文章编号：2095-2163(2020)03-0236-06 中图分类号：TP301.6 文献标志码：A

基于随机游走的级联网络社区发现算法

王 亮，杨海陆，陈德运
（哈尔滨理工大学 计算机科学与技术学院，哈尔滨 150080）

摘 要：随着 Web 2.0 的不断推广以及社交应用的不断普及，在线社交网络结构分析得到了各领域学者的广泛关注。社区是网络中内嵌的密集群组，保证了社区内部用户的强相关性和一致性，因此广泛应用于病毒防护、商品推荐等现实系统。本文提出一种基于随机游走的级联网络社区发现算法，主要解决非直连节点间的相似性度量问题。提出一种基于 2-hop 随机游走的局部可达性计算方法，通过对游走终点一致的节点进行层次聚类，社区结构可在较短的时间内迭代生成。实验结果表明，本方法在级联网络社区发现中具有较高的性能和效率，对星形社区具有较高的匹配性。

关键词：级联网络；社区发现；随机游走；层次聚类

Community detection in cascade networks using random walks

WANG Liang, YANG Hailu, CHEN Deyun
(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

【Abstract】 With the continuous promotion of Web 2.0 and the continuous popularization of social applications, the structural analysis of online social networks has attracted extensive attention from scholars in various fields. Community is a dense group embedded in the network, which ensures the strong correlation and consistency of users within the community. Therefore, it is widely used in virus protection, commodity recommendation, and other practical systems. In this paper, a cascade network community discovery algorithm based on the random walk is proposed to solve the similarity measurement problem of non-directly connected nodes. The paper proposes a local reachability calculation method based on 2-hop random walks. The community structure can be generated iteratively in a short time by hierarchical clustering of nodes with consistent endpoints. The experimental results show that this method has high performance and efficiency in cascaded network community detection and has a high matching to the star community.

【Key words】 cascaded network; community detection; random walks; hierarchical clustering

0 引 言

随着 5G 技术的不断推广以及“互联网+”战略的积极部署，微信、微博等社交网络早已融入人们的日常生活及学习之中。更多的人选择在社交网络进行交流、发表观点，社交网络早已成为现实世界在网络空间的真实缩影，社交网络结构分析也因此成为各学科领域的研究热点和学术前沿问题。

社区结构(Community)是社交网络中的重要结构，社区内部用户之间的链接较为紧密，社区之间的用户链接较为稀疏。从这一角度来看，社区结构可以广泛应用于诸如商品推荐、病毒防护等应用系统。例如，可以对社区内部用户推荐相似的产品，或切断社区之间的联系降低病毒的传播范围等。级联网络是一种特殊的社交网络，微博网络就是典型的级联网络，这种网络中普遍存在星形结构(形成于对高影响力用户的大量关注)，就使得社区内部节点可能不具有直接联系，这给社区识别任务带来了极大的挑战。

目前来看，基于随机游走的动态算法能够有效解决这一问题，其基本原理利用了社区结构对“游走者”的捕获作用。当“游走者”游走到社区边缘时，由于社区内部链接远大于社区间的链接，因此“游走者”有极大的概率再次游走回社区内部。由此可见，位于相同社区的两节点，其随机游走的探测路径应该具有较高的重叠性。Rosvall 等人[1]引入一种信息论方法，揭示加权有向网络中的社区结构。通过将网络随机游走的概率流作为真实系统中的信息流，社区结构可以由概率流压缩生成。Huang 等人[2]探讨随机游走思想在构建社区模块度函数中的应用。在这种方法中，模块度函数被定义为社区

基金项目：黑龙江省自然科学基金面上项目(F2016024)；黑龙江省博士后资助经费(LBH-Z15095)；黑龙江省普通高等学校创新人才培养计划(UNPYSCT-2017094)；国家级大学生创业创新训练计划(201810214020)。

作者简介：王 亮(1997-)，男，本科生，主要研究方向：社会计算；杨海陆(1985-)，男，副教授，硕士生导师，主要研究方向：社会计算、知识图谱、移动互联网安全等。

收稿日期：2019-12-22

引起的马尔可夫随机游走和一个零概率模型的之间的差异。刘阳等人[3]提出一种边权预处理方法，根据多重随机游走对网络连接的遍历情况重新衡量网络边权。预处理后的边权作为网络拓扑的有效补充信息，能够将网络结构去模糊化，从而改善现有算法的社区发现性能。杨海陆等人[4]设计了一种基于 2-hop 互随机游走的异质网络节点相似性度量函数，通过将相似性函数推广到层次聚类并设计相应的相似矩阵校准方案，异质社区识别任务可以在较短的时间内迭代完成。辛宇等人[5]提出一种改进的语义社会网络重叠社区发现随机游走策略，以 LDA (Latent Dirichlet Allocation)算法为基础建立语义空间，实现节点语义信息到语义空间的量化映射。在后续的研究中，Xin 等人[6]提出适应性随机游走概念，将模型推广至动态社区识别任务。在最近的研究中，Okuda 等人[7]提出了一种带约束随机行走相似性度量方法检测图的社区结构。其基本思想是随机游走能够途经相似节点的起始点属于同一社区的可能性较大。

上述方法在社区识别领域得到了普遍认可，但忽略了以下两方面要素。首先，没有对游走长度进行约束，这使得“游走者”能够捕获的信息处于不确定状态；其次，上述方法普遍拒绝“游走者”重复访问已经通过的顶点，而这一情况在密度较高的社区中是普遍存在的。

为了解决这一问题，提出一种基于 2-hop 结构探测的社区识别方法。基本思路是：如果“游走者”的起始点位于相同社区，则“游走者”能够囊括的 2-hop 探测节点集应该具有较高的重叠性。该思想充分考虑了节点的直接相似性和结构相似性，因此识别出的社区结构具有一定的稳定性。

1　定义及算法

定义 1　级联网络是一组由节点、链接关系组成的复杂网络模型 $G=\{V,E\}$，其中 V 为网络节点集合，$E=\{(v_i,v_j)\,|\,v_i,v_j\in V\}$ 为链接集合。

级联网络通常呈现出较强的星形特征，节点度服从幂律分布。微博、知乎等社交网络就是常见的级联网络，其中用户为网络节点，用户之间的社交关系为网络链接。

定义 2　级联网络社区结构被定义为一组节点集 $C=\{v_i,\ldots,v_j\}$。识别社区的目的在于发现网络社区集合 $C=\{C_1,C_2,\ldots,C_k\}$，满足 $C_i\cap C_j\neq\varnothing, i\neq j$。

定义 3　节点 u 的邻居节点 $N(u)$ 定义为：

$$N(u)=\{v\,|\,(u,v)\in E\}, \tag{1}$$

如果节点 u 和节点 v 互为邻居节点，则节点 u 和节点 v 之间必有直接链接。

定义 4　随机游走。从 $G=\{V,E\}$ 中的任意节点开始，随机游走将以转移概率 $P=1/|N(u)|$ 转移到邻居节点，直到遍历全图。

定义 5　节点 u 的探测集合定义为节点 u 能够探测到的 2-hop 随机游走节点集合，即：

$$e(u)=(n_i\,|\,i\in N(u));(n_j\,|\,j\in N(i)), \tag{2}$$

如果 2 个节点位于同一社区，由于节点之间存在强相关性，因此以这两个节点为起始节点的随机游走模型所探测到的终点集应具有较高的重叠性。

如图 1 所示，节点 $n_1\sim n_9$ 的探测集合分别为：

$e(n_1)$：(n_2,n_4)；(n_1,n_5)，(n_1,n_3,n_5).

$e(n_2)$：(n_1,n_5)；(n_2,n_4)，(n_2,n_4,n_6,n_8).

$e(n_3)$：(n_4)；(n_1,n_3,n_5).

$e(n_4)$：(n_1,n_3,n_5)；(n_2,n_4)，(n_4)，(n_2,n_4,n_6,n_8).

$e(n_5)$：(n_2,n_4,n_6,n_8)；(n_1,n_5)，(n_1,n_3,n_5)，(n_5,n_7,n_9)，(n_5,n_9).

$e(n_6)$：(n_5,n_7,n_9)；(n_2,n_4,n_6,n_8)，(n_6)，(n_6,n_8).

$e(n_7)$：(n_6)；(n_5,n_7,n_9).

$e(n_8)$：(n_5,n_9)；(n_2,n_4,n_6,n_8)，(n_6,n_8).

$e(n_9)$：(n_6,n_8)；(n_5,n_7,n_9)，(n_5,n_9).

这里，分号前半段代表以 u 为起始点的 1-hop 探测节点，记为 $e_1(u)$；后半段代表 2-hop 探测节点，记为 $e_2(u)$。研究发现，n_2 和 n_4 的探测集合具有较高的重叠性，因此相似性较高。

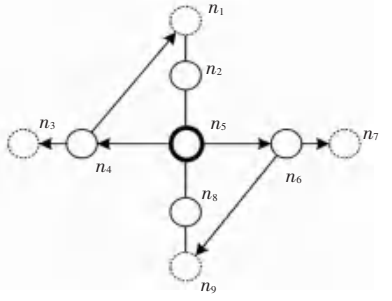

图 1　级联网络中的局部随机游走

Fig. 1　Local random walk in cascade network

定义 6　探测集重叠度。探测集 $e(u)$ 与探测集 $e(v)$ 的重叠度定义为：

$$O(e(u),e(v))=\frac{|e(u)|+|e(v)|-|e(u)\cap e(v)|}{|e(u)|+|e(v)|}, \tag{3}$$

定义 7 节点相似性度量函数。对于任意节点 $u,v \in V$,u 和 v 之间的相似性定义为:

$$sim(u,v) = \frac{O(e_1(u),e_1(v)) + O(e_2(u),e_2(v))}{2}, \quad (4)$$

其中,分子部分前半式度量节点间的直接相关性,后半式度量节点间的结构相关性。

根据式(4),可以计算出图 1 各节点之间的相似关系,例如:

$$sim(n_2,n_4) = \frac{1}{2}\cdot\frac{4}{5} + \frac{1}{2}\cdot\frac{12}{13} = 0.861\ 5, \quad (5)$$

$$sim(n_2,n_8) = \frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{8}{12} = 0.583\ 3. \quad (6)$$

可以看出,虽然 n_2 和 n_4 之间没有直接链接,但相近的游走终点增加了 n_2 和 n_4 位于同一社区的可能性。而对于 n_2 和 n_8 而言,虽然二者同样相邻 2-hop 距离,但将二者划为同一社区显然会降低社区的结构稳定性。

2 基于层次聚类的社区识别算法

本节给出基于层次聚类的社区发现算法 HCD (Hierarchical-based Community Detection)。算法采用贪婪思想,每次操作选取相似性最高的两个节点/社区进行合并。实现过程如算法 1 所示。

算法 1 基于层次聚类的社区发现算法 HCD

输入:级联网络 $G = \{V,E\}$

输出:社区集合 $\boldsymbol{C} = \{C_1,C_2,\ldots,C_k\}$

(1)计算全局相似性度量矩阵 $\boldsymbol{S}$;

(2)初始化 $\boldsymbol{C} = \{C_1,C_2,\ldots,C_n\}$,$C_i = \{V_i\}$;

(3)初始化 $C_{back} = \{C_1,C_2,\ldots,C_n\}$,$C_i = \{V_i\}$

(4)Set $t = 1$;

(5)选定 S_{ab} 最大的社区 C_a,C_b;

(6) $C_k \leftarrow C_a \cup C_b$;

(7) $\boldsymbol{C} = \boldsymbol{C}\backslash\{C_a,C_b\} \cup C_k$;

(8) $C_{back} = C_{back} \cup C_k$;

(9) If $|\boldsymbol{C}| \neq 1$ then

(10) Go to step (17);

(11) Else

(12) $C_{tree} = \{C_{back}^1,C_{back}^2,\ldots,C_{back}^H\}$;

(13) $\boldsymbol{C} = \arg\max_{C\in C_{tree}} EQ(C)$;

(14)Return $\boldsymbol{C}$.

(15)根据算法 2 校准矩阵 $\boldsymbol{S}$;

(16)删除相似性矩阵行列 a,b 并添加行列 k;

(17)Set $t = t + 1$ and go to step (5).

对于算法 1,当两节点合并为社区后,其余节点需要重新计算与该社区之间的相似关系,因此需要校准相似性度量矩阵 $\boldsymbol{S}$。简单的扩展即可将节点之间的相似性度量扩展到社区之间的相似性程度,具体步骤如算法 2 所述。

算法 2 相似性矩阵校准算法

输入:相似性度量矩阵 $\boldsymbol{S}$,新生成社区 C_{uv};

输出:校准后的相似性度量矩阵 $\boldsymbol{\Delta S}$

(1)对于 $e(u)$ 和 $e(v)$,合并 $e_1(u)$,$e_1(v)$ 以及 $e_2(u)$,$e_2(v)$ 中的 u,v 后分别取并集;

(2) $e_1(x)$ 中同时包含 u,v,替换 u,v 为 uv,删除 u,v 的 1-hop 节点集;

(3) $e_1(x)$ 中包含 u,v 中的一个,替换 u,v 为 uv;

(4) $e_1(x)$ 包含 u,v 的 1-hop 节点,如果 $e_2(x)$ 中包含 u,v,删除 $e_2(x)$ 中的 u,v,添加社区 C_{uv};

(5)与 u,v 距离超出 2-hop 的节点不做处理;

(6)Return.

如图 2 所示,假设节点 n_2 和 n_4 合并为社区 C_{24} 可得变化后的探测集合为:

$e(n_1)$:$(n_{2,4})$;(n_1, n_3, n_5).

$e(n_{2,4})$:(n_1, n_3, n_5);$(n_{2,4})$,$(n_{2,4})$,$(n_{2,4}, n_6, n_8)$.

$e(n_3)$:$(n_{2,4})$;(n_1, n_3, n_5).

$e(n_5)$:$(n_{2,4}, n_6, n_8)$;(n_1, n_3, n_5),(n_5, n_7, n_9),(n_5, n_9).

$e(n_6)$:(n_5, n_7, n_9);$(n_{2,4}, n_6, n_8)$,(n_6),(n_6, n_8).

$e(n_7)$:(n_6);(n_5, n_7, n_9).

$e(n_8)$:(n_5, n_9);$(n_{2,4}, n_6, n_8)$,(n_6, n_8).

$e(n_9)$:(n_6, n_8);(n_5, n_7, n_9),(n_5, n_9).

此时重新计算节点间的相似性可得:

$$sim(n_{2,4},n_8) = \frac{1}{2}\cdot\frac{2}{5} + \frac{1}{2}\cdot\frac{6}{10} = 0.500\ 0. \quad (7)$$

略小于合并前的相似程度 0.583 3。

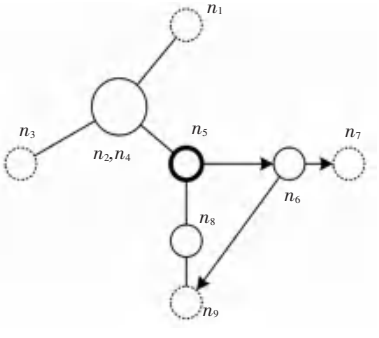

图 2 基于节点合并的社区形成

Fig. 2 Community formation based on node combination

算法的迭代过程最终会生成一个层次化树,因此为了在合适的位置进行分割以获得最优的社区结

构，采用模块度函数度量社区质量，模块度最大的层级即为所求。

算法 1 在生成社区时采用 C_{back} 记录生成社区，采用 C 作为算法计数器，这种策略使得下次合并社区时之前的合并节点仍有机会加入其它社区，因此社区具有重叠性。EQ 函数广泛用于度量重叠社区的质量，定义为：

$$EQ = \frac{1}{D}\sum_{i}\sum_{v\in C_i, w\in C_i}\frac{1}{O_v O_w}\left[A_{vw} - \frac{d_v d_w}{D}\right]. \quad (8)$$

这里 d_v 为 v 的节点度，$D = \sum_{vw} A_{vw}$ 为网络节点的总度数；A_{vw} 为网络邻接矩阵中的元素；O_v 为节点 v 归属的重叠社区个数；C_i 为网络中的第 i 个重叠社区。

3　实验结果与分析

本节给出算法在真实数据集上的运行结果。实验的运行环境为 Intel Pentium G3260 3.3GHz 处理器，4 GB 内存，Windows 7 操作系统，采用 C++与 Matlab R2013b 混合编程。

3.1　人工合成网络社区识别结果

LFR Benchmark 人工合成网络在生成时会根据规则生成 Ground-truth 社区（固有存在的社区结构），因此与固有社区之间的差距越小，算法的性能越优。用归一化互信息度量社区划分结果 C_A 与 C_B 之间的差别，其定义为：

$$NMI(C_A, C_B) = \frac{H(C_A) + H(C_B) - H(C_A, C_B)}{\sqrt{H(C_A)H(C_B)}}. \quad (9)$$

其中，$H(C)$ 表示划分 C 的香农熵，当划分 C_A 与划分 C_B 完全一致时 $NMI(C_A, C_B) = 1$，当划分 C_A 与划分 C_B 完全不同时 $NMI(C_A, C_B) = 0$。

利用 LFR-Benchmark 提供的 Matlab 数据生成器生成了网络规模为 $N = 1\,000$ 以及 $N = 5\,000$ 的复杂网络，其中最小社区尺寸 $c_{\min} = 10$；最大社区尺寸 $c_{\max} = 50$；节点的社区从属度 $o_m = 2$；混合参数为 $\mu = 0.3$；权重分配参数设为 $\mu_w = 0.1$；社区重叠度 γ 选取 0~0.5。比对算法选择了较为典型的 6 种社区识别算法，分别为：GN、FN、LFM、COPRA、CPM 以及 LPA。实验结果如图 3~图 6 所示。

从实验中可以看出，当 $N = 1\,000$ 时，HCD 方法的 NMI 指标接近于 COPRA 方法，略优于 LPA 以及 LFM，优于 GN 以及 FN。主要原因在于 GN 和 FN 面向非重叠社区识别，另外由于级联网络星形结构较为丰富，因此基于模块度优化的社区识别方法性能有所下降。

当 $N = 5\,000$ 时，所有算法性能均有所提升，但总体趋势不变。对于识别出的社区个数，当 $N = 1\,000$ 时 HCD 与 Ground-truth 社区个数较为接近。同样满足这一情况的还有 COPRA 以及 LFM。在 $N = 5\,000$时 CPM 算法同样与 Ground-truth 较为接近，一个可能的原因在于大规模网络社区粒度较低。

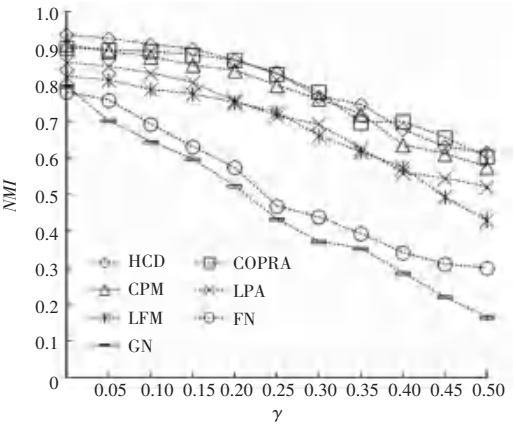

图 3　各算法的 *NMI* 指标（$N = 1\,000$）

Fig. 3　The *NMI* score of each algorithm ($N = 1\,000$)

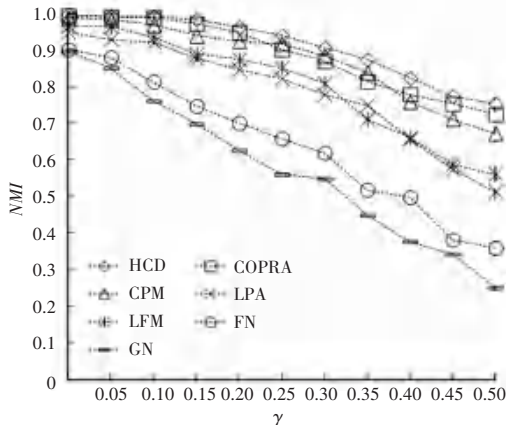

图 4　各算法的 *NMI* 指标（$N = 5\,000$）

Fig. 4　The *NMI* score of each algorithm ($N = 5\,000$)

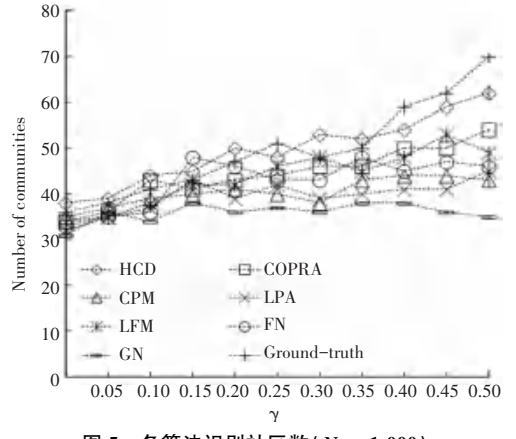

图 5　各算法识别社区数（$N = 1\,000$）

Fig. 5　The number of detected communities ($N = 1\,000$)

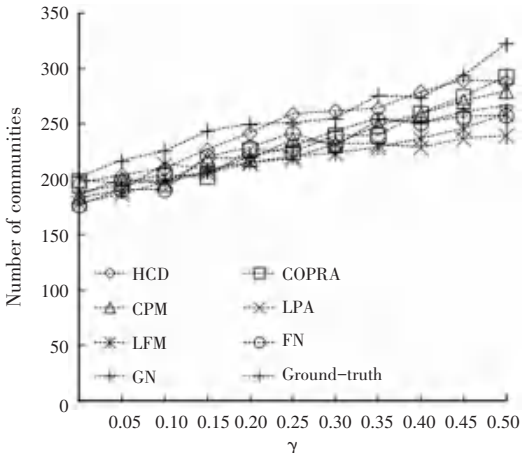

图 6 各算法识别社区数(N = 5 000)

Fig. 6 The number of detected communities (N = 5 000)

3.2 真实网络社区识别结果

本部分选择新浪微博作为实验数据。由于新浪微博 API 存在爬取数量限制,因此使用 Python 编写了面向网页的微博爬虫程序,结果存于 MySQL 数据库中。

收集了 2018 年 10 月~2019 年 9 月共 12 个月用户所发的微博帖子,选取任意网络节点作为初始爬取节点,采用自底向上的方法爬取初始节点的 6 层邻居结构。向上爬取的主要原因在于大多数微博用户的粉丝数量远远大于关注数量,因此如果向下爬取数据,广度优先会造成过大的时间开销。最终爬取到的数据量较为庞大,因此过滤掉了微博数少于 50 的用户以及关注数/被关注数少于 5 的用户。

研究中关注于社区模块度以及社区个数,实验结果如图 7、图 8 所示。

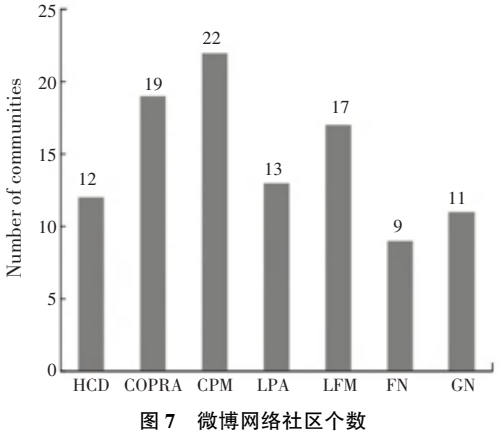

图 7 微博网络社区个数

Fig. 7 Number of communities in micro-blog networks

从实验结果来看,HCD 算法的模块度函数略小于 COPRA,优于其他 5 种方法。原因在于微博网络具有典型的级联特性,星形结构更加清晰,因此社区粒度普遍较低。HCD 识别社区个数较少,这表明 HCD 的局部游走策略是有效的,并且具有较高的效率。

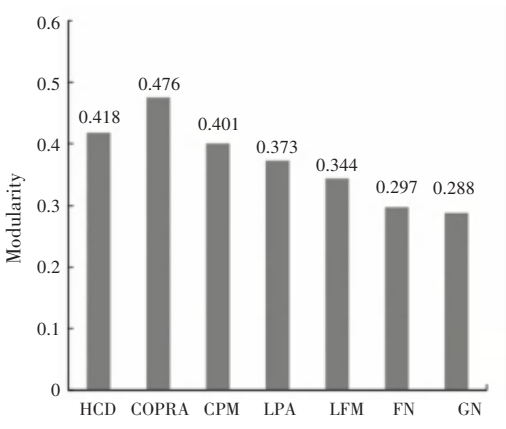

图 8 微博网络社区模块度

Fig. 8 Modularity of communities in micro-blog networks

为了更清晰地给出 HCD 算法的社区识别结果,研究还选择了节点数较少的 Football 网络给出社区识别结果。如图 9 所示,HCD 识别出的社区结构具有较低的粒度,但总体来看,仍然保证了较高的紧密性。

图 9 Football 网络社区识别结果

Fig. 9 Community structure in Football networks

4 结束语

针对基于随机游走的社区识别方法无法有效识别级联网络中的星形社区这一问题,提出一种带约束随机游走的社区识别方法 HCD。首先定义了随机游走的探测集合以及探测集重叠性度量函数,然后设计了一种面向社区迭代的相似性度量函数以及相似性矩阵校准方法,最后采用层次化聚类方法输出社区结果。在人工合成网络和真实网络上的实验结果表明,HCD 算法能够有效地度量非直连节点之间的相似性,在星形社区识别中具有较高的性能和效率。

参考文献

[1] ROSVALL M, BERGSTROM C T. Maps of random walks on complex networks reveal community structure[J]. Proceedings of the National Academy of Sciences of the United States of America, 2018,105(4): 1118.

(下转第 245 页)