

基于社团密合度的复杂网络社团发现算法

陈东明, 王云开, 黄新宇, 王冬琦
(东北大学 软件学院, 辽宁 沈阳 110169)

摘 要: 传统的社团发现算法大多存在划分效果和复杂度相矛盾的问题,为了解决该问题,提出一种新的单社团结构评价标准——社团密合度(group density).在此基础上,设计了一种基于凝聚思想的社团发现算法,该算法通过不断融合小社团,使网络的社团结构向平均社团密合度最大的方向发展,并使用模块度检测算法的划分结果.通过与经典的 GN, Fast Newman, LPA 等算法对多个数据集进行实验对比,验证了本文算法在获得较好的划分效果的同时具有较低的时间复杂度.

关 键 词: 复杂网络; 社团结构; 社团发现; 模块度; 社团密合度

中图分类号: TP 393 **文献标志码:** A **文章编号:** 1005-3026(2019)02-0186-06

Community Detection Algorithm for Complex Networks Based on Group Density

CHEN Dong-ming, WANG Yun-kai, HUANG Xin-yu, WANG Dong-qi
(School of Software, Northeastern University, Shenyang 110169, China. Corresponding author: WANG Yun-kai, E-mail: rickwyk@qq.com)

Abstract: Most of the traditional community detection algorithms cannot balance partitioning effect and complexity well. So, this paper presents a new evaluation standard of single community called group density. Based on the group density, a community detection algorithm based on agglomeration is proposed. The algorithm continues to integrate small communities, and makes the community structure of the network develop in the direction of maximizing average group density. Modularity is employed to detect the partitioning effect of the algorithm. Experimental results demonstrate that the new algorithm outperforms the traditional GN, Fast Newman, LPA algorithms in multiple data sets, which shows that the algorithm proposed has better partitioning effect and lower time complexity.

Key words: complex network; community structure; community detection; modularity; group density

复杂网络是一个综合性的现代学科,它能够抽象地表示事物间的复杂联系.网络由节点和节点之间的连边组成,大多数网络呈现为社团结构^[1].社团划分的目的就是根据网络的拓扑结构将网络节点划分到不同的组,也称为簇、集群或模块.社团划分是一个 NP 问题,通常对网络中每个节点的最终社团归属没有定义,因此,没有算法性能评估的明确标准.

社团发现算法主要分为非重叠社团发现算法和重叠社团发现算法,非重叠社团发现算法中有经典的 GN 算法^[2]、Newman 快速算法(FN)^[3]、Raghavan 等^[4]提出的基于标签传播的 LPA 算法,以及 Newman 等^[5]提出的一种基于元数据的社区发现算法.重叠社团发现算法有 Palla 等^[6]提出的派系过滤算法(clique percolation method, CPM)、Evans 等^[7]提出的基于边聚类的社团发现算法等. Radicchi 等^[8]定义强弱社团的概念,在 GN 算法的基础上提出了改进的局部算法,有效

地提高了计算速度. 然而, 这些算法大多难以在较低时间复杂度要求下保证划分效果.

近年来国内学者也对复杂网络展开了深入研究, 具体体现在以下几个方面: 在社团度量指标这方面, 李珍平等^[9]针对模块度存在的低分辨率问题提出了模块密度的概念. 在社团划分算法方面, 孙鹏岗等^[10]基于模糊传递规则提出了基于模糊聚类的社区发现算法, 韩忠明等^[11]提出了一种基于节点中心度的快速有效的复杂网络社团划分算法, 刘世超等^[12]提出了基于标签传播概率的重叠社区发现算法, 乔少杰等^[13]基于模块度聚类 and 图计算思想提出一种新的面向复杂网络大数据的重叠社区检测算法. 在新兴的动态网络方面, 王莉等^[14]进行了研究和探索, 牛新征等^[15]也提出了一种基于标签的多目标优化的动态网络社团发现算法.

上述算法中的大部分评价指标都基于整个网络, 而缺少评价单一社团性能的指标; 同时大部分算法都具有很高的时间复杂度. 本文提出了社团密合度的概念, 并且以此为评价指标提出了高效的复杂网络社团发现算法, 且准确率较高, 时间复杂度明显降低.

1 模型和方法

1.1 社团密合度

在观察分析网络中的社团结构时, 如果某社团内部的边很多, 而连接到其他社团的边很少, 可以认为这个社团更紧密. 因此, 当一个社团的内部边数与该社团的总边数(内部边数与连接到外部边数之和)之比更高时, 认为这个社团更紧密, 用式(1)表示目标社团内部边数与其总边数的比例:

$$D_1 = \frac{k_{1-in}}{k_{1-in} + k_{1-out}}. \quad (1)$$

式中 k_{1-in} 和 k_{1-out} 分别代表社团内部和外部边数目. 仅仅考虑边的影响并不能准确地表明社团的结构特性, 还需考虑社团内部节点数以及连接到社团总节点数(社团内部节点数与社团邻接节点数之和)对社团划分结果的影响, 因此用式(2)表示同时考虑边和节点影响的社团密合度:

$$D_{1-p} = \sqrt{\left(\frac{k_{1-in}}{k_{1-in} + k_{1-out}}\right)^\alpha \cdot \frac{k_{p-in}}{k_{p-in} + k_{p-out}}}. \quad (2)$$

式中 k_{p-in} 和 k_{p-out} 分别代表社团内部和外部节点数目, α 为正幂系数. 通常在网络中, 社团内部边的数量远大于节点数量, 因此社团内部连边的权

重应该高于节点的权重, 所以本文需要用正幂系数 α 加强边的权重, 增加其在式(2)中的影响. 函数值 D_{1-p} 此时为加强的社团边比例和社团节点比例乘积的几何平均值.

除了考虑社团内的节点与边, 同时也需要考虑一个社团在整个网络中所占比例对结果的影响, 用式(3)代表社团节点数占网络总节点数的影响值:

$$D_p = \left(\frac{k_{p-in}}{p}\right)^\beta. \quad (3)$$

式中: p 代表网络总节点数目; β 为指数幂系数, 其作用是防止在巨大网络中因社团过小而使函数值过小, 以平衡社团节点数与网络总节点数之比.

整理式(2)、式(3), 最终获得社团密合度函数:

$$D = \sqrt{\left(\frac{k_{1-in}}{k_{1-in} + k_{1-out}}\right)^\alpha \cdot \frac{k_{p-in}}{k_{p-in} + k_{p-out}} \cdot \left(\frac{k_{p-in}}{p}\right)^\beta}. \quad (4)$$

当社团中只有一个点时, $k_{1-in} = 0$, 计算得 D 值为 0; 而当社团中包含网络中所有的边和点时, $k_{1-out} = 0$, $k_{p-out} = 0$, $k_{p-in} = p$, 计算得 D 值为 1. 在社团增大的过程中, $k_{1-in} < k_{1-in} + k_{1-out}$, $k_{p-in} < k_{p-in} + k_{p-out}$, 因此得到式(2)的值域为 $(0, 1)$, 同时因为 $k_{p-in} < p$, 可得式(3)的值域为 $(0, 1)$. 得到社团密合度 D 的值域为 $[0, 1]$.

1.2 随机网络下的社团密合度

已知在随机网络下, 任意两个节点 i, j 之间相连的期望值是 $p_i \cdot p_j / (2m)$, 其中 p_i, p_j 代表节点 i, j 的度, m 代表网络中的总边数. 式(5)表示随机网络中社团 k 内部的连边数目:

$$k_{rl-in} = \sum_{ij} \frac{p_{ki} \cdot p_{kj}}{2m}. \quad (5)$$

式中 p_{ki}, p_{kj} 为社团 k 中节点 i, j 的度.

式(6)表示随机网络中社团 k 外部的连边数目:

$$k_{rl-out} = \sum_i \frac{p_{ki} \cdot p_{hj}}{2m}. \quad (6)$$

式中: p_{ki} 表示 k 社团节点 i 的度; p_{hj} 代表网络中不属于社团 k 的节点 j 的度.

假设此时社团外部节点数不变, 仍为 k_{p-out} , 则此时随机社团密合度为

$$D_r = \sqrt{\left(\frac{k_{rl-in}}{k_{rl-in} + k_{rl-out}}\right)^\alpha \cdot \frac{k_{p-in}}{k_{p-in} + k_{p-out}} \cdot \left(\frac{k_{p-in}}{p}\right)^\beta}, \quad (7)$$

随机社团密合度之差为

$$D - D_r = \left(\sqrt{\left(\frac{k_{l-in}}{k_{l-in} + k_{l-out}} \right)^\alpha} - \sqrt{\left(\frac{k_{rl-in}}{k_{rl-in} + k_{rl-out}} \right)^\alpha} \right) \delta.$$
$$(8)$$

其中 $\delta = \sqrt{\frac{k_{p-in}}{k_{p-in} + k_{p-out}}} \left(\frac{k_{p-in}}{p} \right)^\beta$, 且 $k_{p-in} \in (1, p]$,

所以 δ 的值域为 $(0, 1]$. 又 $\sqrt{\left(\frac{k_{l-in}}{k_{l-in} + k_{l-out}} \right)^\alpha} \in (0, 1)$, $\sqrt{\left(\frac{k_{rl-in}}{k_{rl-in} + k_{rl-out}} \right)^\alpha} \in (0, 1)$, 得到密合度之差的值域为 $(-1, 1)$. 经多次实验计算对比, 在 α 取 3, β 取 0.08 的情况下, 合理的社团结构的密合度差值较为明显, 合理差值一般在 $[0.15, 0.5]$ 之间.

1.3 社团密合度在实际网络中的应用

Zachary 空手道俱乐部成员关系网络^[16]是通过对一个美国大学空手道俱乐部进行观测而构建的社会网络, 其中节点表示俱乐部中的成员, 边表示成员之间的联系. 利用经典的 Fast Newman 算法^[3] (FN 算法) 对其进行社团划分. FN 算法的思想是通过迭代合并模块度增值最大的社团进行网络划分, 划分结束后会得到 3 个社团, 如图 1 所示.

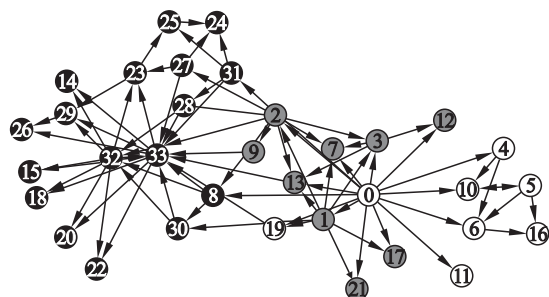


图 1 FN 算法划分结果

Fig. 1 Partitioning result of FN algorithm

记录每一轮划分结果的平均密合度 D_{av} 和模块度 Q , 得到如图 2 所示的折线图. 由图可知, 随着模块度 Q 值不断增大, 平均密合度也随之变大. 前期划分的社团个数较多, 平均密合度保持在较小的范围内. 随着迭代次数增加, 平均密合度也在不断变大, 在最后整个网络合并为一个社团时, 平均密合度变成 1.

计算划分结果的每个社团的密合度与随机密合度, 如表 1 所示.

得到以 1 号节点为中心的社团的密合度与随机密合度之差为 0.172; 以 33 号节点为中心的社团的密合度与随机密合度之差为 0.424; 以 0 号节点为中心的社团的密合度与随机密合度之差为

0.195. 实际划分结果的密合度之差均在合理区间内, 每个社团的密合度值均合理有效.

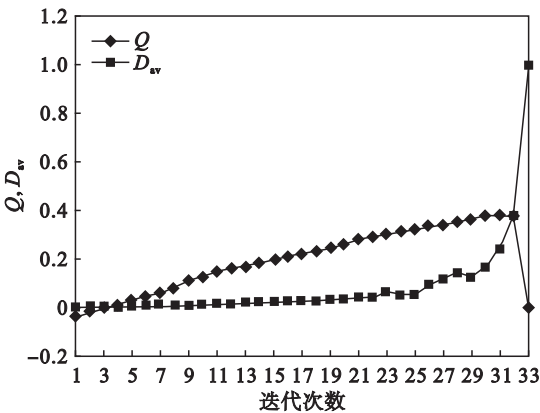


图 2 FN 算法每轮划分平均社团密合度与模块度
Fig. 2 D_{av} and Q of every round in experiment result of FN algorithm

表 1 实验结果
Table 1 Experiment result

社团 标号	社团节点	社团 密合度	随机 密合度
1	1, 2, 3, 7, 9, 12, 13, 17, 21	0.204	0.032
2	8, 14, 15, 18, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33	0.564	0.140
3	0, 4, 5, 6, 10, 11, 16, 19	0.213	0.019

1.4 社团密合度与模块度的比较

模块度是目前常用的一种衡量网络社区结构强度的方法, 该指标最早由 Mark Newman^[2] 提出. 模块度的定义为

$$Q = \sum_i (e_{ii} - a_i^2).$$
$$(9)$$

式中: e_{ii} 表示社团 i 内部的边占网络总边数的比值; a_i 表示网络中所有节点度之和占 2 倍网络边的比值, 即当网络完全随机划分时该社团的边占总边数的比例. 从模块度的定义来看, 它的基本思想是社团内部连接越紧密, 社团间连接越稀疏, 社团划分越合理; 这与社团密合度的评判标准一致.

社团密合度和模块度不同之处在于:

1) 从整个网络角度分析, 模块度是一个全局的网络社团划分评价指标. 社团密合度是网络中单个社团性能指标, 即仅能代表当前网络状态下, 所选社团的密集聚合程度. 因此在社团越少时, 每个社团内部的点越多、边越多、连到社团外部的边越少, 社团密合度就越大; 但这也并不是完全表明社团越大密合度就越大, 例如在空手道俱乐部划分结果里以 0 号为中心的社团的规模小于以 1 号为中心的社团, 但其密合度更大, 这表明以 0 号节

点为中心的社团更紧密,该社团结构更好。

2) 模块度仅考虑了网络中社团内边的影响,并将其与随机网络作比较,并没有考虑社团的节点与社团规模对社团结构的影响;社团密合度添加了社团节点比例这一因素,使计算结果变得更精确。

2 基于社团密合度的社团划分算法

2.1 算法描述

本文提出了一种基于社团密合度和凝聚思想的社团划分算法,简称为 GD (group density) 算法,算法主要思想是通过不断融合两个社团,以使网络社团结构向着平均密合度最大(即密合度增值最大)的方向发展,并计算每一轮的模块度,记录下模块度最大的划分结果;当整个网络变成一个社团时,会得到每一轮的划分结果,将模块度最大的划分结果作为网络的最终划分结果,同时可以看到网络社团结构的变化。

算法伪代码如下:

INPUT: 网络 $G = \{n \text{ 个点}, m \text{ 条边}\}$

OUTPUT: 各个阶段网络划分的结果与最终结果

PROCESS:

```
1: for  $i = 1, 2, \dots, n$  do
2:    $C_i = \{x_i\}$  //  $C_i$  表示社团集合,  $x_i$  表示网络节点
3: end for
4:  $q = n$ ;  $\text{Max}Q = -\infty$  //  $q$  为社团数量,  $\text{Max}Q$  为最大模块度
5: while  $q > 1$  do
6:    $\max = -\infty$  //  $\max$  记录最大平均密合度
7:   choose smallest  $C_{i^*}$ ;
8:   remove  $C_{i^*}$  from  $C$ ;
9:   for  $C_{j^*}$  in  $C$  do
10:    merge  $C_{i^*}$  and  $C_{j^*}$ ;
11:    calculate average GD;
12:    if averageGD >  $\max$  do
13:       $\max = \text{average GD}$ ;
14:       $C_j = C_{j^*}$ ;
15:    end if
16:  end for
17: merge  $C_i$  and  $C_j$ ;
18: calculate  $Q$ ;
19: if  $Q > \text{Max}Q$  do
```

```
20:   result = C // result 为最终划分结果
21: end if
22:  $q = q - 1$ 
23: end while
```

接下来用一个实例演示算法流程. 假设存在一个简单无向网络如图 3 所示。

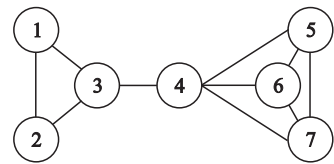


图 3 示例网络
Fig. 3 Sample network

将 GD 算法应用于该网络,首先将每个节点设置为一个社团,然后选择最小的社团中的 {1} 社团,计算将其加入与之邻边的 {2}, {3} 社团后的社团密合度变化,可以看出加入 {3} 社团后的密合度高于加入 {2} 社团的密合度,则 {1} 社团选择加入 {3} 社团;再选择最小社团中的 {2} 社团,计算将其加入与之邻边的 {1, 3} 社团或是 {4} 社团,选择密合度增值最大的 {1, 3} 社团并加入其中;接下来选择最小的社团 {4}, 计算将其加入与之邻边的 {1, 2, 3}, {5}, {6}, {7} 社团后密合度的变化,因为加入 {1, 2, 3} 社团后会使密合度变小,所以将其加入密合度都一样的 {7} 社团;接下来照此方法添加 {5}, {6} 社团,最后会把 {1, 2, 3} 添加到 {4, 5, 6, 7} 社团中,使整个网络变为一个社团. 算法处理流程如表 2 所示。

将模块度最高时得到的划分结果作为最终划分结果,此时网络被划分成两个社团,分别为 {1, 2, 3} 和 {4, 5, 6, 7}, 模块度为 0.355, 划分结果如图 4 所示,圆形和方形各表示一类社团。

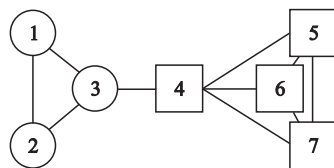


图 4 GD 算法划分结果
Fig. 4 Partitioning result of GD algorithm

2.2 算法复杂度分析

首先为网络中的 n 个节点分别创建社团,时间复杂度为 $O(n)$;然后每轮合并两个社团,循环 n 次直至所有初始社团合并为一个社团为止,循环的算法复杂度为 $O(n)$,此时的总时间复杂度为 $O(n + n)$. 接下来进行社团合并, GD 算法在进行密合度增量计算时并不需要将整个网络的所有

社团两两融合,它只需要找到规模最小社团中有最小密合度的社团,再寻找密合度增值最大的社团进行合并,所以选择最小社团与寻找合并社团

的时间复杂度为 $O(n+n)$. 总时间复杂度为 $O(n+n(n+n))=O(2n^2+n)$, 简化复杂度为 $O(n^2)$.

表 2 算法处理流程
Table 2 Process of GD algorithm

节点编号	1	2	3	4	5	6	7	选择
1		0.083	0.148					加入 {3}
2	0.539		0.539	0.034				加入 {3, 1}
4	0.318	0.318	0.318		0.04	0.04	0.04	加入 {7}
5				0.208		0.059	0.208	加入 {7, 4}
6				0.69	0.69		0.69	加入 {7, 4, 5}
3, 1, 2				1	1	1	1	加入 {7, 4, 5, 6}

注:表中粗体数字为社团密合度.

3 实验分析

使用基于社团密合度的社团发现算法对经典的空手道俱乐部网络 (club)^[16]、海豚网络 (dolphin)^[17]、美国大学生橄榄球联赛网络 (football)^[2]、美国政治书网络 (polbooks)^[18]、悲惨世界人物关系网络 (lesmis)^[19] 数据集进行社团划分,并将划分结果与使用传统经典算法的划分结果进行对比.

3.1 模块度对比实验

针对不同数据集,GD 算法与其他经典算法的最终划分结果的模块度如表 3 所示.

表 3 不同算法划分结果的 Q 值 Table 3 Q values of partitioning results by different algorithms					
算法	数据集				
	club	dolphin	football	lesmis	polbooks
GD	0.419	0.521	0.595	0.552	0.513
FN	0.381	0.495	0.577	0.501	0.502
GN	0.401	0.519	0.600	0.538	0.517
LPA	0.354	0.455	0.574	0.537	0.504

从表中可以看出,空手道俱乐部网络、海豚网络和悲惨世界人物关系网络中 GD 算法划分结果的模块度均高于其他算法 (最高达 9.3%);美国政治书网络和美国大学生橄榄球联赛网络中 GD 算法的模块度明显高于 FN 算法和 LPA 算法,但略逊于 GN 算法,但相比之下,两者差距并不大 (0.8%),且 GN 算法具有更高的算法复杂度.

同时在划分结果方面,针对于空手道俱乐部网络,GD 算法最终划分出了 4 个社团,比 FN 算法结构更精细,证明了社团密合度具有更高分辨

率;针对美国大学生橄榄球联赛网络和悲惨世界人物关系网络,GD 算法分别划分出了 9 个与 6 个社团,更符合真实情况.

3.2 时间对比实验

GD 算法的时间复杂度为 $O(n^2)$, 相比于 GN 算法的 $O(n^3)$ 和 FN 算法的 $O(n(m+n))$, GD 算法具有较低复杂度. 由于 GN 算法复杂度明显高于 GD 算法,这里仅计算 FN 和 GD 算法在上述网络中的运行时间. 对比结果如图 5 所示,可以看出,随着网络规模的增大, FN 算法比 GD 算法有越来越多的运行时间. 由此验证了 GD 算法的高效性.

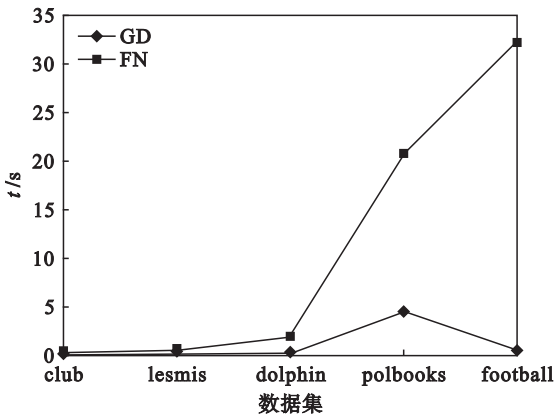


图 5 不同算法运行时间对比
Fig. 5 Comparison of running time with different algorithms

4 结 语

传统社团划分算法普遍存在划分效果和时间的复杂度矛盾的问题. 随着数据量的增加,传统算法无法处理大数据网络社团划分问题. 为了更好地解决这些问题,本文首先提出一种新的单社团评价函数——社团密合度,该函数可以有效地评价

社团结构并应用到算法中;在此基础上,提出了基于社团密合度的社团发现算法.实验证明,该算法能够发现更合理的社团结构,并且同时具有较低的时间复杂度.

参考文献：

[1] Newman M E J. The structure and function of complex networks[J]. *SIAM Review*,2003,45(2):167 – 256.

[2] 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.

[3] Newman M E J. Fast algorithm for detecting community structure in networks [J]. *Physical Review E*, 2004, 69(6):066133.

[4] 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.

[5] Newman M E J, Clauset A. Structure and inference in annotated networks[J/OL]. *Nature Communications*,2016: 11863[2017 – 10 – 08]. <https://www.nature.com/articles/ncomms11863>.

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

[7] Evans T S, Lambiotte R. Line graphs, link partitions, and overlapping communities[J]. *Physical Review E*, 2009, 80(1):016105.

[8] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. *Proceedings of the National Academy of Sciences of the United States of America*,2004,101(9):2658 – 2663.

[9] Li Z,Zhang S, Wang R S, et al. Quantitative function for community detection [J]. *Physical Review E*, 2008, 77(3):036109.

[10] Sun P G. Community detection by fuzzy clustering [J]. *Physica A: Statistical Mechanics & Its Applications*, 2015, 419:408 – 416.

[11] 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[J]. *中国科学:信息科学*, 2016,46(4): 431 – 444.

(Han Zhong-ming,Tan Xu-sheng,Chen Yan,et al. NCSS:an effective and efficient complex network community detection algorithm[J]. *Scientia Sinica Informationis*, 2016, 46(4): 431 – 444.)

[12] 刘世超,朱福喜,甘琳. 基于标签传播概率的重叠社区发现算法[J]. *计算机学报*,2016,39(4):717 – 729.

(Liu Shi-chao, Zhu Fu-xi, Gan Lin. A label-propagation-probability-based algorithm for overlapping community detection[J]. *Chinese Journal of Computers*,2016,39(4): 717 – 729.)

[13] 乔少杰,韩楠,张凯峰,等. 复杂网络大数据中重叠社区检测算法[J]. *软件学报*,2017,28(3):631 – 647.

(Qiao Shao-jie, Han Nan, Zhang Kai-feng, et al. Algorithm for detecting overlapping communities from complex network big data[J]. *Journal of Software*,2017,28(3):631 – 647.)

[14] 王莉,程学旗. 在线社会网络的动态社区发现及演化[J]. *计算机学报*,2015,38(2):219 – 237.

(Wang Li, Cheng Xue-qi. Dynamic community in online social networks[J]. *Chinese Journal of Computers*,2015,38(2):219 – 237.)

[15] 牛新征,司徒钰,余堃. 基于进化聚类的动态网络社团发现[J]. *软件学报*,2017,28(7):1773 – 1789.

(Niu Xin-zheng, Si Wei-yu, She Kun. Evolutionary community detection in dynamic networks [J]. *Journal of Software*,2017,28(7):1773 – 1789.)

[16] Zachary W W. An information flow model for conflict and fission in small groups [J]. *Journal of Anthropological Research*,1977,33(4):452 – 473.

[17] 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.

[18] Gong M, Fu B, Jiao L, et al. Memetic algorithm for community detection in networks[J]. *Physical Review E*, 2011,84(5):056101.

[19] Knuth D E. The Stanford GraphBase: a platform for combinatorial computing [M]. New York: ACM, 1993: 41 – 43.