

一种基于增量式谱聚类的动态社区自适应发现算法

蒋盛益^{1,3} 杨博泓¹ 王连喜^{2,3}

摘 要 针对当前复杂网络动态社区发现的热点问题, 提出一种面向静态网络社区发现的链接相关线性谱聚类算法, 并在此基础上提出一种基于增量式谱聚类的动态社区自适应发现算法. 动态社区发现算法引入归一化图形拉普拉斯矩阵呈现复杂网络节点之间的关系, 采用拉普拉斯本征映射将节点投影到 k 维欧氏空间. 为解决离群节点影响谱聚类的效果和启发式确定复杂网络社区数量的问题, 利用提出的链接相关线性谱聚类算法发现初始时间片的社区结构, 使发现社区的过程能够以较低的时间开销自适应地挖掘复杂网络社区结构. 此后, 对于后续相邻的时间片, 提出的增量式谱聚类算法以前一时间片聚类获得的社区特征为基础, 通过调整链接相关线性谱聚类算法实现对后一时间片的增量聚类, 以达到自适应地发现复杂网络动态社区的目的. 在多个数据集的实验表明, 提出的链接相关线性谱聚类算法能够有效地检测出复杂网络中的社区结构以及基于增量式谱聚类的动态社区自适应发现算法能够有效地挖掘网络中动态社区的演化过程.

关键词 谱聚类, 拉普拉斯矩阵, 拉普拉斯本征分析, 增量聚类, 社区检测, 动态社区

引用格式 蒋盛益, 杨博泓, 王连喜. 一种基于增量式谱聚类的动态社区自适应发现算法. 自动化学报, 2015, 41(12): 2017–2025

DOI 10.16383/j.aas.2015.c150290

An Adaptive Dynamic Community Detection Algorithm Based on Incremental Spectral Clustering

JIANG Sheng-Yi^{1,3} YANG Bo-Hong¹ WANG Lian-Xi^{2,3}

Abstract To tackle the hot issue of dynamic community detection in complex networks, this paper proposes an adaptive dynamic community detection algorithm based on the spectral clustering algorithm of linear link-related approach for static community detection and the incremental spectral clustering. The dynamic community detection algorithm uses the normalized Laplacian matrix to present the relationships of the complex network nodes, and then uses the Laplacian eigen map to map the node into the k -dimensional Euclidean space. In order to solve the problems of the outliers to nodes and the heuristics to identify the number of communities, this paper applies the spectral clustering algorithm of linear link-related approach to detect the community structure of the initial time slice, which makes the process of the complex community detection in an adaptive and computationally inexpensive way. For the subsequent adjacent time slices, the incremental spectral clustering algorithm of this paper aims to detect the dynamic community of complex network adaptively by clustering results for the next time slice based on the foundation of community characteristics of the previous time slice, and adjusting the spectral clustering algorithm of linear link-related approach. Experimental results on a series of datasets show that the spectral clustering algorithm of linear link-related approach can detect the community structure in the complex networks effectively and that the adaptive dynamic community detection algorithm based on incremental spectral clustering can also mine the evolution of dynamic communities effectively.

Key words Spectral clustering, Laplacian matrix, Laplacian eigen analysis, incremental clustering, community detection, dynamic community

Citation Jiang Sheng-Yi, Yang Bo-Hong, Wang Lian-Xi. An adaptive dynamic community detection algorithm based on incremental spectral clustering. *Acta Automatica Sinica*, 2015, 41(12): 2017–2025

收稿日期 2015-05-15 录用日期 2015-09-06

Manuscript received May 15, 2015; accepted September 6, 2015

国家自然科学基金 (61572145), 教育部人文社会科学研究青年项目 (14YJC870021), 广东省科技计划项目 (2014A040401083, 2015A030401093), 广东省普通高校科技创新项目 (2012KJ CX0049), 广东外语外贸大学研究生科研创新项目 (15GWCXXM-40), 广东大学生科技创新培育专项资金 (308-GK151018) 资助

Supported by National Natural Science Foundation of China (61572145), Youth Project of Humanities and Social Sciences of the Ministry of Education (14YJC870021), Science and Technology Project of Guangdong Province (2014A040401083, 2015A030401093), Science and Technology Innovation Project

of Guangdong Province (2012KJ CX0049), Guangdong University of Foreign Studies Graduate Student Research Innovation Project (15GWCXXM-40), Guangdong College Students' Scientific and Technological Innovation to Cultivate Special Funds (308-GK151018)

本文责任编辑 赵铁军

Recommended by Associate Editor ZHAO Tie-Jun

1. 广东外语外贸大学思科信息学院 广州 510006 2. 广东外语外贸大学图书馆 广州 510420 3. 语言工程与计算广东省社会科学重点实验室 广州 510006

1. School of Informatics, Guangdong University of Foreign Studies, Guangzhou 510006 2. Guangdong University of For-

真实世界中存在着多种同构或异构的复杂网络, 复杂网络呈现了节点之间的相互联系. 大量的研究表明, 有效地挖掘出复杂网络中的社区结构能够帮助人们更好地理解网络的拓扑结构和节点的行为特征^[1]. 因此, 近年来涌现出了大量的社区发现相关的研究成果.

社区发现 (Community detection), 又称社区检测, 是以将网络划分为若干个子网络, 并使得同一子网络的节点联系较为密切, 不同子网络的节点联系较为稀疏为目标的过程. 社区发现算法一般分为基于分裂的算法、基于模块度的算法、基于标签传播的算法以及基于谱聚类的算法^[1]. 其中具有代表性的算法包括 GN 算法^[2]、CNM 算法^[3]、BGLL 算法^[4]、LPA 算法^[5]、Semantic-LPA^[6]、CPM 算法^[7]及 NEC 算法^[8]等.

尽管复杂网络静态社区发现的研究已有较为成熟的理论体系, 但对于追踪网络社区的演化, 挖掘网络中的动态社区研究仍是现阶段的热点问题. 当前, 对于同构网络, 动态社区发现算法一般分为基于时空独立评价的方法^[9-10]、基于时空集成评价的方法^[11-15]、基于统一评价的方法^[16-22]以及基于增量聚类的方法^[23-24]. 然而, 基于时空独立评价的方法和基于时空集成评价的方法存在普适性差的问题, 基于统一评价的方法存在优化目标模型时间复杂度较高的问题以及基于增量聚类的方法存在通过降低社区检测精度达到降低时间复杂度的问题.

为解决离群节点影响谱聚类效果、启发式的确定复杂网络社区数量以及解决时空独立评价和时空集成评价的低普适性问题, 本文提出一种基于增量式谱聚类的动态社区自适应发现算法, 该算法通过引入归一化图形拉普拉斯矩阵, 采用拉普拉斯本征映射将网络节点投射到 k 维欧氏空间, 并利用提出的链接相关线性谱聚类算法对初始时间片对应的网络快照进行社区发现, 对后续的时间片, 均以上一时间片的社区摘要为基础, 采用调整的链接相关线性谱聚类进行增量聚类, 以挖掘复杂网络社区的演化过程. 在多个真实网络数据集上的实验表明, 提出的算法是有效和可行的. 本文的主要贡献包括如下几点:

1) 提出一种链接相关的线性谱聚类算法, 解决了离群节点影响谱聚类效果和启发式的确定复杂网络社区数量的问题.

2) 提出一种基于增量式谱聚类的动态社区自适应发现算法. 通过以上一时间片的社区摘要为基础, 解决时空独立评价或时空集成评价存在的低普适性

问题.

本文章节安排如下: 第 1 节介绍背景知识; 第 2 节给出链接相关的线性聚类算法的相关描述; 第 3 节介绍基于增量式谱聚类的自适应动态社区发现算法; 第 4 节进行实验分析; 第 5 节总结全文.

1 背景知识

1.1 拉普拉斯矩阵及其本征映射

假设复杂网络 $G = (V, E)$ 在光滑的流模型上由节点集 $V = v_1, v_2, \dots, v_n$ 和连边集 $E = \{e_{ij} | 0 < i, j \leq n\}$ 组成, 网络邻接矩阵 $A \in \mathbf{R}^{N \times N}$, 定义对角阵 (度分布矩阵) $D \in \mathbf{R}^{N \times N}$, 对角元素 $D_{ii} = \sum_j A_{ij}$, 则标准的拉普拉斯矩阵定义为

$$L = D - A \quad (1)$$

规范化的图形拉普拉斯矩阵定义为

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \quad (2)$$

设规范化的图形拉普拉斯矩阵 $\tilde{L}$ 的本征维度为 m , 则拉普拉斯矩阵的本征分解可通过求解函数式 $\tilde{L}v = \lambda Dv$, 使得本征值从小到大进行排序, 得 $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_m$. 忽略 $\lambda_0 = 0$ 对应的特征向量 v_0 , 选取另外 $k = m - 1$ 个本征向量将复杂网络节点映射到 k 维欧氏空间.

$$v_i \in \mathbf{R}^N \mapsto \tilde{v}_i \in \mathbf{R}^k \quad (3)$$

式中,

$$\tilde{v}_i^T = [v_1(i), v_2(i), \dots, v_k(i)] \quad (4)$$

一般的, 求解拉普拉斯本征映射的时间开销为 $O(n^3)$, 但对于形如复杂网络的稀疏矩阵, 采用文献 [25] 提出的 Lanczos 算法能够将解本征方程的时间降至 $O(ekh + nk^2h + k^2h)$. 其中, k 为保守情况下保留的特征数量, h 为迭代至收敛的次数. 因为假设矩阵为稀疏矩阵, 连边数 e 的量级近似于节点数 n , 而保留的特征数量 $k \ll n$, 则该算法能在近似线性的时间开销获得本征分解的结果.

1.2 面向社区检测的谱聚类

谱聚类算法能够在光滑流模型上对任意形状的簇进行聚类且具有收敛于全局最优的特点, 被广泛应用于图像处理, 生物科学和社会网络分析等多个领域. 尽管对于不同的研究领域, 谱聚类存在多种具体的实现方案, 但谱聚类仍能被形式化的归纳为如下三个步骤:

步骤 1. 根据待处理数据构建表示样本矩阵 L ;

步骤 2. 通过求解本征方程计算矩阵 L 从小到大排列的 k 个非零本征值, 及其对应的本征向量;

步骤 3. 根据样本的特征向量进行一般化的聚类.

对于面向复杂网络社区发现的谱聚类算法, 不同的算法在构建表示样本矩阵的方法以及具体聚类的算法存在着差异. 文献 [8] 在提出模块度的基础上, 引入模块度矩阵表示网络结构, 并通过拉普拉斯映射后对网络进行划分, 实现网络社区结构的进一步发现. 文献 [26] 采用 k -means 算法对比了邻接矩阵、标准的拉普拉斯矩阵、规范化的拉普拉斯矩阵、模块度矩阵以及相关性矩阵在社区检测的效果, 得出邻接矩阵及规范化的拉普拉斯矩阵, 相比于其他三个矩阵能获得更理想的社区检测效果.

1.3 社区检测效果评价指标

Newman 等^[27] 提出的模块度是评价网络社区发现效果常用的指标. 模块度通过刻画社区内部节点之间的紧凑程度来量化复杂网络社区结构的质量. 尽管研究表明模块度存在着网络社区分辨率问题和极端退化现象, 但仍被相关领域学者广泛认可, 并涌现了许多以优化模块度为目标的社区发现算法.

设复杂网络图模型 $G = (V, E)$ 由节点集 $V = v_1, v_2, \dots, v_n$ 和连边集 $E = \{e_{ij} | 0 < i, j \leq n\}$ 组成, 模块度定义为

$$Q = \frac{1}{2M} \sum_{i,j} (A_{ij} - \frac{k_i k_j}{2M}) \delta(c_i, c_j) \quad (5)$$

式中, A_{ij} 表示网络节点 i 和节点 j 的连边权重, k_i 表示节点 i 的度, c_i 表示节点 i 归属的社区, M 为网络 G 连边权重总和, 当 $c_i = c_j$ 时, $\delta(c_i, c_j) = 1$, 反之, $\delta(c_i, c_j) = 0$.

2 链接相关的线性谱聚类算法

针对现有面向复杂网络社区发现的谱聚类算法存在离群节点影响聚类效果和无法自适应地确定网络社区结构数量等问题, 提出一种链接相关的线性谱聚类算法 (Linear approach link relative spectral clustering, LLC). 在给出算法的具体描述前, 先给出算法的相关定义.

设给定网络 $G = (V, E)$, 矩阵 $\tilde{L} \in \mathbf{R}^{N \times N}$ 为网络对应的规范化图拉普拉斯矩阵 (如图 1 所示), 解本征函数 $\tilde{L}v = \lambda Dv$ 得网络一组从小到大排列的特征值向量 $\lambda = \{\lambda_i | i \leq N\}$, 保留特征值向量中非负特征值后计算相邻特征上升梯度 $\tau_i = \lambda_{i+1} - \lambda_i$, 并选取最大特征间隔 $\tau_{\max}$ 对应的特征值下标 k . 根据式 (3) 分解矩阵 $\tilde{L}$ 的本征, 将节点映射到 k 维欧氏空间.

设网络 G 的一个 k 划分 $C = (C_1, C_2, \dots, C_k)$, 社区摘要 C_i 由社区特征、归属社区的节点集以及社

区邻接节点集共同定义:

$$C_i = \{center_i, member_i, adjacency_i\} \quad (6)$$

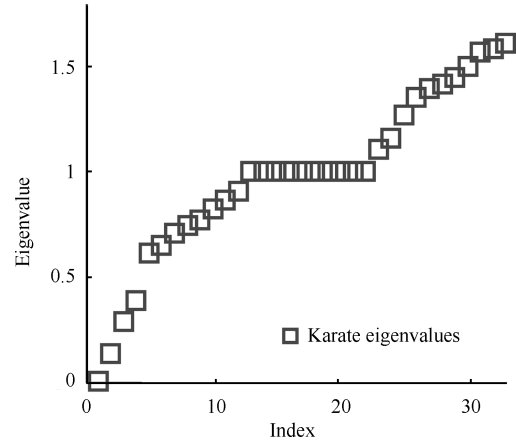


图 1 空手道数据集规范化拉普拉斯矩阵对应的谱
Fig. 1 The spectrum of Karate's normalized Laplacian

令矩阵 $Y \in \mathbf{R}^{N \times k}$ 表示网络 G 投影到欧氏空间的特征矩阵, 设定最大距离参数 ρ , 使得任意样本节点 $y_i \in C_j$ 均满足节点与社区距离 $Dis(y_i, C_j) < \rho$. 一般的, 设 $D_{ij} = Dis(y_i, y_j)$ 参数 ρ 计算公式如下:

$$\rho = \frac{1}{R} \sum_R D_{ij} + \alpha \sqrt{\frac{1}{R} \sum_R D_{ij}^2 - [\frac{1}{R} \sum_R D_{ij}]^2} \quad (7)$$

式中, 随机选取 R 对节点并计算节点对之间距离的均值及标准差, 通过修正参数 α 确定对参数 ρ 的取值.

设网络 G 的一个 k 划分 $C = (C_1, C_2, \dots, C_k)$, 判定节点 $y_i \in C_j$, 当且仅当节点 y_i 和社区 C_j 满足下述条件时成立:

$$Dis(y_i, C_j) = \min_c Dis(y_i, C_j) \quad (8)$$

$$Dis(y_i, C_j) < \rho \quad (9)$$

$$y_i \in C_j(adjacency) \quad (10)$$

也即, 样本节点归属于某一社区必须满足该节点与待归入的邻接社区之间的距离均小于节点与其他社区的距离, 且距离小于设定的最大距离参数 ρ .

下面给出链接相关的线性谱聚类算法的具体实现步骤:

算法 1. LLC

输入. 网络 G 的节点信息 V 及链接信息 E

输出. 社区摘要 C

步骤.

1) 构建网络 G 的邻接矩阵 A 和度分布矩阵 D ;

2) 根据式 (1) 将网络以标准的拉普拉斯矩阵 L 形式化表示;

3) 根据式 (2) 将标准的拉普拉斯矩阵转换为规范化的拉普拉斯矩阵 $\tilde{L}$;

4) 使用 Lanczos 算法对矩阵 $\tilde{L}$ 进行本征分解;

5) 根据特征间隔准则选取映射的前 k 个特征向量, 并构建特征矩阵 Y ;

6) 根据式 (7) 计算最大距离参数 ρ ;

7) 设初始模块度 $IQ = 0$;

8) 初始化选择任一节点作为第一个社区 C_0 ;

9) 选取下一节点, 计算节点与其邻接社区的距离 Dis , 若 $Dis_{\min} < \rho$, 则将节点归入离节点最近的邻接社区, 否则, 若节点未归属任何社区, 则以该节点创建新社区, 反之节点保留在原来社区;

10) 更新社区摘要;

11) 若仍有节点未被划分, 转 9);

12) 计算当前模块度 Q , 若 $Q > IQ$, 设所有节点均未被处理并转 9);

13) 算法结束.

LLC 算法的主要思路是通过迭代的形式将节点归入与节点距离最近的邻接社区或创建新社区, 不断提升模块度值, 从而得到质量的社区发现结果. 通过引入最大距离参数使得算法能够自适应地获得网络社区的数量且避免了离群节点影响社区发现效果. 算法整体可分为两个阶段, 即特征映射阶段和聚类阶段. 在特征映射过程中, 采用了 Lanczos 算法, 如上文分析, 该算法能够在近似线性时间复杂度的情况下计算出稀疏矩阵的本征映射; 在聚类过程中, 计算最大距离参数的时间复杂度为 $O(R)$. 设迭代次数为 I 、划分出的社区数量为 C 以及节点数量为 N , 聚类阶段的时间复杂度为 $O(R + ICN)$. 在实际问题中, 社区数量与节点数量常不在一个量级, 且通常 $C \ll N$, 在实验中, 算法能够在较少的迭代次数中完成对节点的划分, 且有 $IC \ll N$, 因此聚类过程的时间复杂度仍可理解为近线性的时间复杂度. 综上, 提出的链接相关的线性谱聚类算法的时间复杂度近似为 $O(N)$.

3 增量式动态社区发现算法

在链接相关的线性谱聚类的基础上提出一种基于增量式谱聚类的动态社区自适应发现算法 (Adaptive dynamic community detection algorithm based on incremental spectral clustering, ADCD). 在介绍算法的具体描述前, 同样先给出算法重要的相关概念.

短时间平滑性假设 (后文简称平滑性假设) 被广泛应用于动态社区检测研究中. 平滑性假设是假设复杂网络中的社区结构在短时间内的演化是平缓的^[12], 在较短的时间间隔内, 网络中的历史交互信息与当前的交互信息存在一定的相似性. 这意味着在动态社区检测研究可利用历史网络的不同时间片社区摘要来拟合当前时间片的社区摘要, 从而克服

时空独立导致的信息缺失及噪声等问题.

ADCD 算法在平滑性假设的基础上, 对相邻且间隔较短的两个时间片, 以前一时间片社区摘要信息中的社区特征初始化后一时间片各个社区摘要的社区特征并以当前网络的链接信息初始化当前社区的邻接信息, 采用适用于动态社区的链接相关的线性谱聚类算法对网络节点进行增量聚类, 实现对复杂网络的动态社区发现.

算法的具体步骤如下:

输入. 第一时间片的社区摘要, 其余时间片网络 G 的节点信息 V 及链接信息 E

输出. 各个时间片社区摘要 C

步骤.

1) 选取下一时间片作为当前的时间片;

2) 根据当前构建网络 G 的邻接矩阵 A 和度分布矩阵 D ;

3) 根据式 (1) 将网络以标准的拉普拉斯矩阵 L 形式表示;

4) 根据式 (2) 将标准的拉普拉斯矩阵转换为规范化的拉普拉斯矩阵 $\tilde{L}$;

5) 使用 Lanczos 算法对矩阵 $\tilde{L}$ 进行本征分解;

6) 根据特征间隔准则选取映射的前 k 个特征向量, 并构建特征矩阵 Y ;

7) 根据式 (7) 计算最大距离参数;

8) 初始化将 C' 的社区特征作为当前社区集 C 的社区特征, 并根据邻接矩阵 A 更新社区邻接信息;

9) 选取任一节点, 计算节点与其邻接社区的距离 Dis , 若 $Dis_{\min} < \rho$, 则将节点归入离节点最近的邻接社区, 否则, 以该节点创建新社区;

10) 若仍有节点未被划分, 转 9);

11) 更新社区摘要 C 并将 C 赋值给 C' ;

12) 若仍有后续时间片, 转 1);

13) 算法结束.

ADCD 算法在对下一时间片进行增量聚类前, 类似于 LLC 算法将网络以规范化的拉普拉斯矩阵表示并进行本征分解. 在将网络节点投影到欧氏空间后, 根据平滑性假设以上一时间片的社区特征和当前时间片的连接信息分别初始化当前时间片社区摘要中的社区特征及社区邻接信息. 在根据本征矩阵 Y 更新最大距离参数后, 不改变社区特征对网络节点进行自适应的增量聚类, 实现对社区的划分, 当网络所有节点均划分后, 更新社区摘要作为下一时间片的社区发现的基础. 算法整体时间复杂度与 LLC 相似, 设时间片数量为 T , 则算法的时间复杂度为 $O[T(N + R + CN)]$. 事实上, 在处理大规模复杂网络时, 其时间片的数量和节点数量不在同一量级, 因此算法的整体时间复杂度同理可看成是线性时间复杂度 $O(N)$.

4 实验

本文实验分为基于 LLC 算法的静态社区检测

有效性实验和基于 ADCD 算法的动态社区演化实验两个部分. 静态社区检测有效性的验证过程包括参数实验、LLC 算法与当前常用的社区检测算法的对比实验和社区发现可视化实验; 在动态社区演化实验中, 给出了 ADCD 算法与当前常用的两个算法在多个动态网络中发现动态社区演化的对比实验.

4.1 数据集

采用多个真实网络数据集进行实验, 其中静态网络数据集包括: 空手道俱乐部网络数据集 (Karate)^[28]、海豚网络数据集 (Dolphins)^[29]、爵士音乐节合作网络 (Jazz)^[30]、美国大学生美式足球网络数据集 (Football)^[31]、2004 年美国政治家博客网络数据集 (Polblogs)^[32]、电子邮件网络数据集 (Email)^[33] 以及 Celegans 新陈代谢网络数据集 (Cel)^[34]; 动态网络数据集包括: Enron Email 网络数据集^[35]、VAST2008 恐怖分子通信网络数据集^[36] 和 arXiv e-print 引用网络^[37].

4.2 参数实验

本文提出的 LLC 算法和 ADCD 算法均采用了最大距离参数 ρ , 而参数 ρ 受修正参数 α 所控制, 因此, 要验证最大距离参数 ρ 对行社区检测的影响, 仅需对修正参数 α 进行控制实验即可. 当 α 在区间 $[0, 1]$ 变化时, LLC 算法对 Karate、Football 和 Jazz 三个网络数据集进行社区检测的效果如图 2 和图 3 所示.

图 2 和图 3 展示了 LLC 算法产生的模块度 Q 和社区数量随着参数 α 取值的变化. α 取较大值时, LLC 算法发现社区数量较多, 而 α 取较小值时, LLC 算法将整个网络看作一个社区, 算法执行效果都不理想; 当 α 在 $[0.25, 0.8]$ 之间, LLC 算法在模块度指标和社区数量均获得较理想的结果, 即当此时算法能获得正常的社区划分结果.

4.3 算法有效性实验

由于 ADCD 算法是以 LLC 算法为基础, 因

此, 证实 LLC 算法能够有效地获得网络社区划分即能同时得出 ADCD 算法能有效地对静态网络进行社区划分的结论. 在实验过程中, 通过对比分析多个算法划分网络社区的模块度以及划分的社区数量, 验证本文提出的 LLC 算法的有效性. 对比实验的算法包括: GN 算法、CNM 算法、LPA 算法、Similarity 算法^[38]、自旋玻璃模型算法 (Spinglass 算法)^[39]、Newman 提出的模块度矩阵谱聚类算法 (NEC 算法)^[8]. 表 1 给出了各个算法的模块度和社区数量, 除 LLC 算法结果外, 数据均来自于文献 [37]. 表 1 加粗数据表示社区发现得到的模块度最优的数值. 由表 1 可以得出 Spinglass 算法和本文提出的 LLC 算法在大部分网络数据集中均能获得较为理想社区发现结果的结论. 值得说明的是, 在与同为基于谱聚类的 NEC 算法进行对比中, LLC 算法在大部分数据集上能够获得更理想的社区发现结果, 因此可得出 LLC 算法较 NEC 更具有优势的结论.

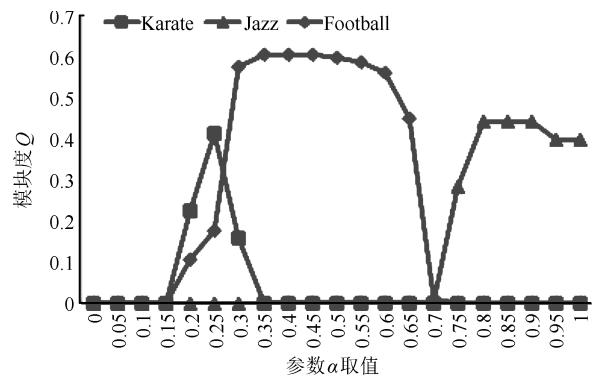


图 2 参数 α 对模块度的影响

Fig. 2 The influence of modularity via α

4.4 可视化实验

为验证 LLC 算法社区发现结果与网络实际情况吻合, 给出算法划分社区的可视化结果. 考虑到节

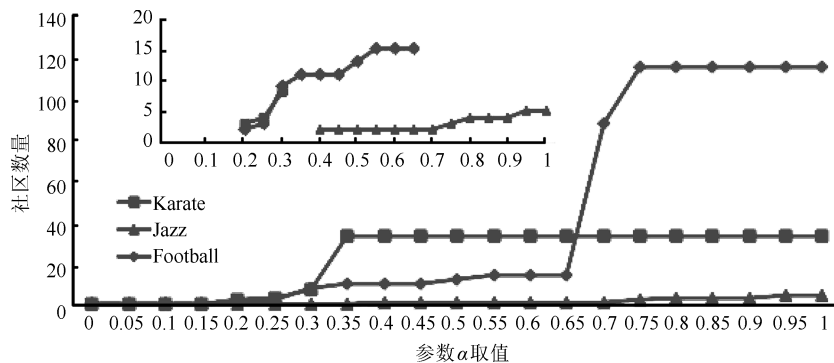


图 3 参数 α 对社区数量的影响

Fig. 3 The influence of community number via α

表 1 各算法社区发现结果

Table 1 Community detection results of different algorithms

Network	Nodes/Edges	GN	CNM	LPA	NEC	Spinglass	Similarity	LLC
Karate	34/78	0.40/5	0.38/3	0.37/3	0.38/5	0.42/4	0.41/4	0.42/4
Dolphins	62/159	0.52/5	0.50/4	0.51/4	0.49/6	0.53/5	0.52/4	0.49/4
Jazz	198/2 742	0.41/39	0.44/4	0.28/2	0.35/8	0.44/5	0.40/4	0.44/4
Football	115/613	0.60/10	0.55/6	0.58/12	0.45/13	0.60/10	0.60/10	0.60/11
Polblogs	1 222/16 714	0.42/205	0.43/10	0.43/3	0.27/62	0.43/16	0.38/2	0.43/2
Cel	453/2 025	0.40/38	0.39/9	0.33/8	0.32/28	0.44/10	0.37/14	0.48/7
Email	1 133/5 451	0.53/61	0.49/16	0.01/4	0.42/45	0.58/12	0.45/16	0.52/6

点过多的网络会产生视觉模糊,无法辨别社区结构的情况,本文对节点数量较少的网络数据集 Karate 和 Jazz 进行可视化,具体如图 4 和图 5 所示。

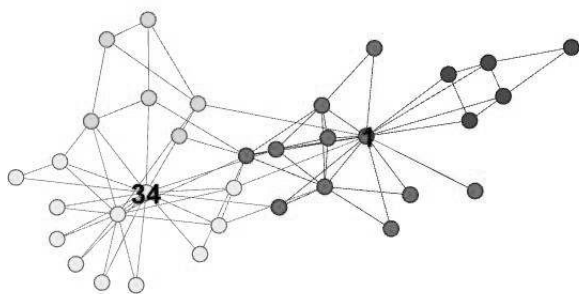


图 4 Karate 社区结构图

Fig. 4 Community graph of Karate network

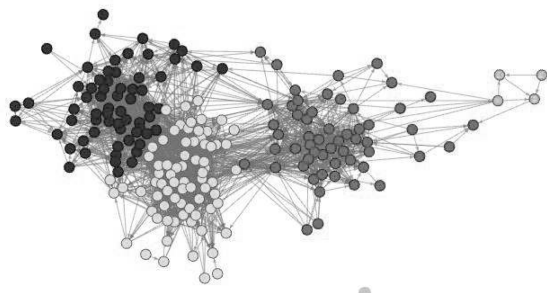


图 5 Jazz 社区结构图

Fig. 5 Community graph of Jazz network

Karate 数据集是美国社会学家 Zachary 观察美国某家空手道俱乐部 34 名成员之间的社会关系而得到的网络数据集。由于俱乐部的主管(节点 34)与教练(节点 1)之间的争执,俱乐部成员分为了两个“党派”,因此,该网络的正确划分应该划分为两个社区结构。如图 4 所示,LLC 算法在发现 Karate 网络中两个党派的基础上将网络进一步划分为 4 个社区,且划分的结果符合网络的现实情况。Jazz 数据集描述的是 198 支爵士乐队在 1912 年至 1940 年之间的合作关系。如图 5 所示,LLC 算法能够将 Jazz 音乐家网络划分出 4 个社区结构,且同一社区的节点联系较为密切,不同社区的节点联系较为稀

疏。综上,本文提出的 LLC 算法能够自适应地对网络进行社区检测,且发现的社区结构与网络真实情况吻合。

4.5 动态社区演化实验

为观察 ADCD 算法挖掘复杂网络中动态社区结构的有效性,将之与常用的动态社区算法 QCA 算法^[40]和 MIEN 算法^[41]进行对比实验。通过观察模块度、社区数量和时间开销三个指标,分析各个算法的效果。具体实验结果如图 6、图 7 和图 8 所示。

Enron Email 网络数据集包含了 Enron 公司在 1999 年 12 月至 2002 年 7 月 151 位高级管理人员的所有邮件交互记录。为模拟网络动态演化的过程,本文选取 Enron Email 网络中 50% 的连边构建基础网络,并以每个时间片增加约 1000 条连边的方案构建 21 个静态时间节点对应的网络结构。在图 6(a)中,本文给出了 3 个算法在动态网络中进行社区检测后计算出的模块度。如该图所示,MIEN 算法得到的模块度在整体上由于另外两个算法,且在各个时间片中都得到最大的模块度值。在图 6(b)给出的社区数量演化情况可以看出,网络整体上可划分为 7 个社区,QCA 算法在检测的整个过程中社区数量没有发生变化,而 MIEN 算法和 ADCD 算法均在个别时刻中发生社区数量的变化。尽管 MIEN 算法较 QCA 算法和 ADCD 算法能得到更高的模块度,但由图 6(c)可以得出,MIEN 算法存在着在社区检测过程中时间开销过大的问题。

arXiv e-print 引用网络包含了超过 255 000 篇论文,且每年有超过 40 000 的增长量。本文采用的数据集时间跨度为 1996 年 1 月至 2003 年 5 月。在实验中,本文以 1996 年和 1997 年两年的数据构建基础网络,并在随后跨时间 1998 年 1 月至 2003 年 1 月以两个月为时间片构建 30 个网络快照。在实验结果中,随着时间的推移,ADCD 发现社区后的模块度明显高于其余两个算法(图 7(a)),MIEN 算法发现社区的数量和时间开销上明显多于其余两个算法(图 7(b)和图 7(c))。

VAST Calling 网络是 IEEE 在 VAST2008 年

竞赛中开放的恐怖分子通话数据集. 该数据集包含了 400 个用户在 10 天内的通话记录. 在对数据进行统计处理后, 以每一天作为一个时间片构建 10 个网络快照, 并以第一个网络快照作为基础网络. 由实验结果 (图 8) 可以得出, QCA 算法和 ADCD 算法的模块度相近且高于 MIEN 算法, QCA 算法在各个时间片检测得到的社区数量均为 10 个社区, 而 ADCD 算法和 MIEN 算法检测出的社区数量随着时间的变化发生增减演化, 在时间开销中, MIEN 算法的时间开销仍然是 3 个算法中最大的, 而 QCA 算法和 ADCD 算法的时间开销较为接近.

综上分析, 本文提出的 ADCD 算法能够与当前较为常用的动态社区检测算法得到相近的社区演化结构, 且在时间开销上能获得一定的优势.

5 结论

针对复杂网络动态社区检测的热点问题, 本文先提出了一种链接相关的线性谱聚类算法 LLC, 然后在 LLC 的基础上提出一种基于增量式谱聚类的动态社区自适应发现算法 ADCD. 通过在多个数据集上的对比实验证明了 LLC 和 ADCD 算法的有效性.

尽管本文提出 LLC 算法和 ADCD 算法具有较好的性能, 但是他们都是针对无向同构网络的. 因此, 在后续研究工作中, 将对 LLC 算法和 ADCD 算法进行更多的实验, 并验证算法在其他网络结构中的可行性, 同时将算法拓展至分布式环境中, 使算法能适用于极大规模的复杂网络中.

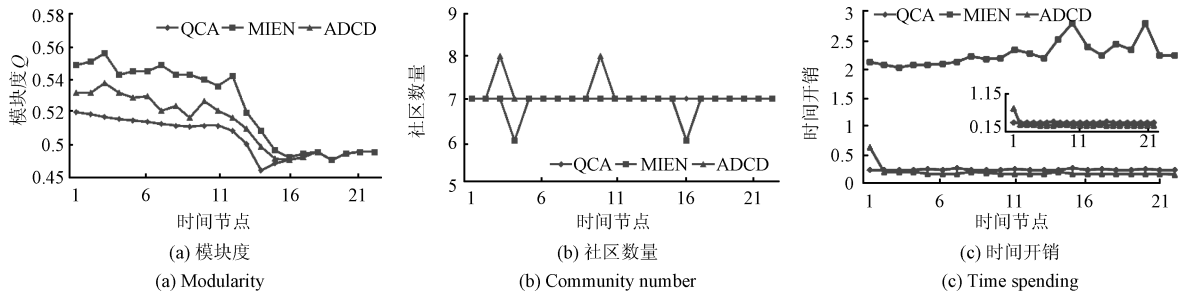


图 6 各算法挖掘的 Enron Email 网络动态社区演化结果

Fig. 6 The results of Enron Email community's evolution by 3 algorithms

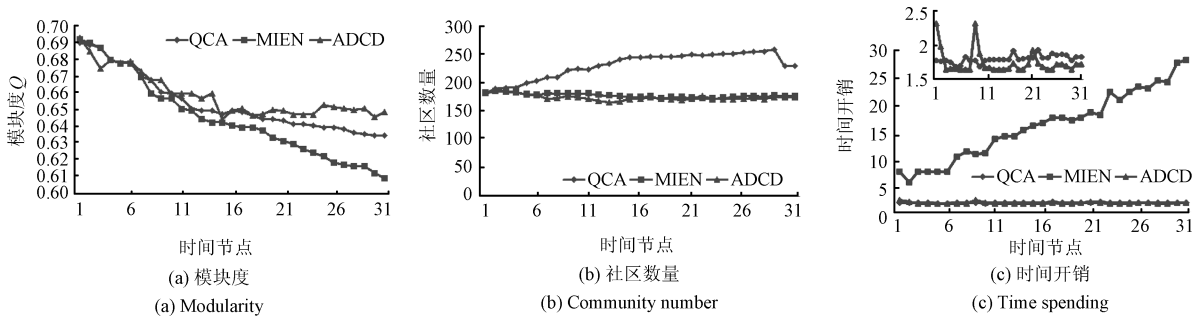


图 7 各算法挖掘的 arXiv 网络动态社区演化结果

Fig. 7 The results of arXiv community's evolution by 3 algorithms

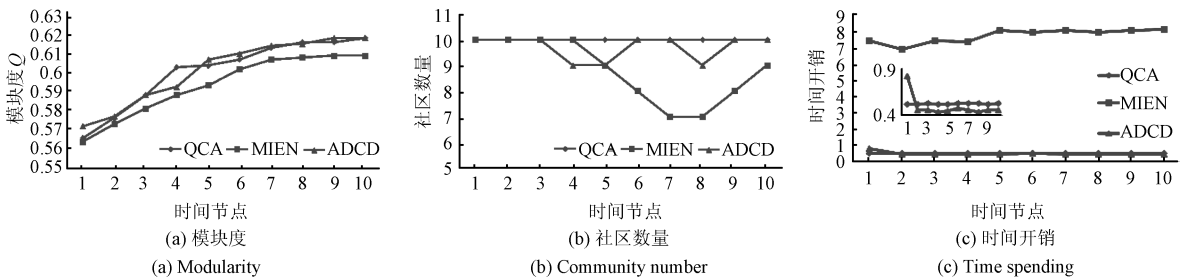


图 8 各算法挖掘的 VAST Calling 网络动态社区演化结果

Fig. 8 The results of VAST Calling community's evolution by 3 algorithms

References

- Fortunato S. Community detection in graphs. *Physics Reports*, 2010, **486**(3–5): 75–174
- Girvan M, Newman M E J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2002, **99**(12): 7821–7826
- Clauaset A, Newman M E J, Moore C. Finding community structure in very large networks. *Physical Review E*, 2004, **70**(6): 066111
- Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, **2008**(10): P10008
- 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
- Xin Yu, Yang Jing, Xie Zhi-Qiang. An overlapping semantic community structure detecting algorithm by label propagation. *Acta Automatica Sinica*, 2014, **40**(10): 2262–2275
(辛宇, 杨静, 谢志强. 基于标签传播的语义重叠社区发现算法. 自动化学报, 2014, **40**(10): 2262–2275)
- 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
- Newman M E J. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 2006, **74**(3): 036104
- Hopcroft J, Khan O, Kulis B, Selman B. Tracking evolving communities in large linked networks. *Proceedings of the national academy of sciences of the United States of America*, 2004, **101**(Suppl 1): 5249–5253
- Palla G, Barabási A L, Vicsek T. Quantifying social group evolution. *Nature*, 2007, **446**(7136): 664–667
- Mucha P J, Richardson T, Macon K, Porter M A, Onnela J P. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 2010, **328**(5980): 876–878
- Chi Y, Song X D, Zhou D Y, Hino K, Tseng B L. Evolutionary spectral clustering by incorporating temporal smoothness. In: *Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2007. 153–162
- Kim M S, Han J W. A particle-and-density based evolutionary clustering method for dynamic networks. *Proceedings of the VLDB Endowment*, 2009, **2**(1): 622–633
- Sarkar P, Moore A W. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter*, 2005, **7**(2): 31–40
- Lin Y R, Chi Y, Zhu S H, Sundaram H, Tseng B L. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. In: *Proceedings of the 17th International Conference on World Wide Web*. New York, NY, USA: ACM, 2008. 685–694
- Zhang H Z, Qiu B J, Giles C L, Foley H C, Yen J. An LDA-based community structure discovery approach for large-scale social networks. In: *Proceedings of the 2007 IEEE Intelligence and Security Informatics*. New Brunswick, NJ, Canada: IEEE, 2007. 200–207
- Zhang H Z, Giles C L, Foley H C, Yen J. Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*. Vancouver, British Columbia, Canada: AAAI, 2007. 663–668
- Hofman J M, Wiggins C H. Bayesian approach to network modularity. *Physical Review Letters*, 2008, **100**(25): 258701
- Xu T B, Zhang Z F, Yu P S, Long B. Evolutionary clustering by hierarchical dirichlet process with hidden Markov state. In: *Proceedings of the 8th IEEE International Conference on Data Mining*, 2008 (ICDM'08). Pisa, Italy: IEEE, 2008. 658–667
- Xu T B, Zhang Z F, Yu P S, Long B. Generative models for evolutionary clustering. *ACM Transactions on Knowledge Discovery from Data*, 2010, **6**(2): 681–690
- Airoldi E M, Blei D M, Fienberg S E, Xing E P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 2009, **9**(5): 1981–2014
- Ho Q R, Song L, Xing E P. Evolving cluster mixed-membership blockmodel for time-evolving networks. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA: AISTATS, 2011. 342–350
- Hartmann T, Kappes A, Wagner D. Clustering evolving networks. *arXiv preprint arXiv: 1401.3516*, 2014.
- Lee P, Lakshmanan L V S, Milios E E. Incremental cluster evolution tracking from highly dynamic network data. In: *Proceedings of the 30th International Conference on Data Engineering (ICDE)*. Chicago, IL, USA: IEEE, 2014. 3–14
- Golub G H, van Loan C F. *Matrix Computations* (Fourth edition). Baltimore, USA: JHU Press, 2012.
- Shen H W, Cheng X Q. Spectral methods for the detection of network community structure: a comparative analysis. *Journal of Statistical Mechanics: Theory and Experiment*, 2010, **2010**(10): P10020
- Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, **69**(2): 026113
- Zachary W W. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 1977, **33**(4): 452–473
- Lusseau D. The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2003, **270**(Suppl 2): S186–S188
- Gleiser P M, Danon L. Community structure in jazz. *Advances in Complex Systems*, 2003, **6**(4): 565–573
- 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)

- 32 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, NY, USA: ACM, 2005. 36–43
- 33 Guimerá R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Self-similar community structure in a network of human interactions. *Physical Review E*, 2003, **68**(6): 065103
- 34 Jeong H, Mason S P, Barabási A L, Oltvai Z N. Lethality and centrality in protein networks. *Nature*, 2001, **411**(6833): 41–42
- 35 Sun J M, Faloutsos C, Papadimitriou S, Yu P S. GraphScope: parameter-free mining of large time-evolving graphs. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007. 687–696
- 36 Grinstein G, Plaisant C, Laskowski S, O'Connell T, Scholtz J, Whiting M. VAST 2008 challenge: introducing mini-challenges. In: Proceedings of 2008 IEEE Symposium on Visual Analytics Science and Technology (VAST'08). Columbus, OH, USA: IEEE, 2008.
- 37 ArXiv Citation Dataset. Kdd cup 2003 [Online], available: <http://www.cs.cornell.edu/projects/kddcup/datasets.html>, Feb, 2003.
- 38 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)
- 39 Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 2006, **74**(1): 016110
- 40 Nguyen N P, Dinh T N, Xuan Y, Thai M T. Adaptive algorithms for detecting community structure in dynamic social networks. In: Proceedings of the 2011 Proceedings IEEE INFOCOM. Shanghai, China: IEEE, 2011. 2282–2290
- 41 Dinh T N, Xuan Y, Thai M T. Towards social-aware routing in dynamic communication networks. In: Proceedings of the

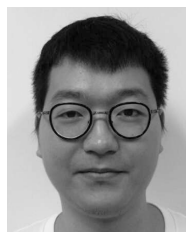
28th International Performance Computing and Communications Conference (IPCCC). Scottsdale, AZ, USA: IEEE, 2009. 161–168



蒋盛益 广东外语外贸大学思科信息学院教授. 主要研究方向为数据挖掘、文本挖掘和用户关系挖掘.

E-mail: jaingshengyi@163.com

(**JIANG Sheng-Yi** Professor at Cisco School of Informatics, Guangdong University of Foreign Study. His research interest covers data mining, text mining, and user relation mining.)



杨博泓 广东外语外贸大学思科信息学院硕士研究生. 主要研究方向为用户关系挖掘和个性化推荐. 本文通信作者.

E-mail: boryee@qq.com

(**YANG Bo-Hong** Master student at Cisco School of Informatics, Guangdong University of Foreign Study. His research interest covers user relation

mining and recommended system. Corresponding author of this paper.)



王连喜 中山大学资讯管理学院博士研究生. 主要研究方向为数据挖掘、特征选择和自然语言处理.

E-mail: wanglianxi2012@163.com

(**WANG Lian-Xi** Ph.D. candidate at School of Information Management, Sun Yet-Sam University. His research interest covers data mining, feature selection and nature language process.)