该社团去除, 该节点返回到候选集中继续参加划分.

2.3 NCSS 算法流程

组合核心节点的选择策略和社区扩张算法, 则可以得到 NCSS 算法, 具体步骤描述如下:

Step 1 设置计算节点中心度算法的收敛阈值和阻尼系数;

Step 2 调用计算节点中心度算法, 获得核心节点列表;

Step 3 依据核心节点列表, 初始化社团结构;

Step 4 调用社团扩张算法, 对已经初始化的社团进行扩展, 要求节点对社团的结构关系强度均大于 0, 更新社团结构, 直到算法收敛;

Step 5 检查划分好的社团, 若社团中节点数低于阈值 (如社团中节点数目少于网络中节点总数目的 5%), 表明社团规模过小, 则将该社团中的所有节点转移到节点候选集中, 并删除该社团结构;

Step 6 重复 Step 4 和 5, 直到复杂网络中每一个节点都划分到一个社团中, 且每个社团的规模符合要求.

2.4 复杂度分析

NCSS 算法划分为两个部分: 第一部分 Step 1∼3 主要负责设置参数, 找出核心节点, 并初始化社团结构; 第二部分 Step 4∼6, 主要通过基于节点对社团依赖度和结构关系强度的节点扩展策略来实现社团结构的发现和更新, 最终获得最佳社团划分, 输出对应的社团, 算法结束.

设复杂网络的节点数为 n; 核心节点数为 k, 由于 2.1 小节对核心节点选择进行了约束, $k \ll n$; 节点平均度为 d, 在真实网络中 d 是一个很小的常量. 首先估计算法第一部分的复杂度:

Step 1 算法相关参数的设置, 为常量复杂度 $O(1)$;

Step 2 计算节点中心度算法时间复杂度最坏情况是 $O(n^2)$, 在网络较为稀疏的情况下, 复杂度会较低. 实际上, PageRank 通常会在 $O(\log n)$ 内收敛 [40];

Step 3 依据核心节点, 初始化社团结构的复杂度为 $O(k)$.

算法第二部分的复杂度:

Step 4 调用社团扩张算法, 对已经初始化的社团进行扩展. 计算节点对社团结构关系强度的复杂度为 $O(kd)$; 选择该节点最大的社团结构关系强度, 实际上是对该节点的社团结构关系强度排序, 所以时间复杂度为 $O(k\log k)$, 对一个社团的邻接节点而言, 它们对该社团的最大结构依赖度通常为常数 c, 在实验过程中常数 $\mathrm{c} \in [0, 30]$. 算法 2 步骤 13 和 14 两个 for 循环遍历所有节点一次, 复杂度为 $O(n)$. 因此社团扩张算法的复杂度为 $O(\max((k^2d^2), (k^2\log k), (\mathrm{c}kn)))$;

Step 5 检查社团规模, 复杂度为 $O(n)$;

Step 6 迭代过程复杂度为 $\mathrm{c1}\cdot O(\max((k^2d^2), (k^2\log k), (\mathrm{c}kn)))$, 其中 c1 为常数.

综合算法各部分分析可知, NCSS 算法的复杂度为

$$O(1+\log n+k+\mathrm{c1}\cdot O(\max((k^2d^2), (k^2\log k), (\mathrm{c}kn)))) = O(kn).$$

为了综合比较 NCSS 算法和目前全局以及局部算法复杂度的差异, 表 1 列出了 NCSS 算法、全局以及局部代表性算法的复杂度. 表 1 中的 n 为复杂网络节点数; m 为复杂网络中边的数量; k 为社团中的关键节点数, 一般的 $k \ll n$; d 为一个节点的平均度; c 为常数.

通过表 1 可以看到, 相对于其他经典算法, NCSS 算法的复杂度较低. 对于结构简单的网络, 社团数量较少, NCSS 算法呈现线性复杂度; 对于结构复杂的网络也具有明显优势.

表 1 NCSS 算法与其他经典算法关于时间复杂度的比较

Table 1 Time complexity of NCSS and other classical algorithms

Algorithm	Time complexity
GN	$O(n^3)$
Newman fast greedy	$O(mn\log n)$
Infomap	$O(n(m+n))$
Leading eigenvector	$O(n^2)$
Label propagation	$O(m+n)$
Spinglass	$O(n^{\frac{3}{2}})$
Walktrap	$O(mn^2)$
louvain	$O(n^2)$
GCE	$O(mn)$
LFM	$O(n^2)$
NCSS	$O(kn)$

3 实验结果与分析

为了客观衡量 NCSS 算法的效果和效率, 我们选择经典的社团发现算法进行对比实验. GN 算法、纽曼快速算法 (FG)、leading eigenvector 算法 (LE)、infomap 算法 (IM)、label propagation 算法 (LPA)、spinglass 算法 (SG)、walktrap 算法 (WT) 和 louvain 算法 (LV) 属于全局社团发现算法, 而 GCE 算法和 LFM 算法属于局部社团发现算法.

我们采用社团划分领域常用的测试数据集进行实验, 数据集包括真实网络数据集和模拟网络数据集. 真实网络数据集包括 Zachary 空手道俱乐部成员关系网络、宽吻海豚网络、美国政治书籍网络、美国大学生橄榄球网络、西班牙大学 email 通信网络和网络科学领域科学家合作网络. 模拟网络数据集主要由 LFR 基准图组成, 选择不同的参数生成不同的 LFR 基准图.

3.1 算法结果评价指标

采用 5 种常用的算法评价指标, 包括准确率 (precision)、召回率 (recall)、综合评价指标 ($F1$)、规范互信息量 (normalized mutual information, NMI) 和模块度 (modularity).

(1) 准确率 (precision). 准确率反映了划分结果中正确节点数量占划分结果节点数量的比例, 计算方法为

$$P=\frac{|L_{result}|\cap|L_{true}|}{|L_{result}|}, \tag{5}$$

其中, $|L_{result}|$ 表示算法得到的社团, $|L_{true}|$ 表示真实社团, 准确率表示社团划分的精度.

(2) 召回率 (recall). 召回率反映了真实社团中被正确划分出的节点的比例, 计算方法为

$$R=\frac{|L_{result}|\cap|L_{true}|}{|L_{true}|}, \tag{6}$$

其中, $|L_{result}|$ 表示算法得到的社团, $|L_{true}|$ 表示真实社团.

(3) 综合评价指标 ($F1$). 准确率和召回率只刻画了算法的一个方面. 准确率高说明划分结果中正确的节点数量占划分结果规模的比例较大, 但是却不能判断是否过分割; 召回率高说明真实社团中有

更多的节点被划分出来, 但是却不能判断是否欠分割, 如极端情况下将整个网络中的所有节点都划分为一个社团, 其 R 值为 1. 评价指标 $F1$ 综合了 P 值和 R 值, 具体计算方法为

$$F1 = \frac{2PR}{P+R}, \tag{7}$$

其中, P 为准确率, R 为召回率.

(4) 规范互信息量 (NMI). NMI (I) 用来描述划分结果与真实结果之间的相关性, 取值范围为 [0, 1], 值越接近 1 表示划分结果越好. 定义如下:

$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} X_{ij}\log\left(\frac{X_{ij}N}{X_{i.}X_{.j}}\right)}{\sum_{i=1}^{C_A} X_{i.}\log\frac{X_{i.}}{N} + \sum_{j=1}^{C_B} X_{.j}\log\frac{X_{.j}}{N}}, \tag{8}$$

其中, A 表示真实网络社团划分; B 表示实验得到的社团划分; C_A 表示社团的实际个数; C_B 算法划分出的社团个数; X 为混合矩阵, 其元素 $X_{ij} = |i \cap j|(i \in A, j \in B, |i \cap j|$ 表示 $i \cap j$ 的 1 范数); $X_{i.}$ 表示混合矩阵 X 第 i 行的和; $X_{.j}$ 表示混合矩阵 X 第 j 列的和.

(5) 模块度 (modularity). 模块度用来评价无标记网络中的社团划分结果, 模块度又称为 Q 函数. Q 函数为社团内实际连接数目与随机连接情况下社团内期望连接数目之差, 用来对社团划分的整体质量做出一个定量的评价. 定义如下:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{i,j} - \frac{k_i k_j}{2m}\right)\sigma(C_i, C_j), \tag{9}$$

其中, k_i 和 k_j 表示节点的度; C_i 表示节点 i 所属的社团; m 表示网络的总边数. 当 $C_i = C_j$ 时, $\sigma(C_i, C_j) = 1$, 否则为 0.

3.2 真实网络数据集

实验采用的真实网络统计信息如表 2 所示. 带标记网络有: Zachary 空手道俱乐部成员关系网络是一个社会网络, 俱乐部分裂成以主管和校长为核心的两派; 宽吻海豚网络记录了 62 头宽吻海豚的社会关系, 网络包含两个小的社团; 美国大学橄榄球联赛网络由 Girvan 和 Newman 收集整理, 数据来自 2000 年美国 115 所大学橄榄球队之间的比赛, 该网络分为 12 个社团, 分别对应联赛的 12 个联盟; 美国政治书籍网络来自亚马逊书店的政治书籍销售记录, 包括 105 个节点和 441 条边, 按照书籍的政治倾向, 该网络被分为三个社团, 分别代表自由、中立和保守. 无标记网络包含西班牙大学 email 通信网络和网络科学领域科学家合作网络, 合作网络描述了来自网络科学领域的 1589 位科学家的合作关系, 共包括 1589 个节点和 2742 条边.

3.2.1 带标记网络

首先 NCSS 和比较算法应用于有标记网络, 用每种算法对不同数据集进行划分, 并计算结果的 P, R, $F1$ 和 NMI 指标值, 结果如表 3 所示. 从表 3 可以看出, 与其他经典算法相比, NCSS 算法在诸如 Karate 跆拳道俱乐部与 Dophin 海豚相互交流关系之类的社区结构性较强的网络中, 综合评价指标 $F1$ 和标准互信息量的表现最好, 在真实网络上 NCSS 算法的精度比其他算法平均提高 6%, 说明 NCSS 算法能够准确发现网络的真实社区结构. 在 Football 和 Polbooks 网络中, 所得到的 $F1$ 和 NMI 均高于平均水平, 优于大部分算法, 仅比 IM 和 SG 算法略差. 根据表 1 可知, IM 和 SG 算法的复杂度均较高. 说明 NCSS 算法在合理的时间内可以完成较好的社团划分.

表 2 实验中用到的真实网络

Table 2 Real networks used in the experiment

Network	Abbreviation	V	E	Labeled
Zachary's karate club [42]	Karate	34	78	T
Dolphin social network [43]	Dolphin	62	159	T
American college football [8]	Football	115	616	T
Books about US politics [10]	Polbooks	105	441	T
Email communication network [44]	Email	1133	5451	F
Coauthorships in network science [2]	Netscience	1589	2742	F

表 3 NCSS 算法与其他算法关于有标记网络划分结果的比较

Table 3 Community detection results of NCSS and other algorithms on labeled networks

Network	Evaluation	GN	FG	IM	SG	WT	LV	LE	LPA	GCE	LFM	NCSS
Karate	P	0.971	0.971	0.971	1	0.941	0.971	1	1	0.971	0.882	1
	R	0.969	0.972	0.972	1	0.938	0.972	1	1	0.969	0.889	1
	$F1$	0.97	0.971	0.971	1	0.939	0.971	1	1	0.97	0.886	1
	NMI	0.836	0.837	0.837	1	0.731	0.837	1	1	0.836	0.580	1
Dolphin	P	0.984	0.935	0.967	0.967	0.951	0.935	0.951	0.967	0.871	0.935	0.984
	R	0.977	0.94	0.954	0.954	0.952	0.909	0.952	0.954	0.818	0.95	0.977
	$F1$	0.981	0.938	0.961	0.961	0.951	0.922	0.951	0.961	0.844	0.942	0.981
	NMI	0.89	0.652	0.814	0.814	0.711	0.693	0.711	0.89	0.509	0.707	0.89
Football	P	0.834	0.573	0.93	0.93	0.869	0.869	0.626	0.869	0.852	0.713	0.869
	R	0.772	0.49	0.891	0.891	0.808	0.808	0.568	0.808	0.8	0.642	0.808
	$F1$	0.802	0.53	0.91	0.91	0.837	0.837	0.595	0.837	0.825	0.675	0.837
	NMI	0.878	0.697	0.936	0.936	0.887	0.884	0.698	0.884	0.891	0.796	0.893
Polbooks	P	0.857	0.838	0.847	0.857	0.847	0.847	0.847	0.857	0.761	0.847	0.848
	R	0.765	0.639	0.721	0.747	0.646	0.646	0.646	0.689	0.665	0.786	0.786
	$F1$	0.808	0.725	0.779	0.798	0.733	0.733	0.733	0.764	0.658	0.745	0.811
	NMI	0.568	0.568	0.571	0.582	0.597	0.597	0.597	0.588	0.331	0.585	0.577

3.2.2 无标记网络

无标记网络上的社团划分难度较大. 将 NCSS 算法和比较算法应用于无标记网络进行社团划分. 由于无标记网络没有人工社团划分结果, 不能使用 P, R, $F1$ 和 NMI 指标来进行评价, 因此采用模块度来进行评价. 各算法在两个无标记网络上模块度的结果如表 4 所示. 由表 4 可知, 在 Email 网络中, 本文算法仅比 SG 算法略差, 与 LV 算法基本持平, 优于其他算法. 而 SG 算法的复杂度较高. 对于 Netscience 网络, 由于网络是非连通的, GN 算法和 SG 算法无法进行社团划分, 表现出对网络类型的依赖性. 在其余的算法比较中, NCSS 算法的模块度最高, 表现最好.

3.3 模拟数据集实验

为了进一步评估这些算法的性能, 我们采用 LFR [45] 人工合成基准网络进行实验, LFR 网络中节点的度分布及社团规模的分布均为幂律分布, 使其更接近真实网络. 实验主要通过设置以下参数来生

表 4　NCSS 算法与其他算法关于未标记网络划分结果的比较

Table 4　Community detection results of NCSS and other algorithms on unlabeled networks

Network	Evaluation	GN	FG	IM	SG	WT	LV	LE	LPA	GCE	LFM	NCSS
Email	Q	0.532	0.506	0.52	0.579	0.53	0.54	0.488	0.532	0.436	0.127	0.536
Netscience	Q	—	0.955	0.931	—	0.955	0.959	0.951	0.917	0.804	0.924	0.956

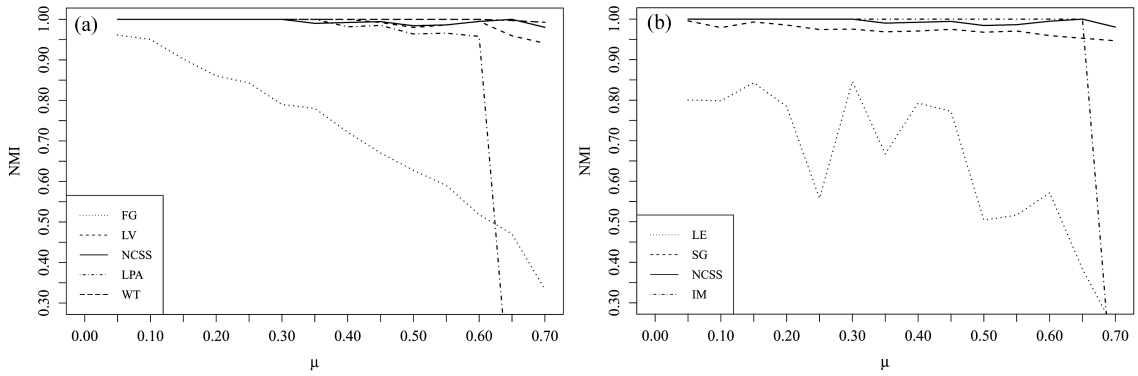


图 1　不同算法在基准网络上的 NMI 随 μ 的变化结果, n=1000, $c_{\min}$=10, $c_{\max}$=50

Figure 1　NMI of NCSS and other algorithms on benchmark networks with increasing μ and $n = 1000$, $c_{\min} = 10$, $c_{\max} = 50$. (a) NCSS versus FG, LV, LPA and WT; (b) NCSS versus LE, SG and IM

成复杂网络:

(1) 网络节点数目 n, 指定网络的节点规模;

(2) 网络节点的平均度 k;

(3) 网络节点允许的最大度数 $k_{\max}$;

(4) 网络拓扑混合参数 μ, 指定社区内节点与社区外节点连接的边数占总边数的比例, 混合参数越大, 则生成的网络结构就越不明显, 它决定了社区的结构特性;

(5) 节点度数指数分布参数 ε_1, LFR 标准网络中的节点度服从参数为 ε_1 的幂律分布;

(6) 社区规模指数分布参数 ε_2, 合成的原始社区规模分布满足参数为 ε_2 的幂律分布;

(7) 最小社区大小 $c_{\min}$, 指定生成的原始社区的最小规模;

(8) 最大社区大小 $c_{\max}$, 指定生成的原始社区的最大规模.

按照文献 [46] 中的实验设计, LFR 基准网络的参数设置如下: 网络规模 n 取值为 1000; 最小社团规模 $c_{\min}$ 取值为 10 或 20; 混合参数 μ 从 0.05 变化到 0.7, 间隔为 0.05. 其他参数保持和文献 [46] 中的一致, 即节点的平均度 k 为 20; 最大度 $k_{\max}$ 为 2.5 倍 k; 最大社团规模 $c_{\max}$ 为 5 倍 $c_{\min}$; 节点度与社团规模的幂律分布指数分别为 $\varepsilon_1 = -2$, $\varepsilon_2 = -1$.

我们生成了两个网络规模相同的 LFR 模拟网络, 两个模拟网络中的社团规模不同. 首先分析不同算法随混合参数 μ 取不同值的划分效果. 图 1 表示在 $n = 1000$, $c_{\min} = 10$, $c_{\max} = 50$, 图 2 表示在 $n = 1000$, $c_{\min} = 20$, $c_{\max} = 100$, 混合参数 μ 分别取不同值时, 得到对应的 NMI 值. 图 1 和 2 的两个子图是 NCSS 和不同算法的比较曲线. 图 1 和 2 的横轴表示 μ 值, 纵轴表示 NMI 值.

从图 1 和 2 上可以看出, 在 μ 值较小的情况下 ($\mu < 0.2$), 9 种算法所表现出来的精度差别不大, 特别是在 $c_{\min} = 20$, $c_{\max} =$ 100 的情况下. 然而随着 μ 值的上升, 网络中社区结构逐渐模糊, 不同算法所能体现出来的性能差别开始变大. 由图 1 和 2 可知, NCSS 对应的 NMI 值一直高于除 WT 以外

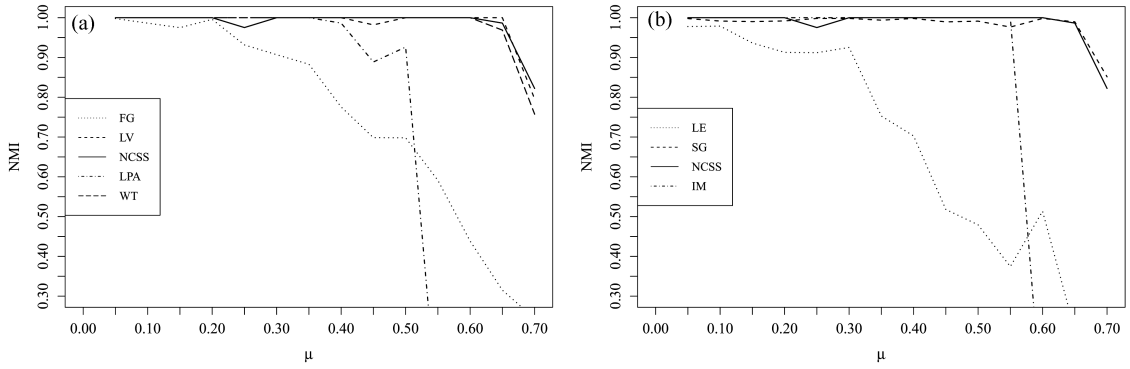


图 2 不同算法在基准网络上的 NMI 随 μ 的变化结果, n=1000, $c_{\min}$=20, $c_{\max}$=100

Figure 2 NMI of NCSS and other algorithms on benchmark networks with increasing μ and $n = 1000$, $c_{\min} = 20$, $c_{\max} = 100$. (a) NCSS versus FG, LV, LPA and WT; (b) NCSS versus LE, SG and IM

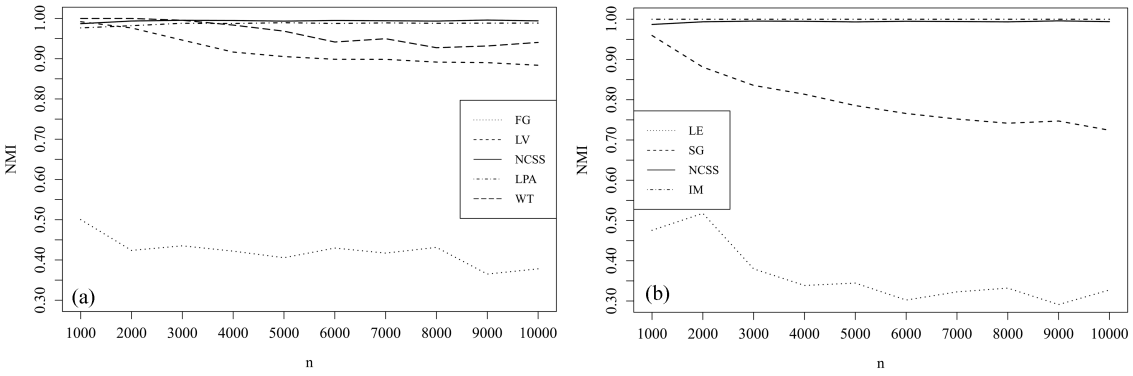


图 3 不同算法在基准网络规模变化下的 NMI 结果, μ=0.65, $c_{\min}$=10, $c_{\max}$=50

Figure 3 NMI values of NCSS and other algorithms on benchmark networks with different scales and μ=0.65, $c_{\min} = 10$, $c_{\max} = 50$. (a) NCSS versus FG, LV, LPA and WT; (b) NCSS versus LE, SG and IM

其他算法, 并且与 WT 的差距最多不超过 0.02. 基于随机游走的 WT 算法的时间复杂度主要依赖网络中边的数目, 时间复杂度为 $O(mn^2)$, 而 NCSS 的时间复杂度仅为 $O(kn)$, 综合分析, NCSS 算法效率优于 WT 算法. 为了更进一步说明 NCSS 算法和其他算法性能差异, 我们固定 μ 为 0.65, 把网络规模按间隔为 1000 从 1000 逐步提升到 10000, 然后对比分析不同算法性能随着网络规模的变化. 图 3 给出了 NCSS 算法和其他算法在网络规模不同下的 NMI 指标变化曲线.

如图 3 所示, 可以发现随着网络规模的增加不同算法的性能差别同样逐渐加大. 原本在网络规模为 1000 时表现最好的 WT 算法, 随着网络规模的增加性能在变差. 而本文所提出的 NCSS 算法依然保持较高的水准, IM 算法和 NCSS 算法在网络规模加大时性能差异不明显, IM 算法和 NCSS 算法优于其他算法, 但是 IM 算法的时间复杂度为 $O(n(m+n))$ 也高于 NCSS 算法的 $O(kn)$, 结果说明 NCSS 算法能够应用于大规模网络上的社团划分.

4 结束语

本文提出了一种基于核心节点中心度和节点与社团之间的结构关系强度的社团划分算法 NCSS.

NCSS 算法从复杂网络的全局出发, 借鉴 PageRank 算法的思想, 定义了节点中心度, 用中心度对节点进行排序, 进而有效选择合适的节点作为潜在的核心节点; 利用复杂网络的局部结构信息, NCSS 算法根据节点与社团间的社团依赖度和结构关系强度建立节点扩张策略, 从各个潜在种子节点开始不断迭代发现其所代表的社团. 算法与其他经典的全局社区发现算法与局部社区发现算法在多种真实网络与模拟基准网络数据集上进行了比较, 实验结果表明算法在大多数数据集上具有一定优势, 不论是有标记网络中的 $F1$ 值和 NMI 值, 还是在无标记网络中的模块度值都处在较高的水准, 同时算法复杂度较小, 执行效率较快. 说明 NCSS 算法不仅能够满足大规模复杂网络的社团划分, 而且能够保持较高精度.

参考文献

1 Newman M E J, Watts D J. Renormalization group analysis of the small-world network model. Phys Lett A, 1999, 263: 341–346

2 Wang G X, Shen Y. Modularity matrix of networks and the measure of community structure. Acta Phys Sin, 2010, 59: 842–850 [王高峡, 沈轶. 网络的模块矩阵及其社团结构指标. 物理学报, 2010, 59: 842–850]

3 Newman M E J. Finding community structure in networks using the eigenvectors of matrices. Phys Rev E, 2006, 74: 036104

4 Shao F, jiang G P. Optimal traffic routing strategy based on community structure. Acta Phys Sin, 2011, 60: 852–858 [邵斐, 蒋国平. 基于社团结构的负载传输优化策略研究. 物理学报, 2011, 60: 852–858]

5 Ino H, Kudo M, Nakamura A. Partitioning of web graphs by community topology. In: Proceedings of the 14th international conference on World Wide Web, New York, 2005. 661–669

6 Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Computation Biol, 2007, 3: e107

7 Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. Bell syst tech J, 1970, 49: 291–307

8 Newman M E J. Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 2006, 103: 8577–8582

9 Shiga M, Takigawa I, Mamitsuka H. A spectral clustering approach to optimally combining numericalvectors with a modular network. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, New York, 2007. 647–656

10 Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 2002, 99: 7821–7826

11 Newman M E J. Fast algorithm for detecting community structure in networks. Phys Rev E, 2004, 69: 066133

12 Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks. J Stat Mech Theory Exp, 2008, 30: 155–168

13 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E, 2007, 76: 036106

14 Pan Y, Li D H, Liu J G, et al. Detecting community structure in complex networks via node similarity. Physica A Stat Mech its Appl, 2010, 389: 2849–2857

15 Zanjani A A H, Darooneh A H. Finding communities in linear time by developing the seeds. Phys Rev E, 2011, 84: 036109

16 Wang X Y, Zhao Z X. Partitioning community structure in complex networks based on node dependent degree. Acta Phys Sin, 2014, 63: 178901 [王兴元, 赵仲祥. 基于节点间依赖度的社团结构划分方法. 物理学报, 2014, 63: 178901]

17 Hour S, Kan L. Structural and regular equivalence of community detection in social networks. In: Proceedings of Advances in Social Networks Analysis and Mining (ASONAM), Beijing, 2014. 808–813

18 Riolo M A, Newman M E J. First-principles multiway spectral partitioning of graphs. J Complex Netw, 2014, 2: 121–140

19 Darst R K, Nussinov Z, Fortunato S. Improving the performance of algorithms to find communities in networks. Phys Rev E, 2014, 89: 032809

20 Tsourakakis C, Gkantsidis C, Radunovic B, et al. Fennel: streaming graph partitioning for massive scale graphs. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, New York, 2014. 333–342

21 Mokross K, Ryder T B, Côrtes M C, et al. Decay of interspecific avian flock networks along a disturbance gradient in Amazonia. Proc Biol Sci, 2013, 281: 20132599

22 Albert R, DasGupta B, Mobasheri N. Topological implications of negative curvature for biological and social networks. Phys Rev E, 2014, 89: 032811

23 Fu L D, Gao L, Ma X K. A centrality measure based on spectral optimization of modularity density. Sci Sin Inform, 2012, 42: 550–560 [付立东, 高琳, 马小科. 基于社团检测的复杂网络中心性方法. 中国科学: 信息科学, 2012, 42: 550–560]

24 Pons P, Latapy M. Computing communities in large networks using random walks. J Graph Algorithms Appl, 2006, 10: 191–218

25 Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E, 2006, 74: 016110

26 Clauset A. Finding local community structure in networks. Phys Rev E, 2005, 72: 026132

27 Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure in complex networks. New J Phys, 2009, 11: 033015

28 Lee C, Reid F, McDaid A, et al. Detecting highly overlapping community structure by greedy clique expansion. arXiv, 2010: 1002.1827

29 Liu X, Yi D Y. Complex network community detection by local similarity. Acta Automatica Sinica, 2011, 37: 1520–1529 [刘旭, 易东云. 基于局部相似性的复杂网络社区发现方法. 自动化学报, 2011, 37: 1520–1529]

30 Chen Q, Wu T T, Fang M. Detecting local community structures in complex networks based on local degree central nodes. Physica A: Stat Mech Its Appl, 2013, 392: 529–537

31 Yuan C, Chai Y. Method for local community mining in the complex networks. Acta Automatica Sinica, 2014, 5: 921–934 [袁超, 柴毅. 复杂网络的局部社团结构挖掘算法. 自动化学报, 2014, 40: 921–934]

32 Perozzi B, Akoglu L, Iglesias Sánchez P, et al. Focused clustering and outlier detection in large attributed graphs. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2014. 1346–1355

33 Moradi F, Olovsson T, Tsigas P. A local seed selection algorithm for overlapping community detection. In: Proceedings of Advances in Social Networks Analysis and Mining (ASONAM), Beijing, 2014. 1–8

34 Zalik K R, Zalik B. A local multiresolution algorithm for detecting communities of unbalanced structures. Physica A: Stat Mech Its Appl, 2014, 407: 380–393

35 Gan W Y, He N, Li D Y, et al. Community discovery method in networks based on topological potential. J Software, 2009, 20: 2241–2254 [淦文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法. 软件学报, 2009, 20: 2241–2254]

36 Wu F, Huberman B A. Finding communities in linear time: a physics approach. Eur Phys J B Condens Mat Complex Syst, 2004, 38: 331–338

37 He D X, Zhou X, Wang Z, et al. Community mining in complex networks — clustering combination based genetic algorithm. Acta Automatica Sinica, 2010, 36: 1160–1170 [何东晓, 周栩, 王佐, 等. 复杂网络社区挖掘 — 基于聚类融合的遗传算法. 自动化学报, 2010, 36: 1160–1170]

38 Okamoto H. Local detection of communities by attractor neural-network dynamics. Artificial Neural Networks, 2015, 4: 115–125

39 Zhang Z Y. Community structure detection in social networks based on dictionary learning. Sci China Inf Sci, 2011, 41: 1343–1355 [张忠元. 基于字典学习的网络社团结构探测算法. 中国科学: 信息科学, 2011, 41: 1343–1355]

40 Adamic L A, Huberman B A. Power-law distribution of the world wide web. Science, 2000, 287: 2115–2115

41 Page L, Brin S, Motwani R, et al. The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, 1998. 161–172

42 Zachary W W. An information flow model for conflict and fission in small groups. J Anthropol Res, 1977, 33: 452–473

43 Lusseau D, Schneider K, Boisseau O J, et al. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Behav Ecol Sociobiol, 2003, 54: 396–405

44 Guimera R, Danon L, Diaz-Guilera A, et al. Self-similar community structure in a network of human interactions. Phys Rev E, 2003, 68: 065103

45 Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Phys Rev E, 2008, 78: 046110
46 Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E, 2009, 80: 056117

NCSS: an effective and efficient complex network community detection algorithm

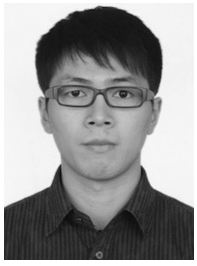
Zhongming HAN*, Xusheng TAN, Yan CHEN & Dagao DUAN

School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048*, China*
*E-mail: hanzm@th.btbu.edu.cn

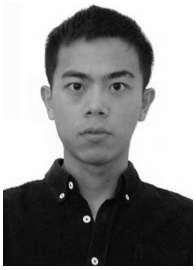
Abstract Community Detection has significant theoretical and practical importance. Global community detection algorithms can achieve high precision, but with worst-case time complexity. Local community detection algorithms mostly have low precision with poor stability. In this paper, a NCSS algorithm is proposed based on node centrality and the structural strength between nodes and local communities in a complex network. The NCSS algorithm ranks core nodes based on global information about the complex network. It then creates initial communities, which can overcome the instability disadvantage of random initial nodes in other local community detection algorithms. During the community-expansion process, the NCSS algorithm calculates only the structural relationship strength between the margin node and related local communities, thus reducing time complexity. Comprehensive experiments are conducted based on different benchmark networks and real complex networks. Typical global community algorithms and local community detection algorithms are selected for comparison. On real networks, the precision of the NCSS algorithm is increased by 6% on average. On benchmark networks, the precision can be increased by as much as 18%.

Keywords complex network, community structure, community detection, key node, structural strength

Zhongming HAN was born in 1972. He received his Ph.D. degree from Donghua University, Shanghai in 2006. Currently, he is an Associate Professor at Beijing Technology and Business University. His research interests include Big Data, Web Data Mining and Social Computing. Dr. HAN is a member of the China Computer Federation and ACM.



Xusheng TAN was born in 1989. He is currently working toward his master's degree in computer science at Beijing Technology and Business University. His research interests include machine learning and social community detection.



Yan CHEN was born in 1991. He is a student member of the China Computer Federation. His current research interests include machine learning and social network analysis.

Dagao Duan was born in 1976. He received his Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing in 2005. Currently, he is an Associate Professor at Beijing Technology and Business University. His research interests include Social Computing and Multimedia Information Processing. He is a member of the China Computer Federation and ACM.