一种基于节点间路径度量的图聚类算法

郑文萍[1),2),3)] 车晨浩[1)] 钱宇华[1),2),3)] 王 杰[1)] 杨 贵[1)]

[1)](山西大学计算机与信息技术学院 太原 030006)

[2)](山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)

[3)](山西大学大数据科学与产业研究院 太原 030006)

摘 要 图聚类算法可以用于发现社会网络中的社区结构、蛋白质互作用网络中的功能模块等,是当前复杂网络研究的热点之一.对网络中节点的相似性和簇发现结果进行合理度量是核心问题.针对此问题,给出了一种基于节点间点不重复路径度量的节点相似性指标.以此为基础提出了一种面向复杂网络的基于"中心-扩展"策略的图聚类算法(A Graph Clustering Algorithm Based on Local Paths between Nodes in Complex Networks,PGC),包括节点相似性计算、中心节点选择、初始簇划分和簇优化四个主要过程.采用点不重复路径对节点相似性进行度量,消除了由大度节点引起较多的点重复路径对节点相似性的影响,提高了算法对大度节点邻域中节点的划分能力.通过与一些经典算法在 11 个真实网络、22 个人工网络数据集上的实验比较分析,结果表明算法 PGC 在标准互信息、调整兰德系数、F 度量、准确度等方面均表现出良好的性能.

关键词 复杂网络;图聚类;簇结构;相似性度量;连通性

中图法分类号 TP301 **DOI 号** 10.11897/SP.J.1016.2020.01312

A Graph Clustering Algorithm Based on Paths Between Nodes in Complex Networks

ZHENG Wen-Ping[1),2),3)] CHE Chen-Hao[1)] QIAN Yu-Hua[1),2),3)] WANG Jie[1)] YANG Gui[1)]

[1)](*School of Computer & Information Technology, Shanxi University, Taiyuan* 030006)

[2)](*Key Laboratory Computational Intelligence & Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan* 030006)

[3)](*Research Institute of Big Data Science and Industry, Shanxi University, Taiyuan* 030006)

Abstract Many complex systems can be modeled as complex networks, such as social network, protein interaction network, citation network, metabolic network etc. Nodes in a complex network often can be grouped into different clusters, called communities. Nodes in the same group form specific functional modules through tight intra-connection, and nodes from different group have relatively loose inter-connection to ensure cooperation among the functional modules of the system. Detecting community structures is crucial to understand the topological structure and dynamic characteristics of networks. Based on analyzing connecting patterns within and between communities, researchers can discover the functional modules and their evolution processes in various complex systems. Many methods have been put forward to detect communities. Among these, core-extension-based methods show good performance in efficiency and effectiveness. There are two essential parts in core-extension algorithms: seed detection and community extension. Seed detection process locates seeds with high centrality. Then, communities can be built from

收稿日期:2018-03-16;在线发布日期:2019-05-30.本课题得到国家自然科学基金项目(61572005)、山西省自然科学基金(201801D121123)和山西省回国留学人员科研基金项目(2017-014)资助.**郑文萍**,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为图论算法、生物信息学.E-mail:wpzheng@sxu.edu.cn.**车晨浩**,硕士研究生,主要研究方向为聚类算法.**钱宇华**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能、机器学习.**王 杰**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为数据挖掘、生物信息学.**杨 贵**,博士研究生,主要研究方向为生物信息学.

the seeds based on node similarity metrics and proper quality function in community extension process. Node similarity metrics play important roles in community detection algorithms. Lots of methods have been proposed to measure similarity of nodes in complex networks. For example Jaccard Index based methods measure nodes' similarity based on their common direct neighbors. Katz Index based methods measure nodes' similarity based on the walks between two nodes. Comparing with Jaccard Index based methods, Katz Index takes advantage of general structure topology information. LS Index measures node's similarity based on the local walks (lengths of walks are no larger than 3) between nodes, and can measure similarity between nodes by using their local connectivity information rather than their direct neighbors. LS Index simplifies the calculation and improve the efficiency, but it is still affected by other structure features such as node's degree and clustering coefficient. For a node with relatively larger degree in a network, it might occur at higher frequencies in paths between two nodes in its direct neighborhood. The nodes in the direct neighborhood of a large-degree node tend to have higher similarities. As a result, LS index based methods tend to group the nodes in the neighborhood of a large-degree node into the same cluster. However, these nodes are often grouped into different clusters in practical networks. In this paper, we propose a graph clustering algorithm, called PGC. We define a novel node similarity index SLP based on vertex non-repetitive paths between nodes. The proposed SLP Index weakens the influence of large-degree nodes on the calculation of nodes' similarity, and can reflect the connectivity degree between two nodes in the network. First, the proposed PGC algorithm calculates nodes' SLP similarity, and determines node weights based on SLP. Second, PGC chooses the node with the highest weight as the first seed node, then selects other seed nodes by considering node weights as well as their similarities with the existing seeds. Then, PGC obtains initial partition by attaching each unseeded node to the seed with the highest SLP similarity with it. Finally, PGC optimizes the initial partition iteratively to maximize the cluster quality evaluation function which is based on complementary entropy. Experimental results show that SLP Index eliminates the influence on the nodes' similarity caused by vertex repetitive paths, and improves the algorithm's ability to cluster the nodes in the neighborhood of large-degree nodes. Compared with other classical graph clustering algorithms on 11 real networks and 22 artificial networks, the proposed algorithm PGC shows a preferable performance.

Keywords　complex network; graph clustering; cluster structure; node similarity; connectivity

1　引　言

现实中很多系统都可以表示成复杂网络，如蛋白质互作用网、基因关联网、新陈代谢网、社会关系网、科学家合作网、交通运输网、电力传输网等[1-4]，其中网络节点表示系统组分，边表示组分间的作用关系. 通过分析这些复杂系统对应网络模型的拓扑特性和网络成分间的复杂关系，研究者可以更深入地理解复杂系统. 而网络科学理论的快速发展也为研究者探索复杂系统提供了新的研究方式.

复杂网络内部通常呈现出与功能关联的簇结构，即可以将网络中的节点分组，组内节点通过紧密的相互联系形成了特定的功能模块，而组间节点的相互联系确保系统各功能模块间的协同工作. 通常，簇内节点间较簇间具有更强的关联关系，如社会网络中普遍存在的社区结构、蛋白质互作用网络中存在的蛋白质复合体等. 对网络成员间进行关联分析而挖掘网络中的簇结构，有助于深入研究各种类型复杂网络的功能模块及其演化特征，对准确地理解并分析复杂系统的拓扑结构及动力学特性具有十分重要的理论意义和应用价值[5].

节点相似性度量在图聚类算法起着重要作用，ISCD+[6]算法和 Chen 等人[7]的算法根据簇内节点间的连通能力定义节点间的相似性，进行了有效的簇发现. 为了更好地对网络中大度节点的局部邻域

进行划分，本文给出了一种基于节点间局部路径连通性的图聚类算法 PGC(Local Paths Based Graph Clustering Method)．算法通过节点间的局部点不重复路径定义节点间相似性，并在此基础上给出了一种基于“中心-扩展”策略的图聚类算法．本文所提出的节点相似性指标消除了由大度节点引起较多的点重复路径对节点相似性的影响，提高了算法对大度节点邻域中节点的划分能力．在真实网络与人工网络上的实验结果表明算法 PGC 在标准互信息、调整兰德系数、F-measure 等方面均表现出良好的性能．

2 相关工作

2.1 节点相似性指标

节点间的相似性度量是图聚类算法的关键问题．基于公共邻居的节点相似性度量在图聚类算法中应用广泛，如共同邻居 CN 指标、Salton 指标、Jaccard 指标、Sorenson 指标等[8-9]．基于公共邻居相似性度量对两个节点的一阶邻域的相似性进行比较，对两个节点在更高阶邻域内的相似关系刻画不足．

基于节点间的连通能力可以在更广泛邻域内对节点间相似性进行度量，如 Katz 指标、LP 指标和 LHN-II 指标等[9]．

Katz 指标利用网络中两节点间的所有路径来度量节点相似性，越短的路径赋予越高的权重，定义为

$$\boldsymbol{S}_{\mathrm{Katz}}=\beta\boldsymbol{A}+\beta^2\boldsymbol{A}^2+\beta^3\boldsymbol{A}^3+\cdots=(\boldsymbol{I}-\beta\boldsymbol{A})^{-1}-\boldsymbol{I} \quad (1)$$

其中，$\boldsymbol{A}$ 为网络的邻接矩阵，$\boldsymbol{I}$ 是单位矩阵．Katz 指标考虑了网络中两节点间的所有路径信息，因此计算代价通常较高．

2009 年 Zhou 等人[10]根据网络中节点间的长度为 2 和 3 的路径数定义节点间的相似性度量 LP 指标．通常 $\beta=0.1$．

$$\boldsymbol{S}_{\mathrm{LP}}=\boldsymbol{A}^2+\beta\boldsymbol{A}^3 \quad (2)$$

2015 年 Chen 等人[7]根据网络中节点间的长度不超过 3 的路径数定义节点间的相似性度量 LS 指标．通常 $\beta=0.4$．

$$\boldsymbol{S}_{\mathrm{LS}}=\boldsymbol{A}+\beta\boldsymbol{A}^2+\beta\boldsymbol{A}^3 \quad (3)$$

其中，具有直接连边的两个节点相似性较 LP 指标明显增加．

基于节点间连通性的指标度量了节点在更高阶邻域内的相似性．然而，网络中的大度节点在一条路径中重复出现的可能性更高，因此大度节点直接邻域内的节点具有更高的相似性．这导致基于节点间连通性的图聚类算法倾向于将大度节点邻域中的节点划分到同一个簇，对大度节点邻域中节点的划分能力不足．

2.2 相关图聚类算法

研究人员提出了许多算法来解决图聚类问题．2002 年 Girvan 和 Newman 提出的 GN 算法[11]是一种分裂型的层次图聚类方法，依次删除网络中介数最高的边，得到聚类结果．GN 算法的基本假设是一条边的介数越高，这条边更有可能连接网络中两个不同的功能模块，所以由这条边连接的两个节点更有可能属于不同簇．然而计算网络中的边介数时间代价高，不适用于规模较大的网络．Newman、Clauset、Blondel 等人提出的 FMM[12]、CNM[13]、BGLL[14]算法是基于模块度最大化的凝聚型层次图聚类方法，迭代选择使当前聚类模块性增长的簇进行合并，直至不能再划分出模块度更高的簇结构为止．最大化模块度可能无法识别出许多实际存在的小规模簇．

2007 年，Raghavan 等人提出了基于标签传播的图聚类算法 LPA(Label Propagation Algorithm)[15]，选择节点邻域内出现次数最多的类标签作为当前节点的类标签，迭代进行标签传播，直至所有节点类标签不再变化．由于 LPA 算法在标签传播和更新过程中存在较大随机性，无法得到稳定的聚类结果．

为了更好的发现实际网络中广泛存在的小规模社区，Bader 等人提出了面向蛋白质互作用网络的复合体检测算法 MCODE[16]，该算法以节点的 k-core 度作为节点权重选择中心节点，与中心节点权重比值高于一定阈值的节点被扩展到当前簇．MCODE 算法没有对簇扩展过程进行评价，仅依赖于节点的 k-core 度，发现的通常是由完全子图合并而成的簇．DPClus[17]、IPCA[18]等算法根据待聚类节点与当前簇的连接紧密程度进行簇扩展，较好地识别网络中簇的稠密部分．实际上，网络中的簇内部连接分布是不均匀的，包含较稠密的核心区域和连接相对稀疏的边界部分．因此，在 DPClus、IPCA 等将一些规模较大的簇分成了内部连接密度不同的多个小簇．

3 背景知识

用图 $G=(V,E)$ 来表示一个复杂网络，其中节点集 $V=\{v_1,v_2,\cdots,v_n\}$ 表示网络中的个体集合且 $|V(G)|=n$，边集 $E=\{e=(v_i,v_j)\,|\,1\leq i,j\leq n\}$ 代表

网络个体间联系的集合. 本文仅对无向简单图进行讨论，将边 $e=(v_i,v_j)$ 简记为 v_iv_j.

记 $\boldsymbol{A}_{n\times n}$ 为图 G 的邻接矩阵，定义为

$$a_{ij}=\begin{cases}1, & v_iv_j\in E(G)\\ 0, & v_iv_j\notin E(G)\end{cases} \tag{4}$$

则 $\boldsymbol{A}$ 的 l 次幂 $\boldsymbol{A}^l(l\geq 1)$ 中的元素 $a_{ij}^{(l)}$ 给出了图 G 中节点 v_i 和 v_j 间长度为 l 的路径数，其中 $a_{ii}^{(l)}$ 为 G 中节点 v_i 的长度为 l 的回路总数.

令 $N_v=\{u\mid uv\in E(G)\wedge u\in V(G)\}$，称 N_v 为节点 v 在 G 中的邻域，则节点 v 的度 $k_v=|N_v|$. 假设 P 是 G 的一个非空子图，其节点集和边集分别为

$$V(P)=\{x_0,x_1,\cdots,x_l\}\subseteq V(G),$$

$$E(P)=\{x_0x_1,x_1x_2,\cdots,x_{l-1}x_l\}\subseteq E(G),$$

称 P 为图 G 中的一条长度为 l 的路径.

对图 $G=(V,E)$ 和 $G'=(V',E')$，如果 $V'\subseteq V$ 且 $E'\subseteq E$，则称 G' 是 G 的子图，记作 $G'\subseteq G$. 若 E' 是由 G 中两个节点都在 V' 中的边组成的，则称 G' 是 G 的导出子图，并记为 $G[V']$，在不引起混淆的情况下记作 $[V']$.

图 G 的节点子集 V' 在 G 中的邻域记作：

$$N_G(V')=\bigcup_{x\in V'}N_x-V' \tag{5}$$

在图 G 中，假设 $\Omega=\{V_1,V_2,\cdots,V_m\}$ 是 V 的一种划分，V_r 称为 Ω 得到的一个簇(或社区). 通常一个簇应该满足弱社区定义[19]，即社区内节点的内部度之和大于该社区节点的外部度之和，如式(6)所示.

$$\alpha\times\sum_{i\in V_r}k_i^{\text{in}}(V_r)>\sum_{i\in V_r}k_i^{\text{out}}(V_r) \tag{6}$$

其中，$k_i^{\text{in}}(V_r)=|\{v_j\mid(v_i,v_j)\in E\text{ 且 }v_j\in V_r\}|$ 为节点 v_i 与簇 V_r 内部节点的连接数，$k_i^{\text{out}}(V_r)=|\{v_j\mid(v_i,v_j)\in E\text{ 且 }v_j\notin V_r\}|$ 为节点 v_i 与簇 V_r 外部节点的连接数. 通常取 $\alpha=1$.

边 e 的介数定义为

$$BC(e)=\sum_{v_i\neq v_j\in V}\frac{g_{ij}^e}{g_{ij}} \tag{7}$$

其中，g_{ij} 为从节点 v_i 到 v_j 的最短路径的数目，g_{ij}^e 为从节点 v_i 到 v_j 的 g_{ij} 条最短路径中经过边 e 的最短路径的数目. 它反映了边 e 在网络信息传输中的控制能力，介数高的边通常连接网络中的不同社区[11].

2004 年 Newman 等人定义了模块性[20]对图聚类算法结果进行评价.

$$Q=\sum_{r=1}^{m}\left[\frac{|E(V_r)|}{M}-\left(\frac{d_r}{2M}\right)^2\right] \tag{8}$$

其中，$M=|E(G)|$，m 是簇个数，$|E(V_r)|$ 是簇 V_r 的内部边数，d_r 是簇 V_r 中所有节点的度数和. 模块度最大化的簇划分结果并不依赖于特定的网络结构，仅与簇内连边数与网络总边数的比值有关. 这导致最大化模块度可能会忽略网络中特殊的子结构，特别是内部连接较少的簇[21].

在簇扩展过程中引入合适的簇评价函数，在簇扩展过程对当前聚类结果进行监督，得到更加合理的簇划分结果. 2017 年 Bai 等人[6]基于互补熵理论提出了一种度量网络中簇发现质量的目标函数，该函数综合考虑簇内紧密程度和簇间稀疏程度对簇发现结果进行评价.

4　基于节点间局部路径的相似性度量

基于节点间连通性的相似性度量与一个图 G 的邻接矩阵 $\boldsymbol{A}$ 的 l 次幂 $\boldsymbol{A}^l(l\geq 1)$ 中的元素 $a_{ij}^{(l)}$ 密切相关，它给出了图 G 中任意节点对间所有长度为 l 的路径(或回路)数. 如 Katz 指标考虑节点间的所有路径数衡量其相似性. 随着网络规模的增大，计算两节点间的所有路径的时间代价急剧增长.

考虑到网络中的节点往往仅与相对较小的邻域范围的节点存在较强的相互作用[22]，为了平衡计算代价和准确率，通常用节点间较短的路径数来衡量节点间的相似性. LP 指标考虑节点间长度为 2 和 3 的路径数来衡量相似性，而 LS 指标则考虑了节点间长度不超过 3 的路径数来衡量相似性.

由于 $\boldsymbol{A}$ 的 l 次幂 $\boldsymbol{A}^l(l\geq 1)$ 中的元素 $a_{ij}^{(l)}$ 包含了节点 v_i 和 v_j 间长度为 l 的路径数. 如果一条路径 P 中各节点互不相同，则称路径 P 为一条点不重复路径.

对 $l=2$ 和 3，有如下两条引理成立.

引理 1.　设 v_i 和 v_j 是简单图 G 中的任意节点，其度分别为 k_{v_i} 和 k_{v_j}，则有

$$a_{ij}^{(2)}=\begin{cases}|N_{v_i}\cap N_{v_j}|, & i\neq j\\ k_{v_i}, & i=j\end{cases},$$

其中 $a_{ij}^{(2)}$ 表示节点 v_i 和 v_j 间所有长度为 2 的路径数.

证明.　令 $P=\{v_iv_x,v_xv_j\}$ 表示节点 v_i 和 v_j 间一条长度为 2 的路径. 若 $i\neq j$，由图 G 是简单图，有 $x\neq i$，$x\neq j$ 且 $v_x\in N_{v_i}\cap N_{v_j}$，因此，$a_{ij}^{(2)}=|N_{v_i}\cap N_{v_j}|$. 此时，$v_i$ 和 v_j 间不存在长度为 2 的点重复路径.

同理，对 $i=j$，由图 G 是简单图，有 $x\neq i$ 且 $v_x\in N_{v_i}$，因此，$a_{ii}^{(2)}=k_{v_i}$. 此时，存在 k_{v_i} 条长度为 2 的起点和终点重复的路径.　　证毕.

引理 2.　设 v_i 和 v_j 是简单图 G 中的任意节点，

其度分别为 k_{v_i} 和 k_{v_j}，则有

$$a_{ij}^{(3)}=\begin{cases}\tau_{ij}^{(3)}+k_{v_i}+k_{v_j}-1, & v_iv_j\in E(G)\\ \tau_{ij}^{(3)}, & v_iv_j\notin E(G)\end{cases},$$

其中 $a_{ij}^{(3)}$ 表示节点 v_i 和 v_j 间所有长度为 3 的路径数，$\tau_{ij}^{(3)}$ 表示 v_i 和 v_j 间所有长度为 3 的点不重复路径数.

证明. 分两种情形证明.

情形 1. 假设 $v_iv_j\in E(G)$，即 v_i 和 v_j 在 G 中是相邻节点. 令 $P=\{v_iv_x,v_xv_y,v_yv_j\}$ 是节点 v_i 和 v_j 间一条存在重复节点的长度为 3 的路径. 由于 G 是简单图，对任意节点 $u\in V(G)$，有边 $uu\notin E(G)$. 由此可得，$i\neq j$ 且 $i\neq x$，$x\neq y$，$y\neq j$.

此时，考虑 $x=j$ 且 $y\neq i$，$x=j$ 且 $y=i$ 和 $x\neq j$ 且 $y=i$ 三种情形.

情形 1.1. 假设 $x=j$ 且 $y\neq i$，则有 $P=\{v_iv_j,v_jv_y,v_yv_j\}$，可得 $v_y\in N_{v_j}\backslash\{v_i\}$. 所以，存在 $k_{v_j}-1$ 条此类型的路径. 如图 1(a)所示.

情形 1.2. 假设 $x=j$ 且 $y=i$，则有 $P=\{v_iv_j,v_jv_i,v_iv_j\}$，仅存在 1 条此类型的路径.

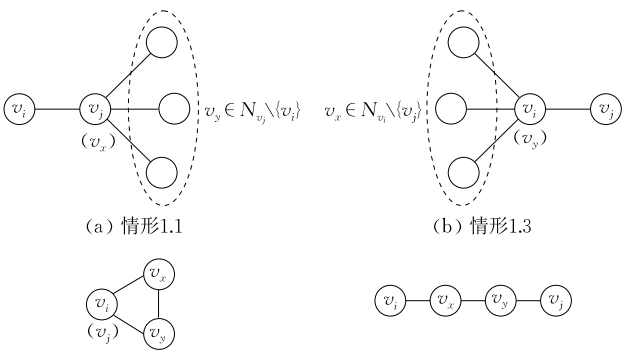
情形 1.3. 假设 $x\neq j$ 且 $y=i$，则有 $P=\{v_iv_x,v_xv_i,v_iv_j\}$，可得 $v_x\in N_{v_i}\backslash\{v_j\}$. 所以，存在 $k_{v_i}-1$ 条此类型的路径. 如图 1(b)所示.

综上可得，当 $v_iv_j\in E(G)$ 时，v_i 和 v_j 间有 $k_{v_i}+k_{v_j}-1$ 条存在点重复的长度为 3 的路径.

情形 2. 假设 $v_iv_j\notin E(G)$，即 v_i 和 v_j 在 G 中不存在连边. 考虑以下两种子情形.

情形 2.1. 假设 $i=j$，即 $P=\{v_iv_x,v_xv_y,v_yv_i\}$ 是长度为 3 的回路. 由 G 是简单图可得 $x\neq y$. 所以，当 $i=j$ 时，P 是长度为 3 的起点与终点重复的路径. 如图 1(c)所示.

情形 2.2. 假设 $i\neq j$，则有 $P=\{v_iv_x,v_xv_y,v_yv_j\}$. 由 G 是简单图可得 i,j,x,y 两两互不相等. 如图 1(d)所示，v_i 和 v_j 间不存在长度为 3 的点重复路径.



(a) 情形1.1 (b) 情形1.3

(c) 情形2.1 (d) 情形2.2

图 1 v_i 和 v_j 间的点重复路径 $P=\{v_iv_x,v_xv_y,v_yv_j\}$

所以，当 v_i 和 v_j 在 G 中不存在连边时，v_i 和 v_j 间不存在长度为 3 的点重复路径.

综合情形 1 和情形 2，可得

$$a_{ij}^{(3)}=\begin{cases}\tau_{ij}^{(3)}+k_{v_i}+k_{v_j}-1, & v_iv_j\in E(G)\\ \tau_{ij}^{(3)}, & v_iv_j\notin E(G)\end{cases},$$

其中 $a_{ij}^{(3)}$ 表示节点 v_i 和 v_j 间所有长度为 3 的路径数，$\tau_{ij}^{(3)}$ 表示 v_i 和 v_j 间所有长度为 3 的点不重复路径数. 证毕.

由引理 2 可得，节点 v_i 和 v_j 间的长度为 3 的点重复路径的数目与它们的度成正比. 节点度越大，经过该节点的点重复路径越多，这导致采用 Katz、LP、LS 等指标对节点相似性进行度量时，网络中的大度节点通常与其邻域中节点的相似性偏高. 如图 2 所示网络 G，根据式(3)有

$$\boldsymbol{S}_{\mathrm{LS}}(v_7,v_8)=4.6>\boldsymbol{S}_{\mathrm{LS}}(v_7,v_{15})=4.2.$$

在 LS 指标下，由于计算了大度节点 v_8 引起的较多点重复路径，节点 v_7 倾向于与度数较大的节点 v_8 有更高的相似性. 而实际上，边 v_7v_8 为图 G 的桥，具有较高的边介数(70). 通常由介数较高的边连接的两个节点更倾向于分属不同的簇[11].

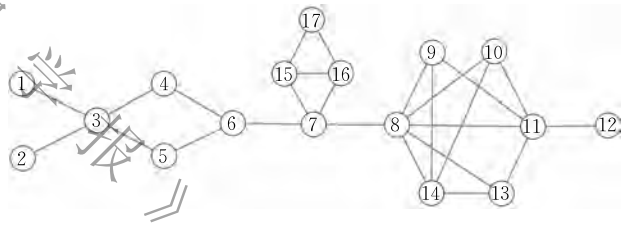


图 2 一个网络实例

为了消除由大度节点引起较多的点重复路径对节点间相似性计算的影响，本文提出基于节点间的局部点不重复路径的相似性度量指标 SLP(Similarity Metrics Based on Local Paths)定义如下.

定义 1. SLP 相似性(Similarity Metrics Based on Local Paths，SLP). 给定网络 $G=(V,E)$，顶点集为 $V=\{v_1,v_2,\cdots,v_n\}$，则网络 G 中节点 v_i 和 v_j 间的局部点不重复路径的相似性 SLP 定义为

$$\boldsymbol{S}_{\mathrm{SLP}}^{G}(v_i,v_j)=\alpha_1a_{ij}+\alpha_2\tau_{ij}^{(2)}+\alpha_3\tau_{ij}^{(3)} \tag{9}$$

其中 a_{ij} 为 G 邻接矩阵元素，$\tau_{ij}^{(l)}$ 为节点 v_i 和 v_j 间长度为 l 的点不重复路径的数目，$\alpha_1,\alpha_2,\alpha_3$ 为自由参数. 通常 $\alpha_1>\alpha_2\geq\alpha_3$ 且 $\alpha_1+\alpha_2+\alpha_3=1$. 本文称 $\boldsymbol{S}_{\mathrm{SLP}}^{G}(v_i,v_j)$ 为节点 v_i 和 v_j 的邻域相似性，且 $\boldsymbol{S}_{\mathrm{SLP}}^{G}(v_i,v_j)$ 越高，v_i 和 v_j 在局部邻域内的连通性越好.

称矩阵

$$\boldsymbol{S}_{\mathrm{SLP}}^{G}=\alpha_1\boldsymbol{A}+\alpha_2\hat{\boldsymbol{A}}^{(2)}+\alpha_3\hat{\boldsymbol{A}}^{(3)}$$

为网络 G 的 $\boldsymbol{S}_{\mathrm{SLP}}^{G}$ 矩阵，其中 $\boldsymbol{A}$ 是 G 的邻接矩阵，$\tau_{ij}^{(l)}$ 是矩阵 $\hat{\boldsymbol{A}}^{(l)}$ 中的元素. 在不引起混淆时，简记为 $\boldsymbol{S}_{\mathrm{SLP}}$.

SLP 指标消除了大度节点引起的点重复路径的影响，可以更准确的表示节点 v_i 和 v_j 间的基于连通能力的邻域相似性. 以图 2 为例，令 $\alpha_1=0.8$，$\alpha_2=0.15$，$\alpha_3=0.05$，有

$$\boldsymbol{S}_{\mathrm{SLP}}(v_7,v_{15})=1>\boldsymbol{S}_{\mathrm{SLP}}(v_7,v_8)=0.8.$$

由介数较高的边所连接的节点 v_7 和 v_8 之间相似性降低.

根据引理 1 和 2 可以得到网络的 $\boldsymbol{S}_{\mathrm{SLP}}$ 矩阵的计算方式.

设 $diag(a_{11}^{(2)},a_{22}^{(2)},\cdots,a_{nn}^{(2)})$ 表示对角元为矩阵 $\boldsymbol{A}^2$ 的对角元素的 n 阶对角阵. 由引理 1 可得

$$\hat{\boldsymbol{A}}^{(2)}=\boldsymbol{A}^2-diag(a_{11}^{(2)},a_{22}^{(2)},\cdots,a_{nn}^{(2)}) \tag{10}$$

由引理 2 可得矩阵 $\hat{\boldsymbol{A}}^{(3)}$ 的元素 $\tau_{ij}^{(3)}$，

$$\tau_{ij}^{(3)}=a_{ij}^{(3)}-a_{ij}(k_{v_i}+k_{v_j}-1) \tag{11}$$

采用图的邻接表存储结构，对节点 v_i 和 v_j，有

$\hat{\boldsymbol{A}}_{ij}^{(2)}=|N_{v_i}\cap N_{v_j}|$，

$\hat{\boldsymbol{A}}_{ij}^{(3)}=|\{(u,v)|u\in N_{v_i},v\in N_{v_j},uv\in E(G)\}|$，

因此，可在 $O(k_{v_i}\times k_{v_j})$ 时间内得到节点对 v_i 和 v_j 间的二步路径数 $\hat{\boldsymbol{A}}_{ij}^{2}$ 与三步路径数 $\hat{\boldsymbol{A}}_{ij}^{3}$. 因此计算图 G 的相似矩阵 $\boldsymbol{S}_{\mathrm{SLP}}$ 的平均时间复杂度为 $O(n^2\times\bar{k}^2)$，其中 $\bar{k}$ 是图 G 的节点平均度.

5　基于节点间路径度量的图聚类算法 PGC

基于上节给出的基于节点间局部路径的相似性度量方法，本节提出一种基于“中心-扩展”策略的图聚类算法(Local Paths Based Graph Clustering Method，PGC)，包括节点相似性计算、中心节点选择、初始簇划分和簇优化四个主要过程.

5.1　节点权重定义

首先根据所提出的 SLP 指标定义网络中的节点权重.

定义 2.　给定网络 $G=(V,E)$ 及 $\boldsymbol{S}_{\mathrm{SLP}}$ 矩阵，对节点 $v_i\in V$，定义其权重为

$$w_G(v_i)=\sum_{v_j\in V}\boldsymbol{S}_{\mathrm{SLP}}(v_i,v_j) \tag{12}$$

由于 $\boldsymbol{S}_{\mathrm{SLP}}(v_i,v_j)$ 表示节点 v_i 和 v_j 在局部邻域内的连通性，因此 $w_G(v_i)$ 表示节点 v_i 在局部邻域内的连通能力. 节点权重越高，表明其在局部邻域范围内连通能力越强，越有可能成为某个簇的中心节点.

5.2　中心节点选择及簇个数确定

考虑到两个不同簇的中心通常相距较远，不同簇中心的相似性应该较低. 因此，本文考虑节点权重及簇中心的分离性，根据式(13)计算确定节点 v_x 成为第 h 个簇中心的可能性.

$$P_h(v_x)=\begin{cases}w_G(v_x), & h=1\\ \dfrac{w_G(v_x)}{\max\limits_{j=1}^{h-1}\boldsymbol{S}_{\mathrm{SLP}}(v_x,c_j)+1}, & h>1\end{cases} \tag{13}$$

其中，c_j 表示第 j 个中心节点，对第 h 个簇，选择 $v_i=\arg\max\limits_{v_j\in V}\{P_h(v_j)\}$ 作为第 h 个初始簇中心 c_h. 首先选择权重最高的节点作为第一个初始簇中心，其余簇中心节点的选择综合考虑节点权重及与已有中心节点的邻域相似性. 节点 v_i 成为第 h 个($h>1$)簇中心节点的可能性与其权重成正相关，而与该节点和已有簇中心节点的相似性成负相关关系.

当簇个数 m 未知时，可以根据网络节点个数选择一个较大的 m 作为初始簇个数(如 $m=n/2$)，得到初始簇划分结果. 其中的簇个数通常会多于真实簇个数，在算法簇优化过程中迭代地将网络划分所得到的小簇进行合并，直到算法趋于稳定.

另一种确定初始簇个数的方法是选择使得 $P_h(c_h)$ 变化率最大的 h 作为最终簇个数 m. 如图 3 所示的 $P_h(c_h)$ 随 h 变化的曲线图中，选择 $m=2$ 作为最终的簇个数.

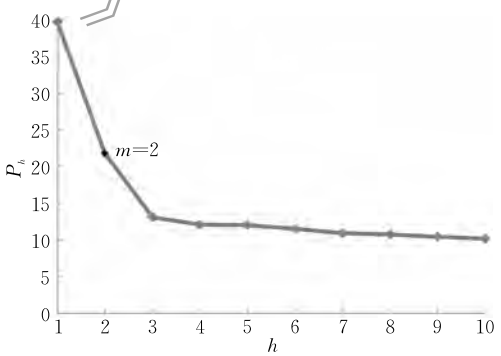


图 3　$P_h(c_h)$ 随 h 变化的曲线图

5.3　簇发现过程

选择簇中心之后，根据定义 1 给出的相似性指标 SLP，计算网络中的其它节点与 m 个簇中心的相似性，并将其分配至与其相似性最高的中心节点所在簇中，得到初始簇划分结果.

为了使算法收敛，选择基于互补熵的簇评价函数[6]作为优化目标对当前簇划分的结果 Ω 进行评价，如式(14)所示.

$$F(\Omega)=\sum_{i=1}^{n}\delta_i I_i \tag{14}$$

在式(14)中，$\delta_i=\sqrt{\sum_{h=1}^{m}\left(\frac{f_{hi}}{\sum_{j=1}^{m}f_{ji}}\right)^2}$ 对当前簇与其它簇的分离程度进行评价，其中 $f_{hi}=\frac{k_{hi}}{n_h}$，$k_{hi}$ 表示节点 v_i 与簇 V_h 之间的连边数，n_h 表示簇 V_h 中的节点个数；而 $I_i=\sum_{h=1}^{m}d_{hi}f_{hi}$ 用来评价簇内部的紧密程度.

算法迭代进行，直至 $F(\Omega)$ 趋于稳定，得到最终的图聚类结果.

5.4 算法描述

本节给出基于“中心-扩展”策略的图聚类方法 PGC. 包括节点相似性计算、中心节点选择、初始簇划分和簇优化四个主要过程，具体如算法 1 所示.

算法 1. 基于节点间路径度量的图聚类算法(PGC).

输入：网络 $G=(V,E)$

输出：图聚类结果 $\Omega=\{V_1,V_2,\cdots,V_m\}$

步骤 1. 计算 G 的节点相似性矩阵，如式(9)所示.

$$\boldsymbol{S}_{\mathrm{SLP}}^{G}(v_i,v_j)=\alpha_1 a_{ij}+\alpha_2\tau_{ij}^{(2)}+\alpha_3\tau_{ij}^{(3)}.$$

步骤 2. 根据相似性矩阵 $\boldsymbol{S}_{\mathrm{SLP}}$ 计算

$$P_h(v_x)=\begin{cases}w_G(v_x), & h=1\\ \dfrac{w_G(v_x)}{\max\limits_{j=1}^{h-1}\boldsymbol{S}_{\mathrm{SLP}}(v_x,c_j)+1}, & h>1\end{cases}$$

得到 m 个初始的中心节点，分别为 $\{c_1,c_2,\cdots,c_m\}$. 令初始簇 $V_i=\{c_i\}(1\leq i\leq m)$.

步骤 3. 对于网络中每一个非中心节点 v_x，将其分配至与其相似性最高的中心节点所在的簇 V_r 中，即 $r=\arg\max\limits_{h=1}^{m}\boldsymbol{S}_{\mathrm{SLP}}(v_x,c_h)$，得到当前簇发现结果 Ω_0. 计算当前簇发现结果的目标函数值 $F(\Omega_0)$；

步骤 4. 对当前划分结果 Ω_0 中不符合弱社区定义的簇，根据相似性矩阵 $\boldsymbol{S}_{\mathrm{SLP}}$，选择与其簇中心相似性最高的簇进行合并，得到新的簇发现结果 Ω_1. 并计算更新后的目标函数值 $F(\Omega_1)$；

步骤 5. 对于当前划分结果 Ω_1，构造每个簇 V_h 在图 G 的导出子图 $[V_h]$，并根据式(9)计算图 $[V_h]$ 的相似性矩阵 $\boldsymbol{S}_{\mathrm{SLP}}^{[V_h]}$.

步骤 6. 根据式(12)计算每个簇导出子图 $[V_h]$ 中的节点的簇内权重 $w_{[V_h]}(v_i)$，并更新簇 V_h 的中心 $c_h=\arg\max\limits_{v_i\in V_h}\{w_{[V_h]}(v_i)\}$.

步骤 7. 对于网络中每一个非中心节点 v_x，将其分配至与其相似性最高的中心节点所在的簇 V_r 中，即 $r=\arg\max\limits_{h=1}^{m}\boldsymbol{S}_{\mathrm{SLP}}(v_x,c_h)$，得到当前簇发现结果 Ω_2. 计算当前簇发现结果的目标函数值 $F(\Omega_2)$；

步骤 8. 令 $\Omega_0=\Omega_2$；

步骤 9. 若到达迭代次数或 $|F(\Omega_0)=F(\Omega_1)|$，算法结束，返回 Ω_0 作为最终簇发现结果. 否则，返回步骤 5.

算法 1 中计算相似矩阵 $\boldsymbol{S}_{\mathrm{SLP}}$ 的代价为 $O(n^2\times\bar{k}^2)$，其中 $\bar{k}$ 是图 G 的平均度. 每次迭代选择中心节点的代价为 $O(n^2)$；令 m 为簇个数，则簇扩展过程代价为 $O(nm)$. 因此算法 1 的总时间复杂度为 $O(n^2\times\bar{k}^2+t\times n^2+t\times n\times m)$，其中 t 为迭代次数. 由于复杂网络通常是稀疏的，因此网络平均度 $\bar{k}\ll n$ 且簇个数 $m\ll n$. 算法 PGC 选择权重最高的节点作为首个初始簇中心，再综合考虑节点权重及其与已有簇中心的邻域相似性依次选择其余簇中心节点，这使簇中心节点的选择更合理，进而使算法具有较快的收敛速度在实际计算中，迭代次数 t 通常不超过 5.

6 实验与结果分析

为评价算法性能，本文在 7 个真实社会网络、10 个 GN benchmark 人工网络和 12 个 LFR benchmark 人工网络[23]和 4 个蛋白质互作用网络上将本文算法与 FMM、LPA、BGLL、MCL[24]、Infomap[25]、ISCD+、Chen 等人的算法进行对比实验. 实验数据集基本情况如表 1 所示.

表 1 评测数据集基本情况

数据集	顶点数	边数	参考社区数
Karate	34	78	2
Dolphins	62	159	2
Polbooks	105	441	3
Football	115	613	12
GN Benchmark	128	1024	4
LFR Benchmark	1000	~20000	~30
Les Misérables	77	254	—
Email	1133	5451	—
Yeast	2375	11693	—
Gavin02	1352	3210	—
Gavin06	1430	6531	—
Krogan_core	2708	7123	—
DIP	4930	17201	—

6.1 评价指标

设 PGC 算法的图聚类结果为 $\Omega=\{V_1,V_2,\cdots,V_m\}$，标签数据集上的原始划分结果为 $O=\{O_1,O_2,\cdots,O_{t'}\}$. 对集合 V_i 和 $O_j(1\leq i\leq m,1\leq j\leq t')$，令 $T_{i,j}=|V_i\cap O_j|$，$b_i=\sum_{j=1}^{t'}T_{i,j}$，$s_j=\sum_{i=1}^{m}T_{i,j}$. 对有标

签数据集，本文选用标准互信息（NMI）[26]与调整兰德系数（ARI）[27]对聚类结果进行评价，定义如式(15)、(16)所示. 划分结果与原始划分的吻合程度越高，NMI 和 ARI 的值越高.

$$NMI=\frac{2\sum_{i=1}^{m}\sum_{j=1}^{t'}T_{ij}\log\frac{nT_{i,j}}{b_i s_j}}{-\sum_{i=1}^{m}b_i\log\frac{b_i}{n}-\sum_{j=1}^{t'}s_j\log\frac{s_j}{n}} \tag{15}$$

$$ARI=\frac{\sum_{i=1}^{m}\sum_{j=1}^{t'}\binom{T_{ij}}{2}-\frac{\left[\sum_{i=1}^{m}\binom{b_i}{2}\sum_{j=1}^{t'}\binom{s_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_{i=1}^{m}\binom{b_i}{2}+\sum_{j=1}^{t'}\binom{s_j}{2}\right]-\frac{\left[\sum_{i=1}^{m}\binom{b_i}{2}\sum_{j=1}^{t'}\binom{d_j}{2}\right]}{\binom{n}{2}}} \tag{16}$$

对无标签社会网络数据集，本文采用式(8)给出的模块性对算法性能进行评价. 对蛋白质互作用网络数据集，选用 F-measure 和 $Accuracy$[28]作为算法评价指标，其定义见式(17)～(20).

$$F\text{-measure}=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{17}$$

其中，

$$Precision=\frac{|\{V_i\in\Omega:\exists O_j\in O\wedge OS(V_i,O_j)\geq\theta\}|}{|\Omega|},$$

$$Recall=\frac{|\{O_j\in O:\exists C_i\in C\wedge OS(C_i,V_j)\geq\theta\}|}{|O|},$$

$$OS(V_i,O_j)=\frac{(T_{ij})^2}{|V_i|\times|O_j|},\ \theta=0.2.$$

此处，OS 称为重叠分数，度量了簇发现结果与标准数据库中复合体的符合程度.

敏感度 Sn 表示识别的在标准复合物中覆盖蛋白质的多少，如式(18)所示.

$$Sn=\frac{\sum_{i=1}^{m}\max_j\{T_{ij}\}}{\sum_{i=1}^{m}|O_i|} \tag{18}$$

真阳性预测值 PPV 表示识别的蛋白质复合物成为真阳性的可能性，如式(19)所示.

$$PPV=\frac{\sum_{j=1}^{t'}\max_i\{T_{ij}\}}{\sum_{j=1}^{t'}b_j} \tag{19}$$

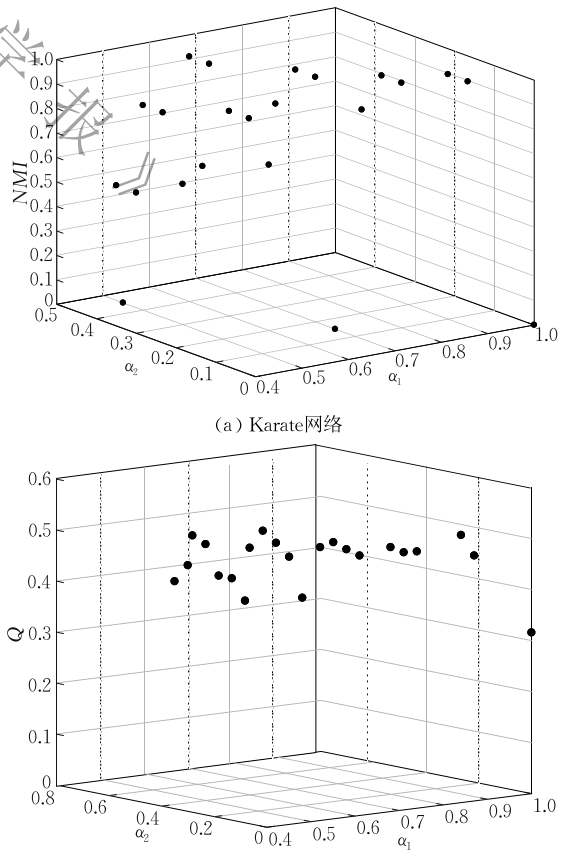
$Accuracy$ 由敏感度 Sn 和真阳性预测值 PPV 构成，如式(20)所示.

$$Accuracy=\sqrt{Sn\times PPV} \tag{20}$$

6.2　参数设置

节点相似性度量在图聚类算法起着重要作用，然而，由于算法 PGC 在聚类过程中对聚簇中心迭代的进行动态调整，因此算法所得到的网络聚类结果对式(9)中涉及的参数 α_1，α_2 和 α_3 敏感度较低. 为进一步验证算法对参数的依赖性，本节在表 1 所列的 11 个真实网络及 22 个人工网络上测试算法 PGC 对参数的敏感性，在带标签网络上以 NMI 为优化目标，无标签网络上以模块性 Q 为优化目标. 由于 $\alpha_1>\alpha_2\geq\alpha_3\geq0$，此处选择 α_1 的变化范围是[0.4，1.0]，α_2 的变化范围是[0，0.45]，$\alpha_3=1-\alpha_1-\alpha_2$.

实验表明，当 $\alpha_1\in[0.6,0.8]$，$\alpha_2\in[0.15,0.35]$，且 $\alpha_3=1-\alpha_1-\alpha_2>0$ 时，算法 PGC 在大多数情况下会达到最优值. 图 4 给出了 Karate 网络[29]（带标签）和 Les Misérables 网络[30]（无标签）上优化目标随参数 α_1 和 α_2 取值变化的散点图. 算法 PGC 在 Karate 网络上，当 $\alpha_1\in[0.8,0.9]$，$\alpha_2\in[0.05,0.15]$ 时取得最高的 NMI 值，此时 $NMI=1$. 在 $\alpha_1\in[0.6,0.8]$ 且 $\alpha_2\in[0.15,0.35]$ 上所获得的图聚类结果 NMI 相差很少（±0.091）.



(a) Karate网络

(b) Les Misérables网络

图 4　参数对算法 PGC 的影响

对 Les Misérables 网络，本文算法在 $\alpha_1=0.6$，$\alpha_2=0.35$ 取得最高的模块性 Q. 然而在 $\alpha_1\in[0.6,0.8]$且 $\alpha_2\in[0.15,0.35]$上所获得的图聚类结果模块性相差很少(±0.0021).

因此，在后续实验结果比较中，对带标签网络本文取 $\alpha_1=0.8$，$\alpha_2=0.15$，$\alpha_3=0.05$；对无标签网络取 $\alpha_1=0.6$，$\alpha_2=0.35$，$\alpha_3=0.05$.

6.3 真实网络实验结果

在本小节中，我们通过在空手道俱乐部(Zachary's Karate Club)[29]、海豚社交网络(Dolphins Social Network)[31]、Polbooks[6]和大学生足球网络(American College Football Network)[32]四个带标签的真实网络以及 Les Misérables、Email[33]、Yeast[34]三个无标签的真实网络上进行实验来对本文所提的 SLP 节点局部路径相似性度量方法以及基于 SLP 的图聚类算法 PGC 进行评测.

首先在 7 个真实数据集上对本文提出的 SLP 节点相似度量的有效性进行测试，实验结果如表 2 所示. 本文算法 PGC 采用基于局部点不重复路径的 SLP 节点相似性度量方法，将本文算法中的节点相似性度量方法改为基于局部路径的 LS 节点相似性度量方法得到对比算法 LS-PGC. 可以看出，在无标签和带标签的真实网络上，本文所给的 SLP 节点相似性度量在各项指标上都取得了优于对比算法的实验结果.

表 2 节点相似性度量实验结果

	Data set														
	Karate			Dolphins			Polbooks			Football			Les Misérables	Email	Yeast
Index	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*	*Q*	*Q*	*Q*
LS-PGC	2	1.0000	1.0000	4	0.4771	0.5686	3	0.6599	0.5416	14	0.8064	0.8872	0.4531	0.3978	0.6139
PGC	2	1.0000	1.0000	2	0.9348	0.8888	2	0.6671	0.5979	11	0.8653	0.9151	0.5224	0.4838	0.6447

将 PGC 算法与经典图聚类算法 FMM、LPA、BGLL、MCL、Infomap、ISCD+和 Chen 等人的算法进行比较实验. 所有实验都取算法在各网络上分别运行 30 次所得聚类结果的平均性能，结果如表 3 所示，其中 m 表示算法最终确定的簇个数. 由于 LPA 和 BGLL 算法结果不稳定，多次运行结果中簇个数相差较大，此处用“—”表示. 可以看出，本文算法 PGC 在大多数情况下性能均优于比较算法. 在 Polbooks 网络上 PGC 得到了 2 个簇，而真实网络包含 3 个簇. 实际上，Polbooks 网络节点表示在 Amazon 在线书店上销售的与美国政治相关的图书，若两本图书被同一用户购买过，则对应节点间存在一条边. 这些图书被分为 3 类：“自由派”、“保守派”和“中间派”，如图 5 所示，菱形点所代表的图书为“保守派”，圆点所代表的图书为“自由派”，正方形的点所代表的图书即为“中间派”. 可以看到，3 类图书中“自由派”和“保守派”这两个簇内部连接比较紧密，簇间连边比较稀疏. 而对于“中间派”，并不具有明显的簇结构. 因此在 Polbooks 网络上划分 2 个簇更加合理，属于“中间派”的节点应该被确定为未聚类节点.

在 3 个无标签数据集 Les Misérables、Email 和 Yeast 上，算法 PGC 在模块性方面与其它算法进行比较，结果见表 4，可以看出 PGC 算法的聚类结果体现了较高的模块性.

表 3 带标签真实网络实验结果对比表

	Data set											
	Karate			Dolphins			Polbooks			Football		
Index	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*	*m*	*ARI*	*NMI*
FMM	2	0.8823	0.8372	3	0.4795	0.6058	3	0.6563	0.5566	5	0.4444	0.6862
LPA	—	0.6632	0.6574	—	0.5106	0.6349	—	0.4921	0.4231	—	0.7390	0.8725
BGLL	—	0.5445	0.6537	—	0.4259	0.5195	—	0.5966	0.5234	—	0.7681	0.8740
MCL	2	0.8823	0.8365	15	0.1447	0.3628	6	0.5884	0.5374	12	0.8967	0.9242
Infomap	3	0.7022	0.6995	6	0.3614	0.5270	5	0.6463	0.5369	10	0.7601	0.8801
ISCD+	2	1.0000	1.0000	4	0.3458	0.4708	3	0.6390	0.5245	13	0.8868	0.9254
Chen	2	1.0000	1.0000	2	0.9348	0.8888	2	0.6671	0.5979	10	0.8154	0.8874
PGC	2	1.0000	1.0000	2	0.9348	0.8888	2	0.6671	0.5979	11	0.8653	0.9151

表 4 无标签真实网络各算法模块性对比

Data set	Index	FMM	LPA	BGLL	MCL	Infomap	ISCD+	Chen	PGC
Les Misérables	*Q*	0.4961	0.2719	0.5461	0.2875	0.5363	0.2540	0.4037	0.5224
Email	*Q*	0.3461	0.0002	0.5522	0.1430	0.5309	0.4563	0.0000	0.4838
Yeast	*Q*	0.7021	0.6639	0.7291	0.4337	0.5269	0.5192	0.5238	0.6447

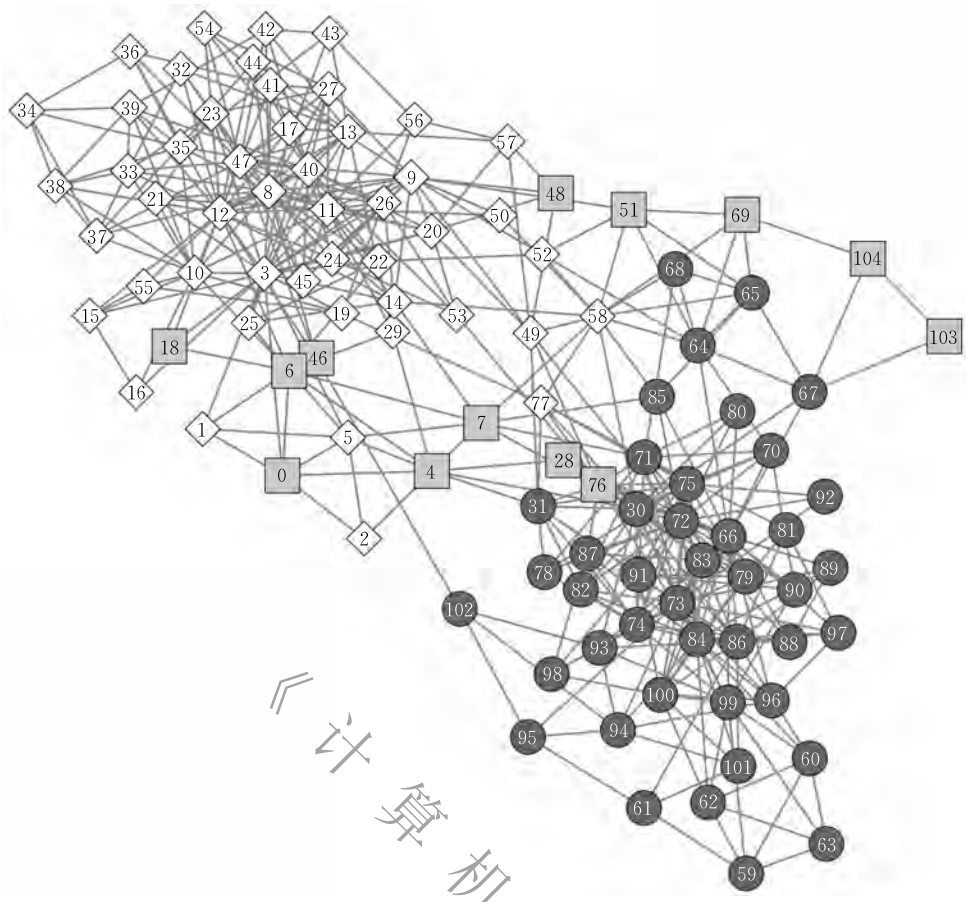


图 5　Polbooks 网络的原始划分

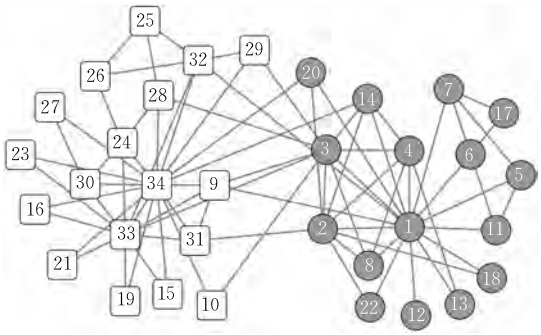
图 6 给出了算法 FMM 和 ISCD+以及本文算法 PGC 在跆拳道网络 Karate 上的最终聚类结果. 由于采用了合理的目标函数，本文算法和 ISCD+取得了与原始网络一致的聚类结果. 而由于 FMM 算法以模块性为目标函数，使得本应属于大簇的节点 10 划分到了小簇中.

图 7 给出了算法 FMM 和 ISCD+以及本文算法 PGC 在海豚社交网络 Dolphins 上的聚类结果. 由于 ISCD+算法采用公共邻居数对节点相似性进行度量，在选择中心节点的时候只考虑节点的度信息，而没有考虑该节点在更大的邻域内的拓扑信息，因此最终得到 4 个簇，大于实际网络的簇个数. 本文算法 PGC 依据节点对间不超过 3 的点不重复路径数定义相似性，考虑了更多的节点局部邻域信息，得到了与原始网络最为一致的划分.

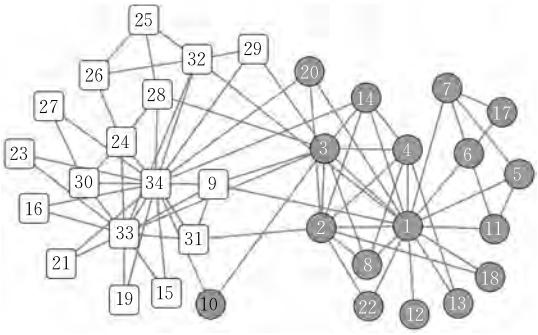
图 8 给出了算法 FMM 和 ISCD+以及本文算法 PGC 在 Polbooks 网络上的聚类结果. 根据之前的分析，在 Polbooks 网络上，有 2 个具有明显簇结构的类别，即“自由派”和“保守派”. 尽管“中间派”在原始网络作为一个簇存在，实际上它并不具有明显的簇结构，簇内连接并不紧密. 本文算法体现了原始网络的这一特征，最终簇个数确定为 2. 与图 5 给出的原始网络划分结果相比较，显然本文算法给出的结果更合理.

Football 数据集是根据美国大学足球联赛在 2000 年一个赛季赛程建立的比赛网络，联赛共 12 个联盟，每个球队属于且仅属于其中一个联盟. 网络中节点代表球队，若两球队之间进行过比赛，则对应节点间存在一条连边，一个联盟中的所有球队看作一个簇，因此，Football 网络中共存在 12 个簇.

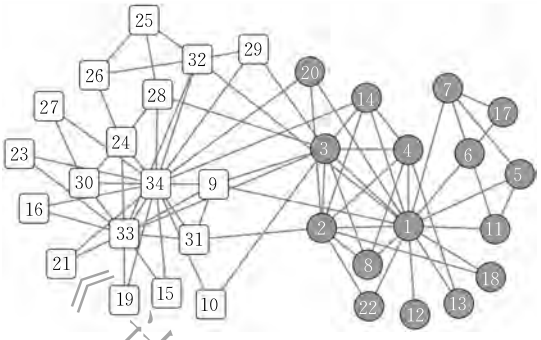
从图 9 中可以看出，原始网络中由节点{36，42，80，82，90}构成的簇没有明显的簇结构，这是由于该联盟中球队之间地理位置距离较远，联盟内部比赛很少，与该联盟比赛较多的都是与其地理距离较近的球队，因此该簇内部连边比较稀疏，导致从拓扑结构方面考虑很难被划分在一个簇. Newman 等人在文献[11]中将该簇中的 4 个节点分别划分为孤立点，最终划分了 17 个簇. 而本文算法 PGC 在 Football 网络上得到了 11 个簇，使得每个簇内部的连接比簇间的连接更加稠密，图 9(b)给出了本文算法 PGC 在 Football 网络上的划分结果.



(a) PGC算法

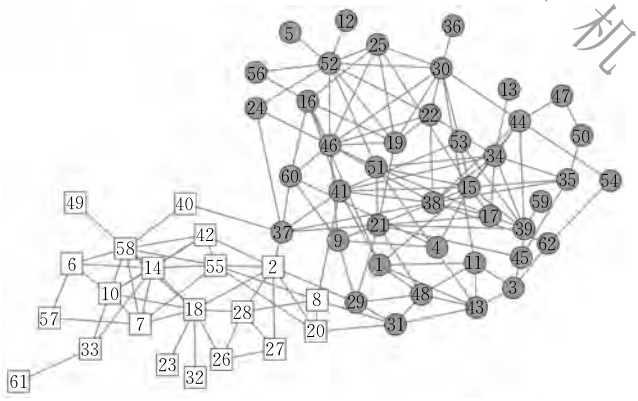


(b) FMM算法

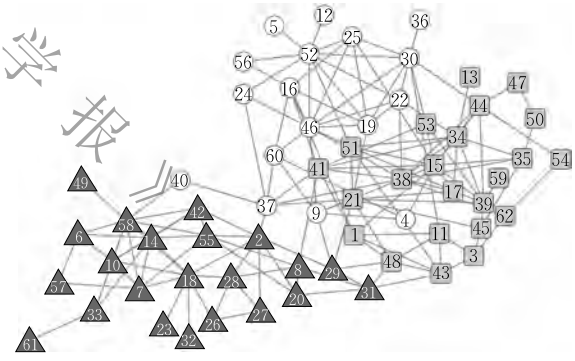


(c) ISCD+算法

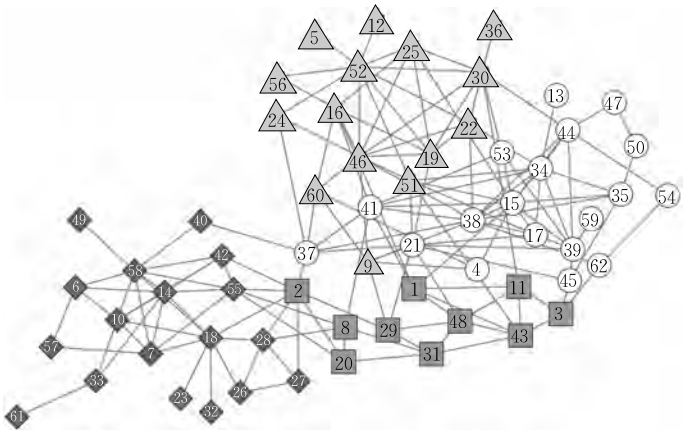
图 6 在 Karate 网络上的聚类结果比较



(a) PGC算法



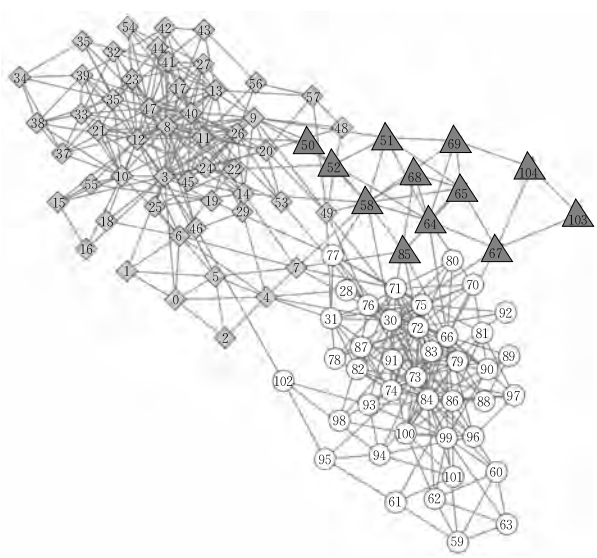
(b) FMM算法



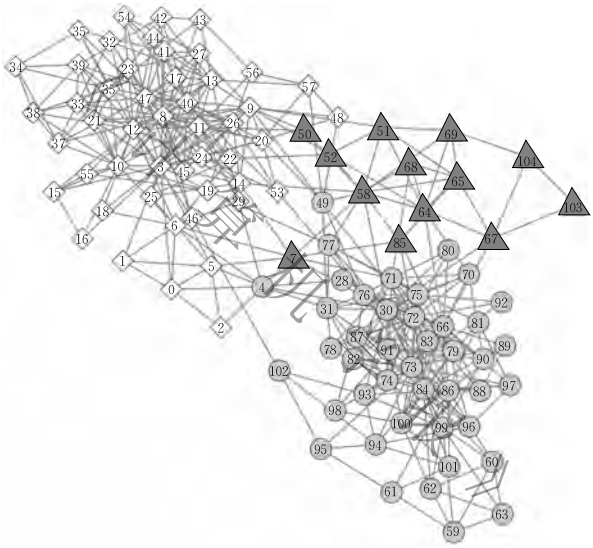
(c) ISCD+算法

图 7 在 Dolphins 网络上的聚类结果比较

(a) PGC算法

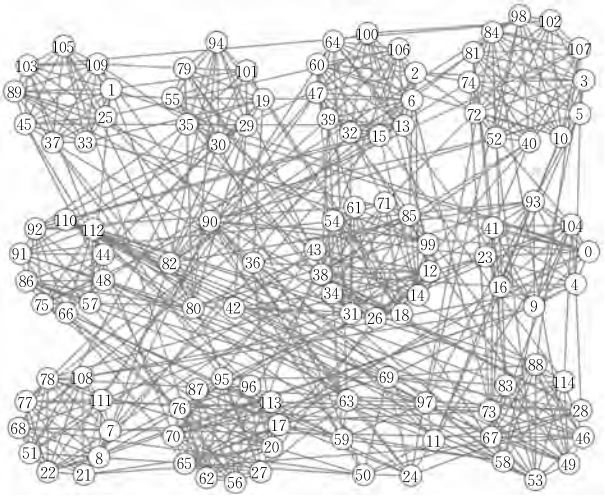


(b) FMM算法

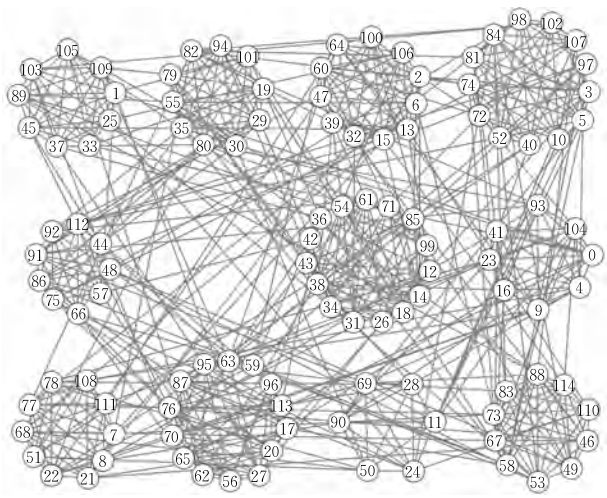


(c) ISCD+算法

图 8　在 Polbooks 网络上的聚类结果比较



(a) Football网络原始划分

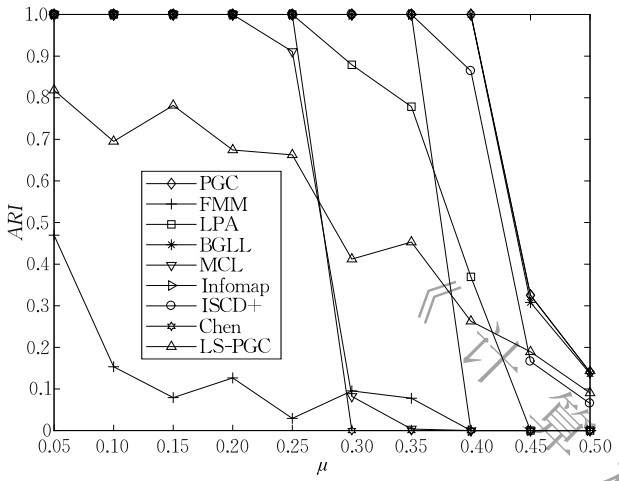


(b) PGC算法聚类结果

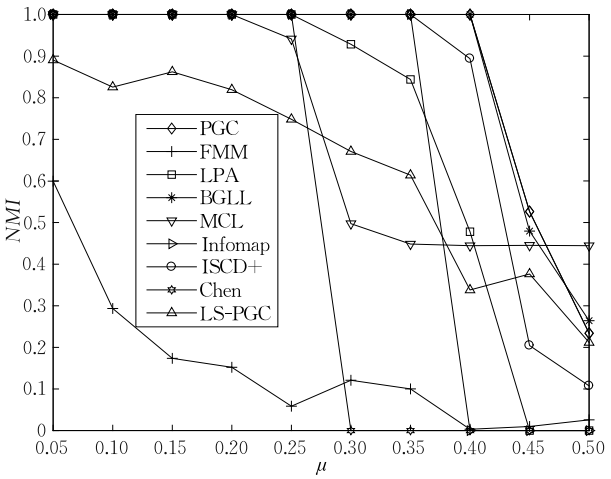
图 9　PGC 算法在 Football 网络上的聚类结果比较

6.4 人工网络实验结果

本节分别构造了不同混合参数的 10 个 GN benchmark 人工网络和 12 个 LFR benchmark 人工网络. GN benchmark 网络包含 128 个节点,平均分到 4 个簇中,节点度均为 16,每个节点簇内连边数相同. 混合参数 μ 是网络中簇间连边占总边数的比例. 网络簇结构越明显,混合参数值 μ 越小. 因此簇发现的难度随 μ 值增加而增大. LFR 网络包括 1000 个节点,平均度为 20,最大度为 50,社区规模区间为 20～50,节点度序列满足指数为 2 的幂律分布,社区规模满足指数为 1 的幂律分布. 在 GN Benchmark 上混合参数 μ 为 0.05～0.5,LFR Benchmark 上,混合参数 μ 为 0.05～0.6. 取各算法在各网络上执行 30 次的结果进行比较,比较结果如图 10 和图 11 所示. 可以看到,本文算法在各混合参数下得到了比较好的图聚类结果,特别是在混合参数进一步增大,簇结构不明显的网络中,PGC 算法依然能够保持较好的图聚类结果.

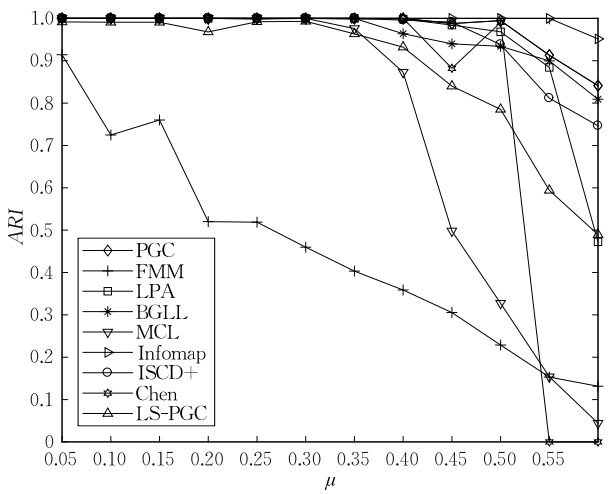


(a) 各算法*ARI*对比结果

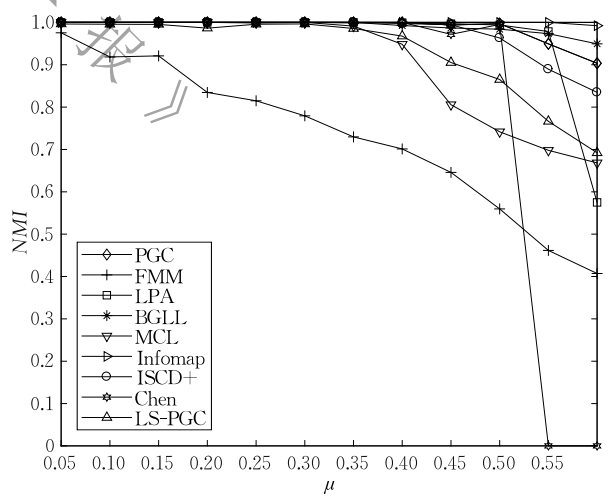


(b) 各算法*NMI*对比结果

图 10 在 GN benchmark 网络上的对比结果



(a) 各算法*ARI*对比结果



(b) 各算法*NMI*对比结果

图 11 在 LFR benchmark 网络上的对比结果

6.5 蛋白质互作用网络数据实验结果

本小节将本文算法 PGC 应用于 Gavin02[35]、Gavin06[36]、Krogan_core[37]和 DIP[38]4 个蛋白质互作用网络上进行蛋白质复合体发现,这些网络的基本信息见表 1. 采用 CYC2008 数据集作为标准复合物对照集[39]. 采用经典的非重叠图聚类算法 FMM、BGLL、MCL、Infomap 和 5 个经典蛋白质复合物发现算法 MCODE、HC_PIN[40]、DPClus、IPCA、ClusterOne[41]作为对比算法,若算法包含参数则选择原作者给定的最优参数值. 采用 *F*-measure 和 *Accuracy* 作为评价指标,定义见式(17)～(20). 算法比较结果见表 5.

表 5　在蛋白质互作用网络数据上算法比较结果

	Data set							
	Gavin02		Gavin06		Krogan_core		DIP	
Index	*F*-measure	*Accuracy*	*F*-measure	*Accuracy*	*F*-measure	*Accuracy*	*F*-measure	*Accuracy*
FMM	0.1708	0.3821	0.1474	0.3993	0.1506	0.4054	0.0170	0.3044
BGLL	0.1652	0.3751	0.1326	0.4027	0.0910	0.3909	0.0131	0.3206
MCL	0.2717	0.4465	0.3448	0.5137	0.3151	0.5162	0.2055	0.4890
Infomap	0.3548	0.4996	0.0913	0.2538	0.3779	0.5823	0.2546	0.5760
PGC	**0.3628**	**0.4795**	**0.3713**	**0.5224**	**0.3889**	**0.5535**	**0.2793**	**0.5494**
MCODE	0.1192	0.2489	0.2497	0.3768	0.2762	0.3548	0.1601	0.2938
HC_PIN	0.2354	0.3756	0.3474	0.4334	0.3043	0.4895	0.0263	0.1636
DPClus	0.3366	0.3797	0.4299	0.4950	0.4438	0.5717	0.3267	0.5504
IPCA	0.4004	0.4404	0.4478	0.5154	0.4430	0.5807	0.3160	0.5260
ClusterOne	0.2708	0.3699	0.3935	0.5242	0.4192	0.5188	0.3712	0.4916

可以看出，本文算法 PGC 与 4 个非重叠图聚类算法相比效果较好，但略逊于 DPClus、IPCA、ClusterOne 等经典蛋白质复合体发现算法. 这是由于蛋白质复合体发现算法可以进行重叠图聚类，能够发现蛋白质互作用网络中存在的大量具有重叠蛋白质的复合体.

7　总结与展望

本文对复杂网络的图聚类算法进行研究，给出了一种基于节点间局部路径连通性的图聚类算法 PGC，包括节点相似性计算、中心节点选择、初始簇划分和簇优化四个主要过程. 算法通过计算节点间的局部点不重复路径定义节点间相似性指标 SLP，根据 $\boldsymbol{S}_{SLP}$ 矩阵定义节点权重，进而选择簇中心；将其余节点分配至与其相似性最高的中心节点所在簇中；以基于互补熵的簇质量评价函数作为目标函数，对簇划分结果进行迭代优化，得到最终簇发现结果. 采用本文所提出的 SLP 指标对节点相似性进行度量，可以消除由大度节点引起较多的点重复路径对节点相似性的影响，从而提高算法对大度节点邻域中节点的划分能力. 在 11 个真实网络和 22 个人工网络上的实验结果表明算法 PGC 在标准互信息、调整兰德系数、*F*-measure 等方面均表现出良好的性能.

本文基于节点间局部路径相似性的图聚类算法 PGC 从拓扑结构角度对社会网络、蛋白质互作用网络中的簇进行发现. 现实世界的复杂网络往往根据应用领域的基本信息在动态演化，如何结合应用领域数据在动态变化的复杂网络上进行簇发现，需要我们进一步探索研究. 针对生物网络，如何结合多源网络信息，将基因调控网络、基因-疾病网络、生物代谢网络和蛋白质互作用网络信息进行融合，挖掘更有价值的功能信息也是值得进一步思考的课题.

参　考　文　献

[1] Rolland T, Taşan M, Charloteaux B, et al. A proteome-scale map of the human interactome network. Cell, 2014, 159(5): 1212-1226

[2] Yang Bo, Liu Ji-Ming, Feng Jian-Feng. On the spectral characterization and scalable mining of network communities. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(2): 326-337

[3] Jia Song-Wei, Gao Lin, Gao Yong, et al. Defining and identifying cograph communities in complex networks. New Journal of Physics, 2015, 17(1): 013044

[4] Li Hui-Jia, Li Hui-Ying, Li Ai-Hua. Analysis of multi-scale stability in community structure. Chinese Journal of Computers, 2015, 38(2): 301-312(in Chinese)
(李慧嘉，李慧颖，李爱华. 多尺度的社团结构稳定性分析. 计算机学报，2015，38(2)：301-312)

[5] Jia Song-Wei, Gao Lin, Gao Yong, et al. Exploring triad-rich substructures by graph-theoretic characterizations in complex networks. Physica A: Statistical Mechanics and Its Applications, 2017, 468: 53-69

[6] Bai Liang, Cheng Xue-Qi, Liang Ji-Ye, et al. Fast graph clustering with a new description model for community detection. Information Sciences, 2017, 388-389: 37-47

[7] Chen Zeng-Qiang, Xie Zheng, Zhang Qing. Community detection based on local topological information and its application in power grid. Neurocomputing, 2015, 170: 384-392

[8] Wang Jie, Liang Ji-Ye, Zheng Wen-Ping. A graph clustering method for detecting protein complexes. Journal of Computer Research and Development, 2015, 52(8): 1784-1793(in Chinese)
(王杰，梁吉业，郑文萍. 一种面向蛋白质复合体检测的图聚类方法. 计算机研究与发展，2015，52(8)：1784-1973)

[9] Wang Xiao-Fan, Li Xiang, Chen Guan-Rong. Network Science: An Introduction. Beijing: Higher Education Press, 2012(in Chinese)

(汪小帆，李翔，陈关荣. 网络科学导论. 北京：高等教育出版社，2012)

[10] Zhou Tao, Lü Lin-Yuan, Zhang Yi-Cheng. Predicting missing links via local information. The European Physical Journal B, 2009, 71(4): 623-630

[11] 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

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

[13] Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. Physical Review E, 2004, 70(6): 066111

[14] Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008

[15] 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

[16] Bader G D, Hogue C W. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 2003, 4(1): 2

[17] Altaf-Ul-Amin M, Shinbo Y, Mihara K, et al. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics, 2006, 7(1): 207

[18] Li Min, Chen Jian-Er, Wang Jian-Xin, et al. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinformatics, 2008, 9(1): 398

[19] Radicchi F, Castellano C, Cecconi F, et al. 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] Newman M E J, Girvan M. Finding and evaluating community structure in networks. Physical Review E, 2004, 69(2): 026113

[21] Fortunato S, Barthélemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1): 36-41

[22] Leskovec J, Lang K J, Dasgupta A, et al. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. Internet Mathematics, 2009, 6(1): 29-123

[23] Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. Physical Review E, 2008, 78(4): 046110

[24] Enright A J, Van Dongen S, Ouzounis C A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research, 2002, 30(7): 1575-1584

[25] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(4): 1118-1123

[26] Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment, 2005, 2005, (9): P09008

[27] Rand W M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, 66(336): 846-850

[28] Wang Jie, Zheng Wen-Ping, Qian Yu-Hua, et al. A seed expansion graph clustering method for protein complexes detection in protein interaction networks. Molecules, 2017, 22(12): 2179

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

[30] Yang Gui, Zheng Wen-Ping, Wang Wen-Jian, et al. Community detection algorithm based on weighted dense subgraphs. Journal of Software, 2017, 28(11): 3103-3114(in Chinese)
(杨贵，郑文萍，王文剑等. 一种加权稠密子图社区发现算法. 软件学报，2017，28(11)：3103-3114)

[31] Lusseau D, Newman M E J. Identifying the role that animals play in their social networks. Proceedings of the Royal Society B: Biological Sciences, 2004, 271(Suppl. 6): S477-S481

[32] 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

[33] Guimerà R, Danon L, Díaz-Guilera A, et al. Self-similar community structure in a network of human interactions. Physical Review E, 2003, 68(6): 065103

[34] Qian Yu-Hua, Li Ye-Bin, Zhang Min, et al. Quantifying edge significance on maintaining global connectivity. Scientific Reports, 2017, 7: 45380

[35] Gavin A, Bösche M, Krause R, et al. Functional organization of the Yeast proteome by systematic analysis of protein complexes. Nature, 2002, 415(6868): 141-147

[36] Gavin A, Aloy P, Grandi P, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature, 2006, 440(7084): 631-636

[37] Krogan N J, Cagney G, Yu Hai-Yuan, et al. Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature, 2006, 440(7084): 637-643

[38] Xenarios I, Salwiński Ł, Duan X J, et al. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Research, 2002, 30(1): 303-305

[39] Pu Shu-Ye, Wong J, Turner B, et al. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Research, 2009, 37(3): 825-831

[40] Wang Jian-Xin, Li Min, Chen Jian-Er, et al. A fast hierarchical clustering algorithm for functional modules discovery in protein Interaction networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(3): 607-620

[41] Nepusz T, Yu Hai-Yuan, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nature Methods, 2012, 9(5): 471-472

ZHENG Wen-Ping, Ph. D., associate professor. Her research interests include graph theory algorithms, bioinformatics.

CHE Chen-Hao, M. S. candidate. His research interest is clustering algorithm.

QIAN Yu-Hua, Ph. D., professor. His research interests include artificial intelligence, machine learning.

WANG Jie, Ph. D. candidate. His research interests include data mining, bioinformatics.

YANG Gui, Ph. D. candidate. His research interest is bioinformatics.

Background

Graph clustering is to group the vertices of the network into clusters taking into consideration edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters. Various graph clustering algorithms have been developed to identify clusters using the information encoded in the network topology. In general, these methods can be classified into two types: global method and local method according to whether they produce clusters based on whole view or partial view of graph topology.

Global approaches exploit the global structure information of networks, such as GN algorithm proposed by Girvan and Newman, Markov clustering algorithm, spectral clustering method. Local clustering methods identify protein complexes by considering local neighbor information in complex networks. To improve computational efficiency, local methods always start by selecting a highly ranked node as seed and then expand the seed to a densely connected group of nodes relying on a local benefit function. MCODE, HC_PIN, DPClus and IPCA etc. are excellent up to date local graph clustering algorithms.

Node similarity metrics and clustering quality evaluation are two key points for graph clustering algorithms. We propose a graph clustering algorithm based on local paths between nodes, abbreviated, PGC. In order to identify bridges in networks, we use local paths, instead of chains, between nodes are used to measure node similarity in algorithm PGC. Besides, we adopt a clustering quality evaluation based on complementary entropy to obtain more accurate clustering output. Compared with other classical graph clustering algorithms on 11 real networks and 22 artificial networks, the proposed algorithm PGC shows a preferable performance.

This research was supported by the Research Project Supported by the Shanxi Nature Science Foundation under Grant No. 201801D121123 and the Shanxi Scholarship Council of China under Grant No. 2017-014. The research was also supported by the National Nature Science Foundation under Grant No. 61572005. The first two foundations aim to study on algorithms in artificial and machine learning to obtain useful information from large scale datasets, such as biological network data set, financial data, and so on. The last foundation aims to study graph algorithms. The authors have made some in-depth researches on computational methods in artificial intelligence, similarity metrics for graph clustering, the graph algorithm design in classical graph hard problems, and complex network modeling problems.