

基于核心图增量聚类的复杂网络划分算法

张新猛¹ 蒋盛益¹

摘要 借鉴基于聚类的无监督入侵检测算法 (Clustering-based method for the unsupervised intrusion detection, CBUID) 聚类原理, 提出一种基于核心图增量聚类的社区划分算法 (Clustering-based method for community detection, CBCD). 本文提出一种社区摘要构建方法, 给出节点与社区相似度的计算公式. 首先, 对由少量高度数节点组成的核心网络采用现有算法进行核心社区划分, 然后, 采用增量方式依据节点与社区相似度, 将剩余节点划分到核心社区中. 算法复杂度主要依赖于网络规模、边的数量及划分的社区个数, 具有线性复杂度. 通过在几个典型真实网络数据集上测试, 所提算法能够有效地进行社区划分.

关键词 复杂网络, 社区摘要, 相似度, 社区发现

引用格式 张新猛, 蒋盛益. 基于核心图增量聚类的复杂网络划分算法. 自动化学报, 2013, 39(7): 1117–1125

DOI 10.3724/SP.J.1004.2013.01117

Complex Network Community Detection Based on Core Graph Incremental Clustering

ZHANG Xin-Meng¹ JIANG Sheng-Yi¹

Abstract This paper references the principle of clustering in clustering-based method for the unsupervised intrusion detection algorithm (CBUID), and proposes a clustering-based method for community detection (CBCD). We propose a method of community summary building, and give the formula of the similarity between node and community. First, it detects communities on the core network composed of a small amount of high-degree core nodes, then partitions the remaining nodes into core community according to the similarity between the node and community incrementally. Its running time mainly depends on the network size, the number of edges and the number of communities, and our algorithm has essentially a linear time complexity. Applications on several common real networks demonstrate that this method is very effective at community detection of networks.

Key words Complex networks, community summary, similarity, community detection

Citation Zhang Xin-Meng, Jiang Sheng-Yi. Complex network community detection based on core graph incremental clustering. *Acta Automatica Sinica*, 2013, 39(7): 1117–1125

自然界与人类社会中诸多系统可以采用复杂网络模型表示, 以节点表示系统中实体, 边表示实体之间的联系, 如人际关系网、论文合作网、电影明星合作网、电子邮件往来网、博客引用网、电话通讯网等. 近年来, 复杂网络已经成为多种学科研究的热点领域之一. 许多复杂网络具有社区结构, 社区可看作网络的一个子图, 社区内部节点连接紧密, 社区间节点连接稀疏. 研究表明社会网络^[1–2]、生物化学网络^[3]都具有明显的社区结构. 网络结构与网络

功能具有紧密的联系, 网络社区结构的研究能够揭示隐藏在复杂网络中的规律并有助于行为预测和控制, 如 Web 社区划分^[4]、蛋白质网络分析和功能预测等^[5].

网络社区结构的研究有较长的历史, 为了快速准确地进行网络社区划分, 人们提出了多种不同的复杂网络划分方法, 如基于连边密度^[6]、谱方法^[7–10]、GN^[7]算法、随机游走算法^[11]等. GN^[7]算法通过反复计算边介数、删除具有最大边介数的边, 自顶向下构建一棵层次聚类树. Newman 提出评价网络模块性的评价函数 (也称 Q 函数)^[7], 作为衡量网络社区划分好坏的一种度量, 并给出基于模块度的谱算法^[8–9]. Q 函数如下算式:

$$Q = \sum_i (e_{ii} - a_i^2) \quad (1)$$

其中, e_{ii} 表示在社区内部边占总边数的比例,

收稿日期 2012-04-16 录用日期 2012-08-12
Manuscript received April 16, 2012; accepted August 12, 2012
国家自然科学基金 (61070061), 教育部人文社会科学研究青年基金项目 (11YJCZH086, 12YJCZH281, 13YJCZH258) 资助
Supported by National Natural Science Foundation of China (61070061), the Youth Project of Humanity and Social Science for Ministry of Education (11YJCZH086, 12YJCZH281, 13YJCZH258)
本文责任编辑 吕金虎
Recommended by Associate Editor LV Jin-Hu
1. 广东外语外贸大学思科信息学院 广州 510006
1. Cisco School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006

$a_i = \sum_j e_{ij}$ 表示一端在社区 i 内的边占总边数的比例. 一个好的网络社区划分, 能够使 Q 值较大, 通常 Q 值处于 $0.3 \sim 0.7$ 之间. FN^[12] 算法采用局部搜索策略, 执行使 ΔQ 值最大的合并策略, 自底向上进行层次聚类. Guimera 等采用模拟退火算法对 Q 值进行优化, 并应用于新陈代谢网络分析上^[13], 但是算法收敛慢, 运行效率低, 且参数设置对聚类结果影响较大^[14]. 另外遗传算法^[15]、基于信息瓶颈方法^[16]也被应用于寻找最优 Q 值. 模块度思想的本质是使更多的边处于社区内部, 受模块度的启发, 文献 [17] 提出以节点与节点之间的共同邻居作为网络划分质量的衡量标准, 同一社区内部的节点之间拥有更多的共同邻居, 定义了共邻矩阵和增益函数, 通过优化增益函数进行社区划分. 社会学研究中, 将具有完全相同邻居的节点称为结构等价, 结构等价的节点很可能处于同一社区, 现实意义是若两个节点与同一个节点相邻, 那么这两个节点相邻的可能性要大于网络中随机两个节点, 两个节点的共同邻居越多, 那么这两个节点相邻的可能性越大. 比如, 在社会关系网络中, 你所认识的两个朋友相互认识的可能性要远大于随机两个人相互认识的可能性, 两个人的共同朋友越多, 这两个人相互认识的可能性越大. 因此, 可以通过比较邻居来判断两个人的相似度. 基于局部相似性的社区划分算法^[18] 以星型邻域子图代替邻居点集来计算两节点的局部相似性.

现实社会中, 有许多大规模网络, 包含海量的节点和边. 比如, 微博网络社区、电话通讯网络等, 同时, 网络是动态增长的, 不断有新的节点及边出现. 目前, 多数社区划分算法都具有较高的时间复杂度, 只适合中小规模网络的社区划分. 以 n 表示网络节点总数, m 表示边总数, GN 算法时间复杂性高达 $O(m^2n)$ 或 $O(n^3)$ (对于稀疏网络), 只适合处理小规模网络. FN (Fast newman) 算法的时间复杂性是 $O(mn)$, Radicchi 等提出连接聚类系数 (Link clustering coefficient) 取代 GN 算法的边介数^[19], 时间复杂性为 $O(m^4/n^2)$, 对于稀疏网络, 它们的算法时间复杂度可达 $O(n^2)$. 文献 [20] 给出一种层次聚类算法, 算法复杂度为 $O(md \log n)$, 若为稀疏网络, 具有线性时间复杂度为 $O(n \log^2 n)$. 若要处理大规模网络数据集, 算法必须具有线性复杂度, 同时能够处理不断新增的节点. 本文提出的算法采用少量高度节点快速构建初始社区划分, 然后, 依据节点与社区的相似度对其他节点进行增量式社区划分, 算法具有线性复杂度.

基于聚类的无监督入侵检测算法 (Clustering-based method for the unsupervised intrusion detection, CBUID)^[21] 采用摘要信息表示一个簇, 判断对象与簇间距离, 将对象划分到相似度最大的簇

中, 具有线性时间复杂度. 基于 CBUID 算法聚类思想, 本文提出基于核心图增量聚类的社区划分算法 (Clustering-based method for community detection, CBCD), 以节点与社区内节点相邻个数作为社区摘要, 给出社区与节点相似度计算方法. CBCD 算法采取三步骤对网络进行社区划分: 首先, 选取度数大于 d_r 的节点构成核心图, 采用现有算法 (本文采取 GN 算法) 对核心图进行初始社区划分, 根据社区中节点的邻居构建社区摘要; 然后, 遍历剩余节点, 计算剩余节点与各个社区的相似度, 将节点加入到相似度最大的社区中; 最后, 重新迭代计算各节点与各社区的相似度, 根据相似度计算结果进行调整所属社区, 并更新社区摘要, 直至没有节点需要进行重新划分. 除初始社区划分外, 算法时间开销主要应用于节点与各社区的相似度计算, 具有较低的时间复杂度. 为验证算法的有效性, 在多个真实的网络数据集上进行了测试, 实验结果表明, 算法能够有效地进行社区划分.

本文第 1 节讨论了社区摘要的定义及节点与社区相似度的计算方法; 第 2 节给出基于相似计算社区划分算法的描述、核心节点度阈值的选取及时间复杂度分析; 第 3 节在多个常用的网络数据集上进行了测试并与现有经典算法做了对比; 第 4 节对算法做了总结, 提出一些待探讨的问题.

1 社区摘要及相似度计算

本文研究的对象为无向无权网络, 以 $G(V, E)$ 表示一个无向无权网络图, 其中, $V = \{v_i | i = 1, \dots, n\}$ 为节点的集合, $E = \{e_{ij} | v_i, v_j \in V\}$ 为边的集合, $|V| = n$, $|E| = m$, 即 n 为节点的总数量, m 为边的总数量.

节点 v_i 的邻集用下式表示:

$$N(v_i) = \{v_j | v_j \in V, e_{ij} \in E\} \quad (2)$$

e_{ij} 为节点 v_i 与 v_j 间的边, $N(v_i)$ 表示所有与节点 v_i 相邻接的节点集合.

图 1 为一个具有社区结构特征的网络示意图, 图 1 中节点 v_4 的邻集为 $N(v_4) = \{v_1, v_2, v_3, v_7\}$.

若将 G 划分为 k 个社区 $C_1, C_2, \dots, C_k$, 设节点 v_i 是社区 C_j 内节点的一个邻居, 社区 C_j 内与节点 v_i 相邻的节点个数作为社区 C_j 对节点 v_i 的支持值, 表示如下式:

$$\text{sup}(C_j, v_i) = |N(v_i) \cap C_j|, \quad v_i \in V \quad (3)$$

$|N(v_i) \cap C_j|$ 表示节点 v_i 的邻集与社区 C_j 交集的节点个数, 含义是社区 C_j 内与节点 v_i 相连的节点个数.

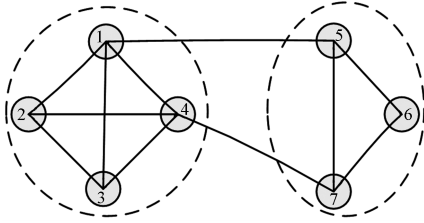


图 1 一个具有社区结构的网络示意图

Fig. 1 A schematic representation of a network with community structure

考虑如图 1 中的网络结构, 可以划分为 2 个社区, 节点 v_1, v_2, v_3, v_4 构成一个社区 C_1 , 另外三个节点 v_5, v_6, v_7 构成一个社区 C_2 , 节点 v_4 与 C_1 中 3 个节点相邻, 与 C_2 中 1 个节点相邻, 所以节点 v_4 在两个社区的支持值分别为

$$\sup(C_1, v_4) = 3$$

$$\sup(C_2, v_4) = 1$$

以社区的邻居及社区内与该邻居节点相连的节点个数作为社区摘要, 以社区摘要表示一个社区, 社区摘要表示如下:

$$CS_j = \{\langle v_i, \sup(C_j, v_i) \rangle | N(v_i) \cap C_j \neq \emptyset\} \quad (4)$$

其中, v_i 是网络 G 中的一个节点且为社区 C_j 中节点的邻居.

基于图 1 中的网络结构, C_1 和 C_2 社区摘要为

$$CS_1 = \{\langle v_1, 3 \rangle, \langle v_2, 2 \rangle, \langle v_3, 3 \rangle, \langle v_4, 3 \rangle, \langle v_5, 1 \rangle, \langle v_7, 1 \rangle\}$$

$$CS_2 = \{\langle v_1, 1 \rangle, \langle v_4, 1 \rangle, \langle v_5, 2 \rangle, \langle v_6, 2 \rangle, \langle v_7, 2 \rangle\}$$

若一个节点的邻居与社区内相邻接的节点数越多, 该节点就越可能属于该社区, 因此, 可以通过节点 v_i 的邻集 $N(v_i)$ 中节点在各社区摘要中的支持值计算节点 v_i 与各社区之间的相似度, 节点 v_i 与社区 C_k 的相似值计算用式 (5) 表示:

$$\text{sim}(v_i, C_k) = \frac{\sum_{v_j \in N(v_i)} \sqrt{\frac{\sup(C_k, v_j)}{|C_k|}}}{d(v_i)} \quad (5)$$

其中, $|C_k|$ 为社区 C_k 中节点的个数, $d(v_i)$ 为节点 v_i 的度, 节点 v_j 为节点 v_i 邻集 $N(v_i)$ 中的节点, $\sup(C_k, v_j)$ 的值越大, 表示社区 C_k 中与 v_j 邻接的节点越多, 节点 v_i 与社区 C_k 越相似. 若 v_j 与社区内所有节点都相邻, 即节点 v_i 与社区中所有节点都有共同的邻居 v_j , $\sup(C_k, v_j)$ 的达到最大值为 $|C_k|$. 若节点 v_i 的所有邻居与社区 C_k 中所有节点都相邻且社区摘要 CS_k 中所有邻居节点支持值均为 $|C_k|$ (这种情况下, 该社区为一个完全图), 那么节点 v_i

与社区 C_k 的相似值 $\text{sim}(v_i, C_k)$ 为 1; 若 v_i 的所有邻居都不和社区中节点相邻, $\text{sim}(v_i, C_k)$ 为 0, 所以 $\text{sim}(v_i, C_k)$ 的值介于 $[0, 1]$. 对比例值的开方是为了消除单个邻居对整个相似度的过度影响, 通过在真实网络数据集上的实际测试, 对比例值的开方能使社区模块度值有所提高.

在图 1 中新加入一个节点 v_8 , 分别与节点 v_3, v_5, v_7 相连, 如图 2 所示.

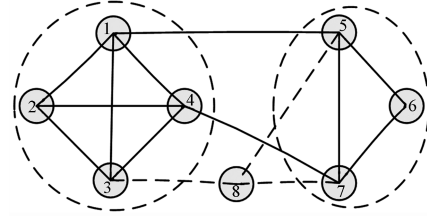
图 2 在图 1 所示网络中新加入一个节点 v_8 Fig. 2 Add a new node v_8 in the network shown in Fig. 1

图 2 中节点 v_8 与现有两个社区 C_1, C_2 的相似度分别为

$$\begin{aligned} \text{sim}(v_8, C_1) &= \frac{\sqrt{\frac{\sup(C_1, v_3)}{|C_1|}}}{d(v_8)} = \frac{\sqrt{\frac{3}{4}}}{3} \\ \text{sim}(v_8, C_2) &= \frac{\sqrt{\frac{\sup(C_2, v_7)}{|C_2|}} + \sqrt{\frac{\sup(C_2, v_5)}{|C_2|}}}{d(v_8)} = \frac{2\sqrt{\frac{2}{3}}}{3} \end{aligned}$$

显然节点 v_8 与社区 C_2 的相似度较大, 应划入到社区 C_2 中.

2 基于核心图的增量社区划分算法

2.1 CBCD 算法描述

CBCD 算法共分三大步骤: 1) 选取网络中度数高于阈值 d_r 的少量节点, 采用现有社区划分算法 (本文采用 GN 算法) 对核心节点初始社区划分, 遍历各社区中核心节点的邻居, 统计邻居与社区内节点相邻接的个数, 构建社区摘要; 2) 遍历剩余节点, 依次比较剩余节点与各社区的相似度, 将节点划分到相似度最大的社区中, 同时更新社区摘要; 3) 重新计算所有节点与各社区的相似度, 将节点重新调整到相似度最大的社区, 同时更新社区摘要, 直至没有节点需要调整, 算法详细描述如下:

输入: 一个网络 $G(V, E)$, 度阈值 d_r , 社区个数 k .

输出: 网络划分 $C_1, C_2, \dots, C_k$.

- 1) 设置一个度阈值 d_r , 选取度高于 d_r 的核心节点, 这部分节点构成一个核心子图 G_c .
- 2) 采用 GN 算法对 G_c 进行初始社区划分, 划分为 k 个初始社区.
- 3) 遍历初始社区内核心节点的在整个网络 G 中的邻居节点, 构建各社区摘要 $CS_1, CS_2, \dots, CS_k$.
- 4) 遍历剩余节点, 计算节点与各社区的相似度, 将该节点划入到相似度最大的社区中, 同时更新该社区摘要.
- 5) 重新遍历所有节点, 计算节点到各个社区的相似值, 若具有相似值最大的社区不是节点原来所在社区, 将该节点移至到相似值最大的社区, 同时更新两个社区摘要.
- 6) 如果有节点更换了社区, 转至 5) 进行再次迭代, 否则结束算法.

通常核心节点重新划分到其他社区的可能性较小, 对于大规模网络, 步骤 5) 可以修改为重新遍历非核心节点, 可以进一步降低时间开销.

2.2 核心图构建

核心图是由网络中度数大于度阈值 d_r 的节点所构成的子图. 核心图应具有两个特征: 首先, 核心图应具有社区结构, 即社区内的核心节点连接紧密, 而社区间的核心节点连接稀疏, 否则无法进行初始核心社区划分; 其次, 核心图每个社区的核心节点应有足够的邻居节点构建社区摘要, 否则未划分节点的邻居节点可能在所有社区摘要中都未出现, 造成无法计算该节点与社区的相似度. 因而 d_r 值设置至关重要, 若 d_r 值过大, 核心节点数量过少, 核心社区摘要信息过少, 可能造成剩余节点错误的划分, 若 d_r 值过小, 核心节点数量过多, 初始社区的划分仍需要大量的时间开销. 如表 1 为 Polblogs 网络在不同度阈值情况下执行情况的比较, 其中, d_r 为度阈值, n_c 为核心节点数, d_c 为核心节点度的和, t_c 为核心图社区划分执行时间, t_o 为剩余节点划分执行时间, Q 为模块度.

Polblogs 网络共有 1 224 个非孤立节点, 从表 1 可以看出随着度阈值减少, 核心节点增多, 由于核心图的社区划分采用现有算法 (本文采用 GN), 核心图的社区划分时间迅速增加, 而剩余节点划分时间开销相当, 显然度阈值越大, 核心节点越少, 算法时间总开销越小. 但是从模块度来看, 阈值过小的情况下, 社区划分可能达不到理想的结果, 比如表 1 中度阈值在 220 的情况下, 模块度为 0.4291, 略低于本算法得到的最优结果 0.4304.

通过实际测试发现, 多数复杂网络可以从两方面考虑度阈值 d_r 的选取. 首先, 若一个节点与本

社区中约 $1/4 \sim 1/3$ 的节点相连, 可以作为核心节点. 设定整个网络节点个数为 n , 需要划分社区个数为 k , 那么核心节点度的阈值 d_r 约为 $n/k/4$. 其次, 所选核心节点的度之和大约是整个网络节点总数的 $2 \sim 3$ 倍, 能够保证核心节点有足够的邻居节点构建初始社区摘要. 表 2 中是实验中用到的其中 4 个网络数据集所选取的核心社区特征, 其中, n 为节点总数, m 为边总数, d_r 为度阈值, n_c 为核心节点数, d_c 为核心节点度的和, p_c 为核心节点度的和与节点总度数和的比, 可以看出核心节点选取基本满足以上两个方面.

表 1 不同度阈值下 Polblogs 网络社区划分执行情况
Table 1 The implementation of Ploblogs network community detection under different thresholds

d_r	n_c	d_c	t_c (ms)	t_o (ms)	Q
300	4	1 566	15	465	0.4289
250	6	2 105	16	467	0.4292
240	7	2 350	16	437	0.4302
220	11	3 260	78	469	0.4291
190	16	4 319	922	438	0.4304
186	17	4 508	1 281	438	0.4304
180	18	4 690	2 140	453	0.4304
170	24	5 734	8 828	406	0.4304
160	28	6 389	19 110	453	0.4304

表 2 四个网络的核心图特征
Table 2 The characteristics of four subgraph of network

Network	n	m	k	d_r	n_c	d_c	p_c
Zachary	34	78	2	5	10	91	58.33 %
Dolphin	62	159	2	8	13	122	38.36 %
Polbooks	105	441	2	13	22	326	36.96 %
Polblogs	1 490	19 090	2	186	17	4 508	11.81 %

2.3 算法复杂度

本文算法时间开销由初始社区划分、初始社区摘要构建、剩余节点划分及迭代 4 个步骤构成.

首先, 对度高于阈值 d_r 的节点所构成的核心图进行初始社区划分. 绝大多数复杂网络具有无标度特性, 复杂网络的度分布多数服从幂律分布^[22], 网络中的高度数节点所占比例很少, 但是高度数节点的度数和却占据了很大的比例, 如表 1 政治博客引用网络中前约 1% 的节点度之和占到总节点度的和约 12%, Eclipse 3.4.2 结构网络中, 最高度前 5% 的节点度数和占到总度数的 17%^[23], 在 AS 图、INET-3.0 模型、修正 BA 无标度网络模型、BA 无标度网

络模型中, 最高度数前 5% 的节点度数和占总度数的比例分别高达约 0.44%、49%、30%、23%^[24].

初次构建社区摘要需要遍历核心节点的所有邻居节点 (基于全网络 G), 时间开销依赖于核心节点的邻居总数量, 假如核心节点的度之和为 d_c , 那么遍历次数为 d_c , 按核心节点选取原则, 通常核心节点度的和约为总节点数的 2~3 倍, 因此, 该步骤时间复杂度约为 $O(n)$.

计算一个节点 v_i 与 k 个社区的相似性需要 k 次遍历该节点的邻居, 时间复杂度为 $d_i k$, d_i 为节点 v_i 的度, k 为社区个数, 假设需要迭代 t 次, 那么时间代价为 $t(\sum_{v_i \in G} d_i k)$, 提取公共因子 k , 可表示为 $tk(\sum_{v_i \in G} d_i)$, $\sum_{v_i \in G} d_i$ 为节点度的和, 节点度的和为边数 m 的两倍, 时间代价也可表示为 $2mtk$, 所以该步骤运算时间复杂度约为 $O(tkm)$. 由于在迭代过程中, 核心节点通常不会重新划分到新社区, 社区摘要变化不大, 因此, 算法收敛较快, 在实际应用中, 可以不考虑核心节点到各社区的重新计算, 时间开销进一步降低. 在本文真实网络数据集的实际测试中, 迭代的次数通常不超过 5 次, 所以算法时间复杂度主要依赖于网络边数和所划分社区个数.

因此, 本文算法除初始化社区外, 时间复杂度为 $O(n + tkm)$, 现实中的复杂网络是稀疏的, 对于稀疏网络, 时间复杂度近似为 $O(n)$.

3 实验分析

为了验证算法的有效性, 采用几个经典的网络数据集进行测试, 为了便于测试对比, 在部分数据集中为每个节点添加了所属社区编号, 以下所有图中用不同形状表示节点所属的真实社区, 以节点位置表示划分结果, 节点标号后括号中数字表示实际所属社区序号.

算法均由 Java 编程实现, 操作系统是 Windows XP (SP3), 计算机配置是: 2 GB 的内存, Intel Core 2 Duo T6400 处理器, 2.00 GHz.

本文测试共用到 10 个网络数据集, 包括空手道俱乐部网络 (Karate)^[25]、海豚网络 (Dolphins)^[26]、美国大学生 2000 赛季美式足球联赛网络 (Football)^[14]、2004 年美国大选政治博客网络 (Polblogs)^[27]、科学家论文合作网络 (Netscience)^[28]、政治书籍网络 (Polbooks, <http://www.orgnet.com/>)、电子邮件网络 (Email)^[29]、爵士音乐家合作网络 (Jazz)^[30]、拟南芥 (*C. Elegans*) 新陈代谢网络 (Cel)^[31]、美国航空网络 (US.air, <http://vlado.fmf.uni-lj.si/pub/networks/data/>).

3.1 Zachary 空手道俱乐部网络

第一个数据集采用 Zachary 空手道俱乐部网络, 它是美国一所大学空手道俱乐部成员间关系的网络. Zachary Wayne 通过两年的观察, 获取了 34 个成员的社会交往关系, 从而得到了一个网络. 在这个网络中, 每个成员代表一个节点, 如果两个成员在俱乐部内或在俱乐部外有社会交往关系, 这两个成员对应的顶点之间有一条边相连. 由于俱乐部领导人的矛盾, 分裂为以校长和主管为中心的两个小俱乐部, 如图 3, 方点和圆点分别表示实际分裂后的各个俱乐部成员.

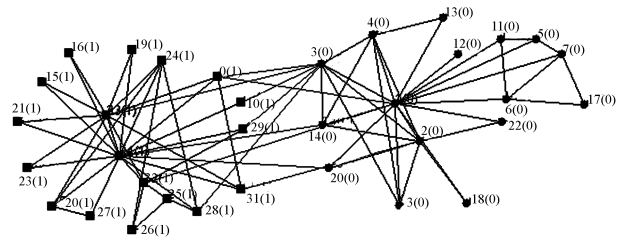


图 3 Zachary 空手道俱乐部网络
Fig. 3 The friendship network from Zachary's karate club

设置核心节点度阈值 d_r 为 5, 共有 10 个节点构成一个子网. 如图 4, 首先采用 GN 算法将核心社区划分为 2 个社区, 分别为 $\{v_1, v_2, v_3, v_4, v_{14}\}$, $\{v_9, v_{24}, v_{32}, v_{33}, v_{34}\}$.

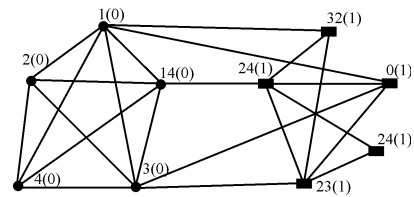


图 4 Zachary 空手道俱乐部网络核心节点组成的子图
Fig. 4 The core subgraph of Zachary's karate club network

最终社区划分的结果如图 3 所示, 所有方点在左边区域, 圆点在右边区域, 显然算法最终划分结果与实际完全相符, 文献 [9, 17] 也取得同样结果. 而若采用 GN 算法^[1] 和基于信息瓶颈算法^[16] 对整个网络进行社区划分, 节点 3 被错误地划分, 节点 3 与实际划分的每个部分联系紧密程度差不多, 本文算法考虑了邻居与社区的关系, 避免了这一节点的错误划分. 在其余节点社区归并过程中, 第一次划分的结果与最终结果一致, 因此仅进行了一次迭代.

3.2 海豚社会关系的网络

第二个数据集是关于生活在新西兰神奇湾 62

只海豚社会关系的网络,网络中节点为海豚,节点之间的边为海豚之间的社会关系,由于一只关键海豚的离开而自动划分为两个较小的族群.如图5,方点和圆点分别表示实际分裂后的族群成员.

设置的核心节点度阈值 d_r 为8,共有核心节点数量为13,所构成的核心社区如图6,由于核心社区是不连通的,采用GN算法只需执行几步即可判断社区已经分离,所以初步社区划分极快.

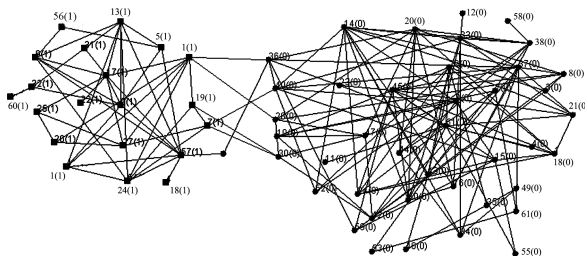


图5 海豚社会网络

Fig. 5 Dolphins network

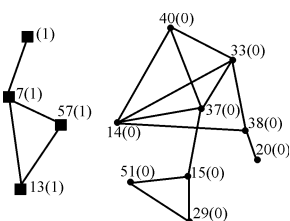


图6 海豚社会网络核心节点构成的子图

Fig. 6 The core subgraph of Dolphins network

最后划分的社区结果如图5所示,只有节点39被错误地划分,两个社区间的边仅为6条,而在文献[9, 17]中被错误划分的节点均为3个,分别为28、30、39,社区间的边为8条.

3.3 政治书籍网络

3.3.1 两社区划分

第三个数据集为政治书籍网络,如图7,方形和三角形节点分别表示具有保守主义和自由主义倾向的书籍,圆形节点表示中立书籍,总节点数105个,网络中的边表示两本书同时被一个读者所购买.

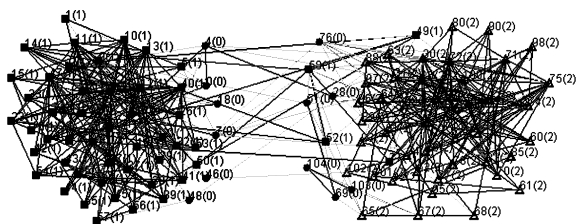


图7 政治书籍网络

Fig. 7 Polbooks network

设置核心节点度阈值12,共选取核心节点22个.所构成的子网络如图8,子图是不连通的,所以采用GN算法能够快速进行社区划分.

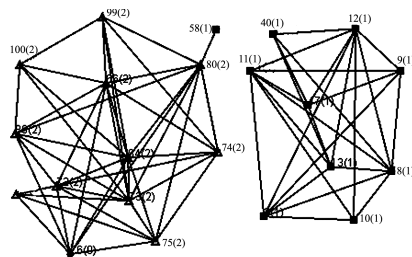


图8 政治书籍网络核心节点构成的子图

Fig. 8 The core subgraph of Polbooks network

将社区划分为2个社区,划分结果如图7,以0、1、2分别表示中立、保守主义和自由主义倾向书籍社区编号,两个社区划分后所包含的各类书籍个数为 $\{ \langle 1, 45 \rangle, \langle 0, 7 \rangle \}$, $\{ \langle 2, 43 \rangle, \langle 1, 4 \rangle, \langle 0, 6 \rangle \}$, $\{ \langle 1, 45 \rangle \}$ 表示实际属于社区1的节点个数为45,不考虑中立书籍,只有社区1中有4个节点被错误的划分,进行一次迭代后算法结束.

在文献[9, 19]中两社区中间共有24条边,本算法划分后社区间有19条边.本算法划分后社区模块度为0.457,而采用GN算法,模块度为0.434.

3.3.2 多社区划分

政治书籍网络实际包含三类书籍,对核心节点采用GN算法划分为3个社区,然后,根据相似计算进行剩余节点社区划分.划分结果如图9所示,圆点、方点与三角形分别表示节点所属的真实社区,所划分的三个社区包含各类书籍的个数为 $\{ \langle 1, 45 \rangle, \langle 0, 6 \rangle \}$, $\{ \langle 2, 5 \rangle, \langle 1, 2 \rangle, \langle 0, 5 \rangle \}$, $\{ \langle 2, 38 \rangle, \langle 1, 2 \rangle, \langle 0, 2 \rangle \}$,三个社区内部边与外部边的个数分别为(205, 19), (26, 21), (179, 22),模块度为0.501,文献[17]算法所得结果模块度为0.480,文献[9]算法所得结果模块度为0.467.

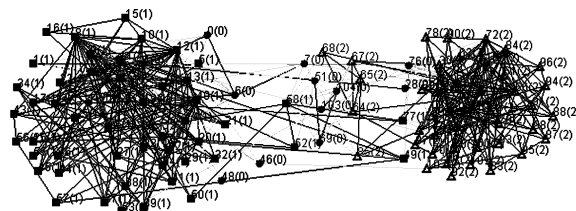


图9 政治书籍网络三社区划分

Fig. 9 The Polbooks network partitioned into three community

3.4 政治博客网络

第4个数据集为政治博客引用网络,网络中节

点表示博客, 分保守派和自由派博客两类节点, 边表示博客间的引用关系, 共 1 490 个节点, 删除 216 个孤立点后, 共有 1 274 个相连的节点. 本实验不考虑博客间的链接方向, 看着一个无向网络, 将网络划分为保守派和自由派两个社区. 设置核心度阈值为 186, 共有核心节点 17 个, 构成的子图如图 10, 核心子图划分为两个社区. 对 1 274 个非孤立点进行划分, 两个社区包含的各类节点个数分别为 $\{(1, 638), \langle 0, 77 \rangle\}$, $\{(1, 14), \langle 0, 545 \rangle\}$, 共有 91 个节点被错误划分, 划分正确率为 92.86 %, 两个社区内部边分别为 9 419 和 8 372, 社区间边为 1 299, 模块度为 0.4304.

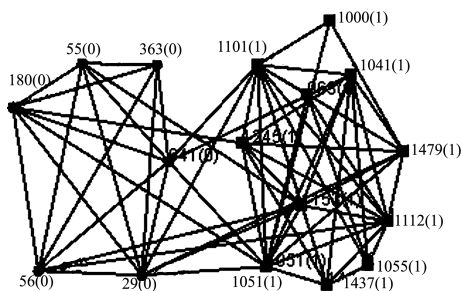


图 10 政治博客引用网络核心节点子图

Fig. 10 The core subgraph of Polblogs network

3.5 多个算法比较

我们对 CBCD 算法在更多数据集上进行了测试, 并与文献 [18] 中不同算法运行的结果进行了对比, 实验结果对比见表 3, 其中, Fastgreedy 指 CNM 算法^[20], Betweenness 指 GN 算法^[7], Label 表示标签传播算法 (Label propagation)^[32], Eigenvector 表示模块度矩阵谱聚类算法^[9], Spinglass 表示自旋玻璃模型算法^[13], Similarity 为基于局部相似度社

区结构发现算法^[18].

表 3 中除 CBCD 算法外, 其他算法社区发现的结果数据均来自于文献 [18]. 表 3 第 1 列展示了网络的基本统计数据, Nodes 对应网络的节点, Edges 对应网络的连接数. 对每种方法, 报告了其社区发现的模块度、社区数目. 例如 Karate 网络用 Similarity 算法得出的模块度为 0.41, 社区数为 4. CBCD 算法列中报告了模块度、社区数目、核心度阈值、核心节点数. 比如 Email 网络用 CBCD 算法得出的结果描述 0.53/8 (70, 32), 表示模块度为 0.53, 社区数目为 8, 核心度阈值为 70, 核心节点数为 32. 从实验结果对比来看, CBCD 算法在 Polblogs 和 Email 网络上社区检测效果最为理想, 选用的核心节点比例很小, 而取得的效果却很好, 这两个网络都具有比较显著的无标度特性, CBCD 算法比较适用于无标度特性明显的网络社区发现. 无标度网络网络节点的连接度没有明显的特征长度, 度呈幂律分布, 不少现实中的网络结构有无尺度的特性, 如 www、新陈代谢网络、电影演员合作网络、美国航班飞机网络等. 但是 CBCD 算法在度分布比较均匀的网络中可能出现选取核心节点困难的问题. 比如在 Football 网络中, 度大于 11 的节点仅有 12 个, 而大于 10 的节点却有 78 个, 若采用 12 个核心节点组成的核心网络进行初始社区划分, 在划分 9 个社区的情况下, 所取得模块度为 0.52. 其实在实际应用中, 若已经知道部分节点属于某些社区, 可以采用已知社区进行构建社区摘要, 然后, 对未知节点进行社区划分. 我们随机取 12 个社区中任意 3 个节点作为每个初始社区成员, 然后, 对其他节点进行社区划分, 取得的模块度可达 0.60. 最后一行为各种算法在所有数据集上测试的模块度的平均值, CBCD 算法总体效果比较突出.

表 3 各算法社区发现结果

Table 3 Community detected by different algorithms

Network	Nodes/Edges	Fastgreedy	Betweenness	Label	Eigenvector	Spinglass	Similarity	CBCD
Karate	34/78	0.38/3	0.40/5	0.37/3	0.38/5	0.42/4	0.41/4	0.37/2(5,10)
Dolphins	62/159	0.50/4	0.52/5	0.51/4	0.49/6	0.53/5	0.52/4	0.52/4(6,38)
Jazz	198/2 742	0.44/4	0.41/39	0.28/2	0.35/8	0.44/5	0.40/4	0.43/3(120,7)
Football	115/613	0.55/6	0.60/10	0.58/12	0.45/13	0.60/10	0.60/10	0.58/9(11,78)
US.air	332/2 126	0.19/4	0.14/124	0.00/3	0.27/16	0.03/24	0.30/10	0.28/6(70,9)
Polblogs	1 222/16 714	0.43/10	0.42/205	0.43/3	0.27/62	0.43/16	0.38/2	0.43/2(186,17)
Cel	453/2 025	0.39/9	0.40/38	0.33/8	0.32/28	0.44/10	0.37/14	0.36/9(40,31)
Email	1 133/5 451	0.49/16	0.53/61	0.01/4	0.42/45	0.58/12	0.45/16	0.53/8(70,32)
Netscience	379/914	0.84/19	0.84/18	0.81/38	0.78/33	0.84/21	0.82/17	0.81/12(6,111)
均值		0.47	0.47	0.37	0.41	0.48	0.47	0.48

4 结论

本文将 CBUID 算法聚类的思想应用于社区结构划分中, 提出一种基于核心图增量聚类的社区发现算法 CBCD, 给出社区摘要的构建方法和节点与社区相似度计算公式. 算法首先对少量高度节点构成的核心网络进行社区划分, 然后, 将剩余节点加入相似度最大的社区中, 时间开销主要依赖于网络中边的数量, 具有线性时间复杂度.

现实中的复杂网络往往是大规模、稀疏、动态的, 而高度数核心节点却是少量的, 可以将 CBCD 算法平凡地推广到这类复杂网络的社区发现问题中, 不仅能快速进行社区划分, 而且能够采用增量方法判断不断新增的节点应划分的社区.

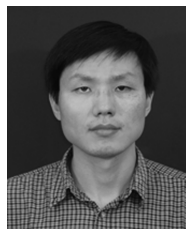
CBCD 算法尚存在一些待深入探讨的问题. 首先是核心节点选取, 虽然本文提出的核心节点选取的两个原则在实验所采用的网络上适用, 但在更多不同类型网络中, 核心节点选取方法也许存在较大差异, 尤其是高度数节点较为集中在某些社区的或节点度比较均匀的情况下, 很难选出核心节点; 其次, 部分具有社区结构特征的网络中, 核心节点所构成的核心图是否也呈现相似社区结构也是一个有待探讨的问题; 另外, 对于动态网络, 新增的节点若是核心节点或核心节点消亡, 可能会引起社区的产生、合并、分裂与消亡, 本算法尚未能对社区进行动态重新划分.

CBCD 算法目前是针对无权无向网络, 下一步将基于加权有向网络探讨如何构建社区共邻摘要及节点与社区的相似度计算方法, 在更多网络数据集上进一步探讨不同类型网络中核心节点的选取. 同时很多真实复杂网络处于不断动态演化过程中, 比如电信通讯网络和微博网络等, 如何改进该算法以适应不断动态变化的复杂网络也是一个富有挑战的研究方向.

References

- 1 Zhao Y P, Levina E, Zhu J. Community extraction for social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, **108**(18): 7321–7326
- 2 Kelley S, Goldberg M, Magdon-Ismael M, Mertsalov K, Wallace A. Defining and discovering communities in social networks. *Handbook of Optimization in Complex Networks*, 2012, **57**(2): 139–168
- 3 Ángeles S M, Boguñá M, Sagués F. Uncovering the hidden geometry behind metabolic networks. *Molecular BioSystems*, 2012, **8**(3): 843–850
- 4 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: ACM, 2005. 661–669
- 5 Farutin V, Robison K, Lightcap E, Dancik V, Ruttenberg A, Letovsky S, Pradines J. Edge-count probabilities for the identification of local protein communities and their organization. *Proteins: Structure, Function, and Bioinformatics*, 2006, **62**(3): 800–818
- 6 Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 2005, **435**(7043): 814–818
- 7 Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**(2): 026113
- 8 Newman M E J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**(23): 8577–8582
- 9 Newman M E J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006, **74**(3): 6104–6126
- 10 Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*. New York, USA: ACM, 2011. 635–644
- 11 Shiga M, Takigawa I, Mamitsuka H. A spectral approach to clustering numerical vectors as nodes in a network. *Pattern Recognition*, 2011, **44**(2): 236–251
- 12 Newman M E J. Fast algorithm for detecting community structure in networks. *Physical Review E*, 2004, **69**(6): 6133–6138
- 13 Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, **74**(1): 6110–6122
- 14 He Dong-Xiao, Zhou Xu, Wang Zuo, Zhou Chun-Guang, Wang Zhe, Jin Di. Community mining in complex networks clustering combination based genetic algorithm. *Acta Automatica Sinica*, 2010, **36**(8): 1160–1170
(何东晓, 周栩, 王佐, 周春光, 王喆, 金弟. 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**(8): 1160–1170)
- 15 Jin Di, Liu Jie, Yang Bo, He Dong-Xiao, Liu Da-You. Genetic algorithm with local search for community detection in large-scale complex networks. *Acta Automatica Sinica*, 2011, **37**(7): 873–882
(金弟, 刘杰, 杨博, 何东晓, 刘大有. 局部搜索与遗传算法结合的大规模复杂网络社区探测. 自动化学报, 2011, **37**(7): 873–882)
- 16 Shen Hua-Wei, Cheng Xue-Qi, Chen Hai-Qiang, Liu Yue. Information bottleneck based community detection in network. *Chinese Journal of Computers*, 2008, **31**(4): 677–686
(沈华伟, 程学旗, 陈海强, 刘悦. 基于信息瓶颈的社区发现. 计算机学报, 2008, **31**(4): 677–686)
- 17 Guo Chong-Hui, Zhang Na. Partition methods based on common neighbors matrix for the community structure in complex network. *Systems Engineering – Theory & Practice*, 2010, **30**(6): 1077–1084
(郭崇慧, 张娜. 基于共邻矩阵的复杂网络社区结构划分方法. 系统工程理论与实践, 2010, **30**(6): 1077–1084)
- 18 Liu Xu, Yi Dong-Yun. Complex network community detection by local similarity. *Acta Automatica Sinica*, 2011, **37**(12): 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**(12): 1520–1529)
- 19 Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**(9): 2658–2663
- 20 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, **70**(6): 6111–6117

- 21 Jiang S Y, Song X Y, Wang H, Han J J, Li Q H. A clustering-based method for unsupervised intrusion detections. *Pattern Recognition Letters*, 2006, **27**(7): 802–810
- 22 Deng W B, Li W, Cai X, Wang Q P. The exponential degree distribution in complex networks: non-equilibrium network theory, numerical simulation and empirical data. *Physica A: Statistical Mechanics and Its Applications*, 2011, **390**(8): 1481–1485
- 23 Li Hui, Zhao Hai, Xu Jiu-Qiang, Li Bo, Li Peng, Wang Jia-Liang. Research on hierarchy of large-scale software macro-topology base on k -core. *Acta Electronica Sinica*, 2010, **38**(11): 2635–2643
(李辉, 赵海, 徐久强, 李博, 李鹏, 王家亮. 基于 k -核的大规模软件宏观拓扑结构层次性研究. 电子学报, 2010, **38**(11): 2635–2643)
- 24 Zhang Xin, Zhao Hai, Wang Li-Fei, Li Chao. Analysis on the internet AS-level topology. *Journal on Communications*, 2008, **29**(7): 50–61
(张昕, 赵海, 王莉菲, 李超. AS 级 Internet 拓扑分析. 通信学报, 2008, **29**(7): 50–61)
- 25 Zachary W W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, **33**(4): 452–473
- 26 Lusseau D. The emergent properties of a dolphin social network. *Proceedings of the Royal Society B: Biological Sciences*, 2003, **270**(S2): S186–S188
- 27 Adamic L A, Glance N. The political blogosphere and the 2004 U.S. election: divided they blog. In: *Proceedings of the 3rd International Workshop on Link Discovery*. New York, USA: ACM, 2005. 36–43
- 28 Newman M E J. Scientific collaboration networks, I: Network construction and fundamental results. *Physical Review E*, 2001, **64**(1): 016131
- 29 Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. *Physical Review E*, 2003, **8**(6): 065103
- 30 Gleiser P M, Danon L. Community structure in jazz. *Advances in Complex Systems*, 2003, **6**(4): 565–573
- 31 Jeong H, Mason S, Barabási A L, Oltvai Z N. Lethality and centrality in protein networks. *Nature*, 2001, **411**(6833): 41–42
- 32 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007, **76**(3): 036106



张新猛 广东外语外贸大学讲师, 主要研究方向为数据挖掘和自然语言处理。本文通信作者。

E-mail: javad0902@163.com

(**ZHANG Xin-Meng** Lecturer at Guangdong University of Foreign Studies. His research interest covers data mining and natural language processing.

ing. Corresponding author of this paper.)



蒋盛益 广东外语外贸大学教授, 博士。主要研究方向为数据挖掘和自然语言处理。E-mail: jiangshengyi@163.com

(**JIANG Sheng-Yi** Ph.D., professor at Guangdong University of Foreign Studies. His research interest covers data mining and natural language processing.)