计算机科学
COMPUTER SCIENCE

一种基于局部路径信息的重叠社区发现算法

郑文萍, 王宁, 杨贵

引用本文

郑文萍, 王宁, 杨贵.一种基于局部路径信息的重叠社区发现算法[J]. 计算机科学, 2022, 49(12): 155-162.

ZHENG Wen-ping, WANG Ning, YANG Gui. Overlapping Community Detection Algorithm Based on Local Path Information [J]. Computer Science, 2022, 49(12): 155-162.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220500190

一种基于局部路径信息的重叠社区发现算法

郑文萍[1,2,3]　王　宁[1]　杨　贵[1]
1 山西大学计算机与信息技术学院　太原 030006
2 计算智能与中文信息处理教育部重点实验室(山西大学)　太原 030006
3 智能信息处理研究所(山西大学)　太原 030006

摘　要　重叠社区发现是复杂网络分析的主要任务之一。针对现有的基于局部扩展和优化的重叠社区发现方法受初始种子节点选择影响较大、适应度函数无法度量节点间多样的连接方式等问题,提出了一种基于局部路径信息的重叠社区发现算法(Local Path Information-based Overlapping Community Detection Algorithm,LPIO)。首先选取局部极大度点作为初始种子节点,并根据社区内节点邻域标签一致性更新社区的种子节点集,避免初始种子节点对算法性能的影响;然后为度量稀疏网络中节点间多样的连接方式,给出了基于局部路径信息的社区适应度函数,扩展种子节点集得到社区结构;最后计算未聚类节点与社区种子集之间的点不重复路径数量,得到未聚类节点与已有社区间的距离,为未聚类节点分配社区。在4个有标签网络和8个无标签网络上,与7个经典重叠社区发现算法进行对比,实验结果表明,所提算法在重叠标准互信息(*ONMI*)、F_1分数、扩展模块度(*EQ*)等方面表现良好。

关键词:重叠社区发现;局部扩展和优化;社区适应度;局部路径信息
中图法分类号　TP391

Overlapping Community Detection Algorithm Based on Local Path Information

ZHENG Wen-ping[1,2,3],WANG Ning[1] and YANG Gui[1]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education,Shanxi University,Taiyuan 030006,China
3 Institute of Intelligent Information Processing,Shanxi University,Taiyuan 030006,China

Abstract　The detection of overlapping communities is one of the main tasks of complex network analysis. The performance of most existing methods based on local expansion and optimization are greatly affected by the selection of initial seed nodes and the community structure significance measurement. Aiming at these problems,an overlapping community detection algorithm is proposed based on local path information(LPIO). First,the local maximum degree points are selected as initial seeds,which will be updated according to the label consistency of node's neighborhood in the community to reduce the influence of the selection of initial seeds. To measure the various connection patterns between nodes in networks,a community fitness function is defined based on local path information to obtain community structures from seed nodes. Finally,unclustered nodes are assigned to proper communities according to the number of non-repetitive paths between the unclustered nodes and the community seed sets. Comparative experiments on 4 labeled networks and 8 unlabeled networks with 7 classic overlapping community detection algorithms show that the proposed algorithm performs well on overlapping standard mutual information(*ONMI*),F_1 score,and extended modularity(*EQ*).

Keywords　Overlapping community detection,Local expansion and optimization,Community fitness,Local path information

现实世界中许多复杂系统都可以抽象成网络,如通信网络、电力网络、经济网络和社交网络等[1]。社区结构是复杂网络的重要特征,即网络中存在若干个社区,社区内部节点间连接相对稠密,社区之间连接相对稀疏。不同社区间通常会有一些公共节点,这样的社区结构被称为重叠社区结构,社区间的公共节点被称为重叠节点[1]。重叠节点通常对应网络中有多种功能的实体,如生物网络中具有多种生物功能的基因、社交网络中属于多个社会群体的人员等。发现复杂网络中的重叠社区结构有助于研究人员探索复杂系统实体间丰富的相互作用模式,准确理解系统的组织原则。

大多数基于局部扩展和优化策略的社区发现方法根据社区内节点间的直接连边密度或三角形密度来评价社区的内部

到稿日期:2022-05-19　返修日期:2022-08-27
基金项目:国家自然科学基金(62072292);山西省1331工程项目
This work was supported by the National Natural Science Foundation of China(62072292) and 1331 Engineering Project of Shanxi Province,China.
通信作者:郑文萍(wpzheng@sxu.edu.cn)

连接紧密程度，进而发现连接相对稠密的社区结构。然而，当社区内部连接相对稀疏时，无法形成较多的三角形连接，利用连边密度或三角形密度无法有效度量社区内节点间多样的连接方式。针对此问题，本文提出了一种基于局部路径信息的重叠社区发现算法，利用社区内相邻节点间的路径连通信息给出了基于局部路径支持度的社区适应度评价函数，对种子节点集进行扩展得到内部路径连通性较强的社区结构。在社区发现过程中，根据节点邻域的社区标签的一致性动态更新社区种子节点，以消除初始种子节点选择对算法性能的影响。在4个有标签网络和8个无标签网络上，与7个经典重叠社区发现算法的比较结果表明，LPIO能较好地发现网络路径连通性较好的社区结构。

1 相关工作

重叠社区发现已成为图聚类问题的重要研究热点，目前已经提出了一些成功的重叠社区发现算法，如基于派系过滤的、基于标签传播的以及基于局部扩展和优化的方法等[1]。

最早的重叠社区发现算法是由 Pallal 等于 2005 年提出的派系过滤算法 CPM[2]，通过合并网络中重叠度高的 K-团来完成社区发现，社区间的公共节点即为重叠节点。2018 年 Jabbour 等[3]提出用弦图快速寻找网络中的最大团，以导电性为社区适应度函数，扩展由最大团构成的初始社区。基于派系过滤的方法假设社区内节点间形成很多三角形连接，这一假设不适用于识别内部连接比较稀疏的社区结构。同时，寻找网络中的最大团是 NP 困难问题，因此基于派系过滤的方法仅适用于发现小规模网络中连接稠密的社区结构。

2010 年，Gregory 提出了 COPRA[4]算法，利用标签传播算法进行重叠社区发现，根据邻居节点中社区标签的出现频率定义节点的社区归属度，将出现频率大于给定阈值的标签作为节点的社区标签。COPRA 算法有接近线性的时间复杂度，在实际中得到了广泛的应用。然而，当有多个标签出现频率最高的社区时，COPRA 算法会为节点随机分配社区，这导致算法运行结果的稳定性较低；同时，COPRA 假设所有邻居节点对待确定标签节点的社区归属影响是相同的，这也导致社区发现过程忽略了重要邻居对节点社区归属的影响。2011 年 Xie 等提出了基于信息广播机制的标签传播算法 SLPA[5]，利用历史标签信息来提高社区发现结果的稳定性，动态记录标签传播过程中节点的历史标签及出现频率，将节点分配到列表中出现频率高于给定阈值的社区中。2019 年 Lu 等提出了 LPANNI[6]算法，根据节点邻域中三角形连接密度定义标签更新顺序，利用节点邻域内的三角形密度及与邻居节点的路径连通信息定义邻居节点的影响力，根据邻居节点影响力确定标签更新策略，提高了大规模网络中社区发现的准确性和稳定性。

基于局部扩展和优化的重叠社区发现方法根据节点局部信息发现可能的社区中心节点，并以这些节点为种子节点进行社区扩展。由于这类算法无须获取网络的全局信息，因此更适用于大规模网络或动态网络的社区发现问题[7]。2009 年 Lancichinetti 等提出了 LFM[8]算法，随机选择一个未分配社区的节点作为种子，以社区内部连边占比为社区适应度函数，迭代地扩展种子节点，得到社区内部连边较多的社区结构。由于初始种子节点选择随机性大，因此 LFM 算法的社区发现结果不稳定。2010 年 Lee 等提出了 GCE[9]算法，选取网络中所有 K-团作为种子，以社区内部边所占比例为适应度函数扩展社区，由于寻找网络的所有 K-团代价较高，GCE 算法不适用于发现大规模网络中的社区结构。2017 年 Wang 等提出了 LOCD[10]算法，根据 k 步邻域内邻居节点的个数选择种子节点，利用社区内节点间的 2-路径连通性定义社区适应度函数来发现社区。2018 年 Rezvani 等[11]利用网络中三角形连接模式定义社区适应度函数，以发现三角形连接密集的社区结构。2020 年 Choumane 等提出了 Core_expansion 算法[12]，选择与邻居节点间 2-路径连通性高的节点集作为种子集，并根据未聚类节点与社区的 2-路径连接信息扩展社区。通常，基于局部扩展和优化的算法种子节点(集)一旦选定就不会改变，因此算法性能受初始种子节点(集)选取质量的影响较大。

利用三角形连接模式可以有效识别网络中内部连边比较稠密的社区结构，但无法识别内部连接相对稀疏的社区。实际上，在稀疏网络中社区内节点间的路径连通性比社区间节点的连通性更强，即处于同一社区的节点间往往通过多条短路径相互连接。基于此，本文提出了一种基于局部路径信息的重叠社区发现算法，主要包括种子节点选择与更新、社区扩展、未聚类节点处理3个主要部分。为了减小初始种子节点选择对算法性能的影响，LPIO 算法在社区发现过程中根据邻居节点的社区标签一致性迭代更新社区种子节点；根据社区内节点间的局部路径连通性定义了社区适应度函数，选择能提高社区适应度的节点扩展种子节点；最后根据未聚类节点与社区种子集间的局部路径数量确定其社区归属。

2 基础知识

2.1 基本概念和术语

复杂网络可以表示为一个图 $G=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 是节点集，E 是边集，记 $n_G=|V|$，$m_G=|E|$。除非特别声明，本文仅考虑无向无权图。

节点 v 的直接邻居节点的集合称作该节点的邻域，记作 $N(v)=\{u|u\in V,uv\in E\}$。记节点 v 的度 $d_v=|N(v)|$。节点子集 $S(\subseteq V)$ 的邻域表示为 $N(S)=\bigcup_{v\in S}N(v)-S$，定义子集 S 的体积为 S 中所有节点的度数和，记作 $Vol(S)=\sum_{v\in S}d_v$。若一个节点 v 的度不小于其邻域内所有节点的度，则称 v 是图 G 的一个局部极大度节点。

对于图 $G'=(V',E')$，若 $V'\subseteq V$ 且 $E'\subseteq E$，则称 G' 是 G 的一个子图。若 $E'=\{uv|u\in V',v\in V',uv\in E\}$，则称 G' 是节点集 V' 的导出子图，记作 $G[V']$。若 E' 可以表示为 l 条连续边的集合，即 $E'=\{u_0u_1,\cdots,u_{i-1}u_i,\cdots,u_{l-1}u_l\}$，则称 G' 为图 G 中一条长度为 l 的路径，若路径上节点两两不同，则称 G' 为图 G 中的一条长度为 l 的点不重复路径。

假设 $\Omega=\{C_1,C_2,\cdots,C_k\}$，其中 $C_i\subseteq V(G)$ 是图 G 的节点子集，且 $V=C_1\cup C_2\cup\cdots\cup C_k$，则称 Ω 为图 G 的一种社区结构，C_i 是一个社区。若 $C_i\cap C_j=\emptyset(1\leqslant i\neq j\leqslant k)$，则称 Ω 为图 G 的一种非重叠社区结构；否则称其为图 G 的一种重叠社区结构。

2.2 社区适应度

网络中的社区通常对应于网络中行使特定功能的模块，

社区内节点间连接相对紧密，社区间节点间的连接相对稀疏。为了发现网络中内部连接相对紧密的社区结构，研究人员根据节点间的不同连接模式定义了社区适应度以衡量社区结构的显著程度。

Radicchi 等[13]用社区 C 内部节点间直接连边的密度定义社区适应度，即：

$$\rho(C)=\frac{2|E_C|}{|V_C|(|V_C|-1)} \tag{1}$$

其中，V_C表示社区 C 的节点集合，E_C表示图 G 关于顶点子集 V_C的导出子图 $G[V_C]$的边集合。

Lancichinetti 等[8]提出的 LFM 算法中，将依附于社区节点的边中位于社区内部的边所占比例作为社区适应度函数，即：

$$f(C)=\frac{k_C^{\text{in}}}{(k_C^{\text{in}}+k_C^{\text{out}})^{\alpha}} \tag{2}$$

其中，$k_C^{\text{in}}=2|E_c|$表示社区 C 中节点内部度之和，k_C^{out} 表示社区 C 内节点与社区外节点关联的边数，α 是社区规模控制参数。k_C^{in} 和 k_C^{out} 也分别被称为社区 C 的内部度和外部度。

Kannan 等[14]提出的社区导电性指标 $con(C)$通过社区内节点与社区外节点的通信能力来度量社区内部连接的紧密程度，式(3)给出了社区导电性的定义。

$$con(C)=\frac{k_C^{\text{out}}}{\min\{Vol(C),Vol(V-C)\}} \tag{3}$$

社区 C 的导电性将社区向外连边的数量与社区体积进行比较，导电性越低，向外连边相对较少，社区结构越明显。

上述适应度函数是基于社区内连边数量来度量社区结构显著性的，社区内部连边越多，社区结构越明显。然而，现实网络通常是连接稀疏的，且社区规模分布不平衡，基于社区内连边数的社区适应度函数无法度量小规模社区或内部连边较少的社区结构的显著性。

Newman[15]提出了模块度指标，以随机网络为零模型评价网络中社区结构的显著性。图 G 的一种社区结构 $\Omega=\{C_1,C_2,\cdots,C_k\}$的模块度定义为：

$$Q(\Omega)=\sum_{c\in\Omega}\left(\frac{|E_C|}{|E|}-\left(\frac{Vol(C)}{2|E|}\right)^2\right)$$

其中，E 是图 G 的边集合，E_C表示图 G 关于社区 C 的节点子集 V_C导出子图 $G[V_C]$的边集合。模块度指标对社区内部节点间连边情况与相同规模随机网络的连边情况进行比较，来衡量每个社区 C 的显著性。

$$Q(C)=\frac{|E_C|}{|E|}-\left(\frac{Vol(C)}{2|E|}\right)^2 \tag{4}$$

以随机网络为参照，模块度指标 $Q(C)$在一定程度上消除了连边稀疏性对社区适应度的影响，以最优化模块度为基础的社区发现算法在实际中得到了广泛应用[16]。然而，基于模块度优化所发现的社区存在精度限制问题[17]，通常可以发现规模分布比较均匀的社区结构，对于小规模社区和非平衡分布的社区结构发现能力不足。

模块度指标通过社区内部节点对的直接连边数来衡量社区结构显著性，忽略了社区内节点间更复杂的连接关系。而在实际网络中，具有相同边数的社区往往因其内部连接方式的差异而具有不同的社区结构显著性。图 1 给出了两个节点数及边数相同的社区 C_1和 C_2，节点 v_{21}为未聚类节点，与 C_1和 C_2连边数相同。表 1 中，第 1－4 行分别给出了由式(1)－式(4)所计算的社区适应度，其中第 2 列和第 4 列是 C_1和 C_2的社区适应度值，第 3 列和第 5 列是将 v_{21}分别加入 C_1或 C_2所得社区的适应度值。可以看到，由于社区节点数和边数相同，式(1)－式(4)给出的适应度函数无法区分两个社区的结构显著性，也无法区分节点 v_{21}的加入对社区结构的影响。但实际上，尽管 C_1和 C_2边数相同，但在 C_1内部节点间有更多的短路径相互连通，比 C_2的路径连通性更好。而在加入节点 v_{21}后，C_2的节点间通过 v_{21}形成了更多短路径相互连接，而 v_{21}对 C_1中节点间的连通性影响较小，因此社区 $C_2\cup\{v_{21}\}$应比 $C_1\cup\{v_{21}\}$的适应度值更高。表 1 中的第 5 行给出了本文所定义的基于局部路径连通性的社区适应度函数 $LS_l(C)$，如式(5)所示，对以上情况进行了更好的度量。第 2.3 节给出了 $LS_l(C)$的详细定义。

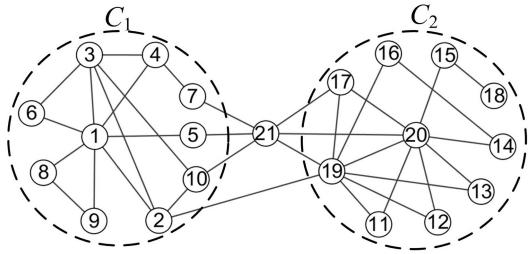

图 1 具有两个社区的网络示意图

Fig. 1 Network with two communities

表 1 节点 v_{21}对社区适应度的影响

Table 1 Impact of v_{21} on community fitness

社区适应度	C_1	$C_1\cup\{v_{21}\}$	C_2	$C_2\cup\{v_{21}\}$
$\rho(C)$	0.311	0.309	0.311	0.309
$f(C)$	0.875	0.894	0.875	0.894
$con(C)$	0.142	0.117	0.142	0.117
$Q(C)$	0.137	0.158	0.137	0.158
$LS_l(C)$	31.000	32.000	5.111	73.000

2.3 局部路径支持度

社区内节点间通常会通过一些特殊连接模式相互连通，如直接连边或三角形模式，利用这些特殊连接模式衡量社区显著性可以有效发现网络中的社区结构。然而，当社区内部连接相对稀疏时，社区内节点间的直接连接或者三角形连接不足以刻画社区结构。路径连通性是对社区结构的一种更通用的描述，社区内节点间的路径连通性通常比社区间节点的连通性更强，即处于同一社区的节点间往往通过多条短路径相互连接。然而，已有的社区适应度指标无法利用连边密度或三角形密度有效度量社区内节点间的路径连通性。针对此问题，本文首先提出了一种基于局部路径支持度的社区适应度函数，对社区内相邻节点间的路径连通性进行度量。

定义 1 和定义 2 分别给出了相邻节点对 u 和 v 在一条节点间点不重复路径 $P_{u,v}$上关于社区 C 的内部路径支持度和外部路径支持度。

定义 1 假设 $u,v\in G$ 且 $uv\in E$，$P_{u,v}(\neq\{u,v\})$是从 u 到 v 的一条点不重复路径。则在路径 $P_{u,v}$上相邻节点 u 和 v 关于社区 C 的内部路径支持度定义为：

$$L_{P,C}^{\text{in}}(u,v)=\frac{|x|x\in C,x\in P,x\neq u,x\neq v|}{|len(P)|-1}$$

其表示从 u 到 v 的点不重复路径 P 上属于社区 C 的节点所占的比例。

定义 2 假设 $u,v\in G$ 且 $uv\in E$，$P_{u,v}(\neq\{u,v\})$是从 u 到

v 的一条点不重复路径。则在路径 $P_{u,v}$ 上相邻节点 u 和 v 关于社区 C 的外部路径支持度定义为：

$$L_{P,C}^{\text{out}}(u,v)=1-L_{P,C}^{\text{in}}(u,v)$$

其表示从 u 到 v 的点不重复路径 P 上不属于社区 C 的节点所占的比例。

定义 3 和定义 4 利用相邻节点对 u 和 v 之间长度不超过 l 的点不重复路径，分别给出了 u 和 v 关于社区 C 的内部路径支持度和外部路径支持度。

定义 3 假设 $u,v\in G$ 且 $uv\in E$，则相邻节点 u 和 v 关于社区 C 的 l-路径内部支持度定义为：

$$S_{l,C}^{\text{in}}(u,v)=\sum_{P_{u,v}(\neq\{u,v\},len(P_{u,v})\leqslant l}L_{P,C}^{\text{in}}(u,v)$$

其表示所有从 u 到 v 的长度不超过 l 的点不重复路径上关于 C 的内部支持度之和。

定义 4 假设 $u,v\in G$ 且 $uv\in E$，则相邻节点 u 和 v 关于社区 C 的 l-路径外部支持度定义为：

$$S_{l,C}^{\text{out}}(u,v)=\sum_{P_{u,v}(\neq\{u,v\},len(P_{u,v})\leqslant l}L_{P,C}^{\text{out}}(u,v)$$

其表示所有从 u 到 v 的长度不超过 l 的点不重复路径上关于 C 的外部支持度之和。

由于社区结构是网络的一种局部结构特征，通常局限在节点的一个相对较小的局部邻域内，社区节点往往通过短路径连通，因此本文假设 $l\leqslant 3$。

基于定义 1—定义 4，定义 5 给出了基于局部路径支持度的社区适应度函数的定义。

定义 5 设 $\Omega=\{C_1,C_2,\cdots,C_k\}$ 是 G 的一种社区结构，则基于局部 l-路径支持度的社区 C 的适应度函数的定义如式(5)所示：

$$LS_l(C)=\frac{\sum_{x\in C,y\in C,xy\in E}S_{l,C}^{\text{in}}(x,y)}{\sum_{x\in C,y\in C,xy\in E}S_{l,C}^{\text{out}}(x,y)} \tag{5}$$

为了定义的完整性，令 $LS_l(\varnothing)=0$。

基于局部 l-路径支持度的适应度函数可以有效度量社区内部节点间连接的紧密程度。该适应度函数值越大，社区 C 内节点之间的连接更多的是出现在社区内部，更少与社区外的其余节点发生联系。

用定义 5 计算图 1 所示社区 C_1 和 C_2 的社区适应度，结果如表 1 的第 5 行第 2 列和第 4 列所示，可以看出，由于 C_1 中节点间有更多短路径连接，因此 $LS_l(C_1)>LS_l(C_2)$。表 1 中第 5 行第 3 列和第 5 列给出了将节点 v_{21} 分别加入 C_1 和 C_2 后的基于局部路径支持度的社区适应度，可以看出，$LS_l(C_2\cup\{v_{21}\})>LS_l(C_1\cup\{v_{21}\})$，这是由于 C_2 的节点间通过 v_{21} 形成了更多短路径相互连接。

3 基于局部路径信息的重叠社区发现算法

本节提出了一种基于局部路径信息的重叠社区发现算法，主要包括种子节点选择与更新、社区扩展、未聚类节点处理 3 个主要部分。在社区扩展阶段，定义了基于局部路径支持度的社区适应度函数；在未聚类节点处理阶段，根据局部路径连通性定义节点与社区间的距离，以发现节点间路径连通性较强的社区结构。

3.1 种子节点选择与更新

社区发现质量受初始种子节点的影响较大，而传统的基于局部扩展和优化的社区发现算法中，种子节点一旦选定将不再改变，从而影响社区发现质量。通常将节点度、介数、聚集系数、K 核等中心性指标最高的节点作为初始种子进行社区扩展，然而获取这些节点需要利用网络的全局信息，不适用于结构动态变化的网络。此外，网络中的社区规模分布通常是不平衡的，某些社区中可能不包含中心性指标高的节点。实际上，确定种子节点时应只考虑节点局部邻域内的中心性，当节点 v 在其局部邻域内的中心性指标高于其他节点时，则有可能成为种子节点。为了消除初始种子节点选择对社区扩展过程的影响，根据当前社区发现结果迭代更新各个社区的种子节点，可提高社区发现质量。

基于以上考虑，本文首先选择在局部邻域范围内的极大度节点作为初始种子节点，并在社区扩展过程中，根据社区内节点邻域的社区标签的一致性动态调整社区种子节点集。算法 1 给出了初始种子节点的选择过程。

算法 1 初始种子节点选择过程 InitSeed(G)

输入：网络 G=(V,E)

输出：初始种子节点集 I

1. 令 I=∅，CI=V，CS=V；
2. 若 CI=∅，则转步骤 7；
3. 从 CI 中随机选取节点 v；
4. 若 $d_v<1$，则 CI=CI−{v}，转步骤 2；
5. 若 v 是局部极大度点，即 $d_v\geqslant\max_{u\in N(v)\cap CS}\{d_u\}$，则令 I=I∪{v}，CI=CI−{v}−N(v)，CS=CS−{v}−N(v)，转步骤 2；否则转步骤 6；
6. CI=CI−{v}，转步骤 2；
7. 返回初始种子节点集 I，结束。

算法 1 选择局部极大度节点作为初始社区种子。考虑到度同配性高的网络中大度节点之间有边连接的概率更大，局部极大度节点间有连接的可能更大，这可能会导致所选择的初始种子节点为网络中的大度节点，从而使得位于网络稀疏区域的小度节点没有机会成为种子。针对此问题，算法 1 在将局部极大度节点 v 确定为种子节点的同时，将 v 及其邻域 $N(v)$ 中的节点从候选种子集中删除。

为了消除初始种子节点选择对社区发现结果的影响，对种子节点扩展得到社区后，本文将社区的种子节点更新为社区内邻域社区归属一致性高的节点。若节点邻域中的节点属于同一个社区，则该节点的社区归属也是相对确定的，将社区归属确定性高的节点确定为种子节点，有助于提高社区发现质量。基于以上考虑，定义 6 给出了节点集邻域标签熵的定义。

定义 6(邻域标签熵) 设 $\Omega=\{C_1,C_2,\cdots,C_k\}$ 是图 G 的一种社区结构，令社区 $C_i(1\leqslant i\leqslant k)$ 的类别标签为 i，$l(S)\subseteq\{1,\cdots,k\}$ 为 S 中节点的类别标签集合，则节点(集) S 的邻域标签概率分布 $\boldsymbol{P}_{N(S)}$ 定义为：

$$\boldsymbol{P}_{N(S)}(l=i)=\sum_{u\in N(S)}\frac{\delta(i\in l(u))}{|N(S)|\times|l(u)|}(i=1,\cdots,k)$$

其中，$\delta(\cdot)$ 为指示函数，如果 $\cdot$ 为真，则 $\delta(\cdot)=1$，否则 $\delta(\cdot)=0$。S 的邻域标签熵定义为：

$$EN(S)=-\sum_{1\leqslant i\leqslant k}\boldsymbol{P}_{N(S)}(l=i)\log(\boldsymbol{P}_{N(S)}(l=i)) \tag{6}$$

节点邻域标签熵可对节点 v 的邻域内社区标签的一致性进行度量，邻域标签熵越低，则其邻域内的社区标签一致性越高，该节点的社区归属确定性也越高。图 2 给出了跆拳道

网络(Karate)的两个真实社区 C_1 和 C_2。其中，节点 v_5 邻域内所有节点属于社区 C_2，而节点 v_9 邻域内节点分别属于社区 C_1 和 C_2，因此 v_5 的社区归属较 v_9 更确定。利用定义 6 计算 v_5 和 v_9 的邻域标签熵，可得 $EN(v_5)=0$，$EN(v_9)=0.971$。

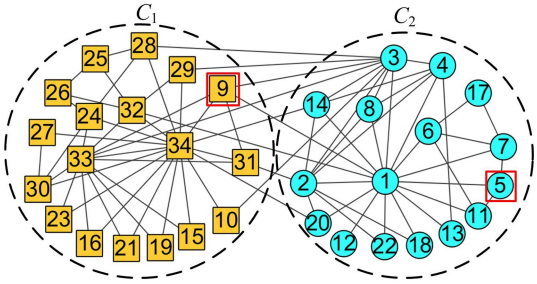

图 2 Karate 网络的真实社区划分结果

Fig. 2 Ground-truth community segmentation results of Karate network

在复杂网络中，大度节点可能行使多种网络功能而归属于多个社区，以大度节点作为种子进行社区扩展，可能会把包含大度节点的多个社区识别成一个社区。若大度节点邻域内节点社区标签种类较多，则其邻域标签熵会偏高，无法被选取为种子节点。因此，选取社区中邻域标签熵最小的节点集作为新种子集进行社区扩展，可以将社区范围限制在有效范围内，避免产生过大的社区。算法 2 给出了社区种子节点集的更新过程。

算法 2 社区种子节点集更新 Update(G,C)

输入：网络 G=(V,E)，社区 C

输出：种子节点集 I_C

1. 对于每个节点 v∈C，根据式(6)计算其邻域标签熵 EN(v)；
2. 令 $I_C=\{u|u\in C, EN(u)=\min\limits_{v\in C}\{EN(v)\}\}$；
3. 返回种子节点集 I_C，结束。

3.2 社区扩展过程

得到社区的种子集后，需要对种子节点进行扩展得到社区。对于每个种子集 *Seed*，在由种子集 *Seed* 扩展得到的导出子图 $G[Seed\cup N(Seed)]$ 上运行标签传播算法[18]，得到其局部邻域内的社区结构 $\Omega_{Seed}=\{C_{S_1},\cdots,C_{S_m}\}$。对 Ω_{Seed} 中的社区，根据定义 5 计算各社区的适应度，若两个社区的合并可提高社区适应度，则合并这两个社区。

大度节点对社区的路径连通性影响更大，当社区中存在大度节点时，社区合并过程可能会受到大度节点的影响而产生过度合并现象，将该社区与周围的小社区合并成一个很大的社区。为了避免社区过度合并，不仅应考虑社区内节点间的路径连通性，还应考虑待合并的两社区间节点的标签一致性，在社区扩展过程中合并那些节点标签一致性较高的社区。定义 7 给出了基于标签一致性的社区间距离定义。

定义 7(社区间距离) 设 $\Omega=\{C_1,C_2,\cdots,C_k\}$ 是图 G 的一种社区结构，令社区 C_i $(1\leqslant i\leqslant k)$ 的类别标签为 i，$l(v)\subseteq\{1,\cdots,k\}$ 为节点 v 的标签集合。设 S_i 和 S_j 分别为社区 C_i 和 C_j 的种子集，$\boldsymbol{P}_{N(S_i)}$ 和 $\boldsymbol{P}_{N(S_j)}$ 分别为 S_i 和 S_j 的邻域标签概率分布，则社区 C_i 与 C_j 间的距离定义为：

$$dis(C_i,C_j)=\frac{D(\boldsymbol{P}_{N(S_i)}\parallel\boldsymbol{P}_{N(S_j)})+D(\boldsymbol{P}_{N(S_j)}\parallel\boldsymbol{P}_{N(S_i)})}{2} \quad (7)$$

其中，$D(\boldsymbol{P}_{N(S_i)}\parallel\boldsymbol{P}_{N(S_j)})=\sum\limits_{x\in l(\Omega)}\boldsymbol{P}_{N(S_i)}(l=x)\log\dfrac{\boldsymbol{P}_{N(S_i)}(l=x)}{\boldsymbol{P}_{N(S_j)}(l=x)}$。

根据定义 7 给出的社区间距离定义，将由标签传播算法得到的种子节点集 *Seed* 邻域的社区结构 $\Omega_{Seed}=\{C_{S1},\cdots,C_{Sm}\}$ 中的社区进行合并，得到本轮迭代的社区分配结果。

算法 3 社区扩展 Extend(G,S)

输入：网络 G=(V,E)，社区结构 Ω，种子集 $S=\{s_1,\cdots,s_{|S|}\}$，社区间距离阈值 α=50

输出：社区结构 $\Omega=\{C_1,C_2,\cdots,C_k\}$

1. 令 i=0，Ω=∅；
2. 令 i=i+1，若 i>|S|，则转步骤 10，否则利用标签传播算法得到 s_i 局部邻域社区结构 $\Omega_{s_i}=\{\bar{C}_1,\cdots,\bar{C}_m\}$，令 t=0；
3. t=t+1；若 t>m，则转步骤 2；
4. 若 Ω=∅，则令 $\Omega=\{\bar{C}_t\}$，转步骤 3；
5. 令 $C_{max}=\arg\max\limits_{c_i\in\Omega,\bar{C}_t\cap N(C_i)\neq\emptyset}\{LS_l(\bar{C}_t\cup C_i)-LS_l(C_i)\}$，其中 $LS_l(\cdot)$ 根据式(5)计算得到；
6. 若 $LS_l(\bar{C}_t\cup C_{max})-LS_l(C_{max})<0$，则转步骤 3；
7. 根据式(7)计算 $dis(C_{max},\bar{C}_t)$；
8. 若 $dis(C_{max},\bar{C}_t)<\alpha$，则 $C_{max}=C_{max}\cup\bar{C}_t$；否则 $\Omega=\Omega\cup\{\bar{C}_t\}$；
9. 转步骤 3；
10. 结束。

以上算法中，步骤 2 利用标签传播算法对种子节点集进行社区扩展，而标签传播算法具有接近线性的时间复杂度[22]，因此对种子节点集进行一次社区扩展的时间复杂度近似为 $O(|E|)$。步骤 5 将社区 C_i 与使社区适应度增加最大的已有社区进行合并，通过计算社区内任意 2 个节点间的长度不超过 l 的最短路径可得到社区适应度，时间代价不超过 $O(|V|\times|E|)$。步骤 7 计算社区间距离需考虑每个节点的社区标签，时间代价为 $O(|V|)$。因此，算法 3 的总时间复杂度为 $O(|V|\times|E|)$。

3.3 未聚类节点处理

若某节点距离种子集较远或者节点度较小，则可能无法被扩展入任一社区，成为未聚类节点。如图 3 所示，无颜色节点为 Karate 网络中的未聚类节点，橙色节点和蓝色节点分别为发现的两个社区。复杂网络中的节点度通常服从幂律分布，因此存在大量度很小的节点，导致未聚类节点个数较多。对未聚类节点的处理在大规模稀疏网络的社区发现问题中至关重要。

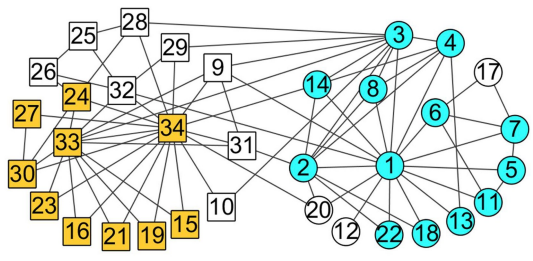

图 3 Karate 网络上的未聚类节点(电子版为彩图)

Fig. 3 Unclustered nodes on Karate network

通常未聚类节点与已有社区的直接连边数较少，无法利用直接连边确定其社区归属。此时，考虑利用未聚类节点与已有社区的种子集之间的点不重复路径数量来确定其与社区

的连接紧密程度。一个未聚类节点 u 与社区 C 的种子集 $Seed$ 之间存在的点不重复路径数量越多，则 u 与 C 的连接越紧密，u 属于 C 的概率更大。

定义 8 令 $L_l(u,v)$ 是图 G 中节点 u 与 v 间所有长度为 l 的点不重复路径，则未聚类节点 u 与社区 C_i 的距离定义为：

$$dis_u(u,C_i)=\sum_{v\in Seed_i}(|L_2(u,v)|+|L_3(u,v)|) \tag{8}$$

其中，$Seed_i$ 为社区 C_i 的种子节点集。

根据式(8)得到未聚类节点 u 与已有社区 C_i ($i=1,\cdots,|\Omega|$) 的距离 $dis_u(u,C_i)$ ($i=1,\cdots,|\Omega|$)，并对其降序排列，从中选择使 $dis_u(u,C_i)$ 变化率最大的 $dis_u(u,C_i)$ 作为确定 u 社区归属的阈值 h_u，若 $dis_u(u,C_i)>h_u$，则将 u 分配到社区 C_i。图 4 给出了未聚类节点阈值的确定过程示意图。

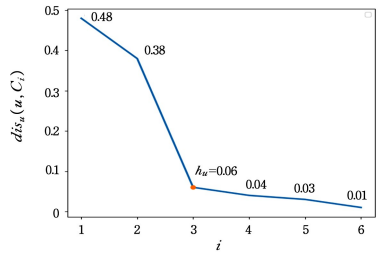

图 4 未聚类节点 u 的阈值确定

Fig. 4 Selection of threshold of unclustered node u

3.4 算法框架

算法 4 给出了基于局部路径信息的重叠社区发现算法 LPIO 的框架。

算法 4 基于局部路径信息的重叠社区发现算法

输入：网络 G=(V,E)

输出：社区结构 Ω={C_1,C_2,…,C_k}

1. 初始种子节点集 I=InitSeed(G)，S=I；
2. Ω=∅；
3. Ω= Extend(G,S)；
4. 对 Ω 中的每个社区 C_i($1\leqslant i\leqslant|\Omega|$)，利用算法 2 更新 C_i 的种子节点集 S_i= Update(G,C_i)；
5. 若 S≠{S_i|i=1,…,|Ω|}，则 S={S_i|i=1,…,|Ω|}，转步骤 2；
6. 令未聚类节点集 $V_{un}=V-\bigcup_{C\in\Omega}C$；
7. 若 $V_{un}=\emptyset$，则转步骤 13；
8. 对每个节点 u∈V_{un}，
9. 计算 $dis_u(u,C_i)$(i=1,…,|Ω|)；
10. 令社区归属阈值 h_u 为变化率最大处的 $dis_u(u,C_i)$ 值；
11. 对于 i=1,…,|Ω|，若 $dis_u(u,C_i)>h_u$，则 $C_i=C_i\cup\{u\}$；
12. 转步骤 6；
13. 返回社区发现结果 Ω，结束。

算法 4 中，步骤 1 选择局部极大度节点作为初始种子节点集，判断一个节点 v 是否为局部极大度节点的代价为 $O(d_v)$，因此初始化种子节点集的时间代价为 $O(|E|)$；步骤 3 对种子节点集进行一次社区扩展的时间代价为 $O(|V|\times|E|)$；步骤 4 根据节点的邻域标签熵更新种子集，其中计算每个节点的邻域标签熵的代价为 $O(|E|)$；步骤 8－步骤 11 处理未聚类节点集 V_{un}，假设未聚类节点平均度为 $\bar{d}_{un}$，则计算未聚类节点与其他社区间距离的代价为 $O(|V_{un}|\times\bar{d}_{un}^3)$。由于未聚类节点的度数通常很小，因此可把 $\bar{d}_{un}$ 看做常数。假设种子集更新次数为 t，则算法 4 的总时间复杂度为 $O(t\times|V|\times|E|)$。

4 实验结果

本节在 4 个有标签网络和 8 个无标签网络上，将本文提出的基于局部路径信息的重叠社区发现算法 LPIO 与 CPM[2]，SLPA[5]，LPANNI[6]，LFM[8]，DEMON[19]，Core_expansion[12]，Node_perception[20] 7 种经典重叠社区发现算法进行了比较。数据集的基本情况如表 2 所列。

表 2 数据集的基本情况

Table 2 Basic information of datasets

数据集	节点数	边数	社区数
Karate	34	78	2
Dolphins	62	159	2
Polbooks	105	441	3
Football	115	613	12
Les	77	254	—
Jazz	198	2742	—
USAIR	332	2126	—
NetScience	379	914	—
Email	1133	5451	—
Power	4941	6594	—
PGP	10680	24316	—
Ca-Astroph	18771	198050	—

4.1 评价标准

在有标签的网络上，使用 McDaid 等[25]提出的重叠标准互信息(ONMI)指标和 F_1 分数评价算法性能[21]。假设 n 为网络的节点个数，$\Omega=\{S_1,S_2,\cdots,S_K\}$ 为检测到的社区结构，$\Gamma=\{O_1,O_2,\cdots,O_T\}$ 为真实社区结构，ONMI 的计算式为：

$$ONMI=\frac{I(\Omega;\Gamma)}{\max(H(\Omega),H(\Gamma))}$$

其中：

$$H(\Omega)=\sum_{1\leqslant i\leqslant K}\left(-\frac{|S_i|}{n}\log\frac{|S_i|}{n}-\frac{n-|S_i|}{n}\log\left(\frac{n-|S_i|}{n}\right)\right)$$

$$H(\Gamma)=\sum_{1\leqslant i\leqslant T}\left(-\frac{|O_i|}{n}\log\frac{|O_i|}{n}-\frac{n-|O_i|}{n}\log\left(\frac{n-|O_i|}{n}\right)\right)$$

$$I(\Omega;\Gamma)=\frac{1}{2}[H(\Omega)-H(\Omega|\Gamma)+H(\Gamma)-H(\Gamma|\Omega)]$$

重叠标准互信息(ONMI)可以衡量算法发现的社区结构 Ω 和网络真实社区结构 Γ 的一致程度，ONMI 越大，发现的社区分布更符合真实社区分布。

对于无标签网络，采用扩展模块度 EQ[21]进行评价，其定义为：

$$EQ=\frac{1}{2|E|}\sum_{k=1}^{m}\sum_{v,w\in C_k}\frac{1}{O_vO_w}\left(a_{vw}-\frac{d_vd_w}{2|E|}\right)$$

其中，v 和 w 表示节点，m 为社区数目，O_v 表示节点 v 所属的社区数，a_{vw} 为邻接矩阵元素。通常，EQ 的值越大，社区结构越明显。

4.2 实验结果与分析

4.2.1 有标签网络的实验结果比较

表 3 和表 4 分别列出了跆拳道网络(Karate)、海豚社交网络(Dolphins)、美国政治书网络(Polbooks)、大学生足球网络(Football)4 个有标签网络的算法比较结果，用 $ONMI$ 值和 F_1 分数评价社区发现质量。

表 3 有标签网络上的 *ONMI* 实验结果

Table 3 Experimental results of *ONMI* on labeled networks

Network	CPM	SLPA	LPANNI	LFM	Core_expansion	DEMON	Node_perception	LPIO
Karate	0.1653±0	0.1933±0.1524	0.6277±0	0.6381±0.0579	0.4459±0	0.1321±0	0.1225±0	**0.9170±0**
Dolphins	0.2751±0	0.3165±0.1247	0.3252±0	0.5776±0.1287	0.1444±0	0.1206±0.1691	0.0955±0.0118	**0.6753±0.2125**
Polbooks	0.3252±0	0.2952±0.0467	0.2546±0	0.3096±0	0.1003±0	0.2258±0.0782	0.1042±0.0189	**0.4964±0.0077**
Football	0.1603±0	0.6224±0.0997	**0.7343±0**	0.6095±0.0494	0.3936±0	0.3013±0.0006	0.2039±0.0028	0.3400±0.0078

表 4 有标签网络上的 F_1 分数实验结果

Table 4 Experimental results of F_1 Score on labeled networks

Network	CPM	SLPA	LPANNI	LFM	Core_expansion	DEMON	Node_perception	LPIO
Karate	0.4466±0	0.4959±0.1504	0.7633±0	0.8750±0	0.7800±0	0.5750±0	0.3041±0.0334	**0.9850±0**
Dolphins	0.4800±0	0.4996±0.1148	0.4116±0	**0.9099±0**	0.2309±0	0.3700±0.4100	0.1151±0.0701	0.7908±0.1277
Polbooks	0.3690±0	0.4777±0.0774	0.3012±0	0.5766±0	0.1515±0	0.3400±0.0600	0.1866±0.0376	**0.7895±0.1215**
Football	0.5700±0	0.7799±0.0918	**0.8949±0**	0.8160±0.0185	0.4541±0	0.4412±0	0.5176±0.0231	0.5737±0.0424

LPIO 算法在 Karate，Dolphins，Polbooks 3 个网络上的重叠标准互信息 *ONMI* 和 F_1 分数明显优于对比算法，在 Football 网络上 *ONMI* 值有所下降。实际上，Football 网络节点表示一个美国大学生足球队，边表示球队间的比赛关系，社区代表一个美国大学生足球联盟。大多数联盟内部球队间的比赛多，因此对应社区内部连接紧密。但有些联盟球队较少，地域分散，导致联盟内比赛较少，如图 5 中节点数较少的几个社区，这些社区内部的连通性较低，而与社区外节点间形成了比较多的局部连通路径。这部分社区节点被 LPIO 算法识别为重叠节点。图 5 和图 6 分别给出了 Football 网络的真实社区结构和算法 LPIO 发现的社区结构，图 6 中黄色节点为重叠节点。可以看出，LPIO 能够正确识别联盟内比赛较多的社区，将连接比较松散的社区中的节点识别为重叠节点。

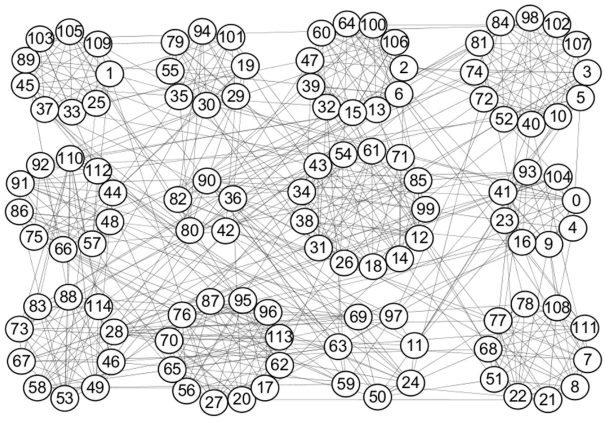

图 5 Football 网络真实社区结构

Fig. 5 Ground-truth community structure of Football network

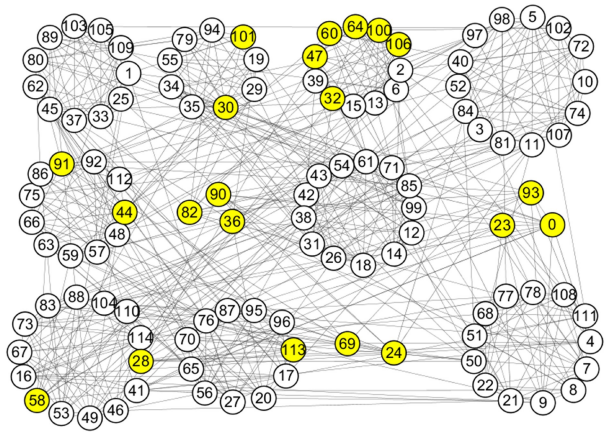

图 6 LPIO 算法在 Football 网络上的结果

（电子版为彩图）

Fig. 6 Communities detected by LPIO on Football network

4.2.2 无标签网络的实验结果比较

表 5 列出了在 Les，Jazz，USAIR，NetScience，Email，Power，PGP，Ca-Astroph 8 个无标签网络上算法的对比结果，用扩展模块度 EQ 作为评价指标，最后一行给出了各算法在实验网络上所发现社区的扩展模块度平均值。可以看出，所提算法 LPIO 在大多数网络上扩展模块度较高，这说明算法能较好发现网络中的重叠社区结构。当网络社区结构中存在较多重叠节点时，该社区结构的扩展模块度较低。由于算法 LPIO 识别出更多重叠节点，这导致其扩展模块度值较算法 SLPA 和 LPANNI 低，特别在小规模网络上，重叠节点对扩展模块度的影响更显著。

表 5 无标签网络上的扩展模块度 *EQ* 实验结果

Table 5 Experimental results of *EQ* on unlabeled networks

Network	CPM	SLPA	LPANNI	LFM	Core_expansion	DEMON	Node_perception	LPIO
Les	0.1846±0	0.3241±0.1079	**0.5152±0**	0.3386±0.0445	0.3847±0	0.0202±0.0241	0.3041±0.0082	0.2966±0.0593
Jazz	0.0028±0	0.3705±0.0736	**0.4310±0**	0.2853±0.0016	0.1675±0	0.000016±0	0.1151±0.0352	0.2129±0.0685
USAIR	0.0414±0	0.0645±0.0447	0.1047±0	0.0593±0.0071	**0.1192±0**	0.0204±0	0.1018±0.0019	0.1177±0.0113
NetScience	0.6287±0	0.7356±0.0173	0.7295±0	0.7239±0.0255	0.6206±0	0.5074±0.1213	0.5641±0	**0.7863±0.0055**
Email	0.0302±0	**0.4062±0.0157**	0.2663±0	0.1200±0.0072	0.1849±0	0.0302±0.0023	0.1065±0.0032	0.3138±0.0024
Power	0.1577±0	0.6470±0.0050	0.6207±0	0.5581±0.0470	0.5688±0	0.0849±0.0041	0.1299±0	**0.8466±0.0022**
PGP	0.3843±0	**0.7438±0.0417**	0.6814±0	0.5467±0.0502	0.4028±0	0.2653±0.0057	0.3156±0.0009	0.7083±0.0116
Ca-Astroph	0.0792±0	**0.5221±0.0026**	0.4539±0	0.2553±0.0344	0.2310±0	0.0947±0.0002	0.2161±0.0008	0.3230±0.0021
Average	0.1886±0	**0.4767±0.0385**	0.4753±0	0.3609±0.0271	0.3349±0	0.1278±0.0197	0.2316±0.0062	0.4506±0.0203

结束语 针对现有的基于局部扩展和优化的重叠社区发现方法受初始种子节点选择影响较大，适应度函数无法度量

节点间多样的连接方式等问题，本文基于局部扩展和优化策略提出了一种基于局部路径信息的重叠社区发现算法。首先定义了节点的邻域标签熵对社区内节点邻域标签一致性进行度量，选取邻域标签熵低的节点更新社区的种子节点集，这有助于在社区扩展过程中确定合理的社区边界，避免初始种子节点对社区发现结果的影响；给出了基于局部路径信息的社区适应度函数，以便在社区扩展过程中考虑网络中节点间丰富的连接模式；最后根据未聚类节点与已有社区间的路径连通信息确定未聚类节点的社区归属。在 4 个有标签网络和 8 个无标签网络上，与 7 个经典重叠社区发现算法进行了对比实验，结果表明，所提出的算法 LPIO 在重叠标准互信息 *ONMI*、F_1 分数、扩展模块度 *EQ* 等方面表现良好。

从局部路径连通性角度进行复杂网络社区发现，可以发现内部路径连通性强的社区结构。在实际网络中，路径连通性可能仍无法准确刻画不同网络中社区的连接模式。如何发现网络中具有特定连接方式的模体，进而发现重叠社区结构是值得深入研究的课题。

参 考 文 献

[1] CHENG F, WANG C, ZHANG X, et al. A local-neighborhood information based overlapping community detection algorithm for large-scale complex networks[J]. IEEE/ACM Transactions on Networking, 2020, 29(2): 543-556.

[2] PALLA G, DERÉNYI I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[3] JABBOUR S, MHADHBI N, RADAOUI B, et al. Detecting highly overlapping community structure by model-based maximal clique expansion[C]// 2018 IEEE International Conference on Big Data(Big Data). IEEE, 2018: 1031-1036.

[4] GREGORY S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics, 2010, 12(10): 103018.

[5] XIE J, SZYMANSKI B K, LIU X. SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process[C]// 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011: 344-349.

[6] LU M, ZHANG Z, QU Z, et al. LPANNI: Overlapping community detection using label propagation in large-scale complex networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(9): 1736-1749.

[7] LUO W, ZHANG D, NI L, et al. Multiscale local community detection in social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(3): 1102-1112.

[8] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New journal of physics, 2009, 11(3): 033015.

[9] LEE C, REID F, MCDAID A, et al. Detecting highly overlapping community structure by greedy clique expansion[C]// Proceedings of the 4th International Workshop on Social Network Mining and Analysis(SNA-KDD). 2010: 33-42.

[10] WANG X, LIU G, LI J. Overlapping community detection based on structural centrality in complex networks[J]. IEEE Access, 2017, 5: 25258-25269.

[11] REZVANI M, LIANG W, LIU C, et al. Efficient detection of overlapping communities using asymmetric triangle cuts[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 30(11): 2093-2105.

[12] CHOUMANE A, AWADA A, HARKOUS A. Core expansion: A new community detection algorithm based on neighborhood overlap[J]. Social Network Analysis and Mining, 2020, 10(1): 1-11.

[13] RADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences, 2004, 101(9): 2658-2663.

[14] KANNAN R, VEMPALA S, VETTA A. On clusterings: Good, bad and spectral[J]. Journal of the ACM(JACM), 2004, 51(3): 497-515.

[15] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences, 2006, 103(23): 8577-8582.

[16] CARNIVALI G S, VIEIRA A B, ZIVIANI A, et al. CoVeC: coarse-grained vertex clustering for efficient community detection in sparse complex networks[J]. Information Sciences, 2020, 522: 180-192.

[17] FORTUNATO S, BARTHELEMY M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

[18] RAGHAVAN U N, ALBERT R, KUMARA S. Near linear time algorithm to detect community structures in large-scale networks[J]. Physical Review E, 2007, 76(3): 036106.

[18] COSCIA M, ROSSETTI G, GIANNOTTI F, et al. Demon: a local-first discovery method for overlapping communities[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 615-623.

[20] SOUNDARAJAN S, HOPCROFT J E. Use of local group information to identify communities in networks[J]. ACM Transactions on Knowledge Discovery from Data(TKDD), 2015, 9(3): 1-27.

[21] CHAKRABORTY T, DALMIA A, MUKHERJEE A, et al. Metrics for community analysis: A survey[J]. ACM Computing Surveys(CSUR), 2017, 50(4): 1-37.

ZHENG Wen-ping, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include graph theory algorithms and bioinformatics.

（责任编辑：喻藜）