基于拓扑结构的密度峰值重叠社区发现算法

封云飞[1] 陈红梅[1,2]
(西南交通大学信息科学与技术学院 成都 611756)[1]
(西南交通大学云计算与智能技术高校重点实验室 成都 611756)[2]

摘 要 现代网络科学的不断发展,为人们的生活提供了极大的便利。对复杂网络的研究是推动现代网络科学发展的重要动力,而社区是研究复杂网络的重要结构。已有的社区发现方法大多是高度复杂的,这不利于有效挖掘复杂网络。为了研究更高效的社区发现算法,文中将近年来被提出的密度峰值聚类算法应用于社区发现中,对密度峰值算法进行改进,提出了一种高效的社区发现算法。将密度峰值算法应用于社区发现存在一些问题,由于复杂网络数据结构具有特殊性,其数据大多以拓扑图或邻接矩阵的形式存储,因此将密度峰值聚类算法应用到社区发现中的核心问题是如何有效地计算网络中各节点间的距离、节点局部密度和选择中心节点。针对该问题,文中通过网络拓扑图中各节点及其邻居节点的度来计算每一个节点的局部密度,通过节点间的相似度来度量节点间的距离,并对距离进行离散化处理,以便选取社区中心节点;定义了核心跳变值来更精确地选取社区中心,防止大社区吞并小社区;基于 LFR 人工网络和真实网络数据集,将所提算法与已有算法进行比较,并采用扩展的模块度、调整兰德系数以及归一化互信息对实验结果进行评估。真实网络中的实验结果表明了所提算法具有不错的效果,且在一些真实场景中具有明显优势;在人工网络中,所提算法同样具有优势,同时其相比其他算法更加稳定。

关键词 社区发现,重叠社区,密度峰值,拓扑结构,数据挖掘
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.180901644

Topological Structure Based Density Peak Algorithm for Overlapping Community Detection

FENG Yun-fei[1] CHEN Hong-mei[1,2]
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)[1]
(Key Laboratory of Cloud Computing and Intelligent Technology, Southwest Jiaotong University, Chengdu 611756, China)[2]

Abstract With the continuous development of modern network science, people's lives have been greatly facilitated. The study on complex networks is an important driving force for the development of modern network science, and the community is an important structure for research on complex networks. Most of the existing community discovery methods are highly complex, so this paper applied the density peak clustering algorithm proposed in recent years to community discovery, and proposed an efficient community discovery algorithm. Due to the particularity of the data structure of complex network, the complex network data are stored in the form of topological graphorad jacency matrix, and thus how to effectively calculate the distance between nodes and the local density of each node, and how to select the core nodes of the community are the key problems when density peak clustering algorithm is applied to community discovery. In light of this, this paper calculated the local density of each node by the degree of each node and its neighbor nodes in the network topological graph, measured the distance between nodes by the similarity between nodes, and discretized the distance to expediently select the core nodes for this algorithm. Moreover, the core hop values were defined to accurately select community centers, thus avoiding the large communities annex small communities. Experiments were carried out on the LFR artificial network dataset and the real network dataset with the evaluation indexes of the extended modularity, the adjusted rand coefficient and the normalized mutual information. Experimental results in real networks show that the proposed algorithm has good effects and obvious advantages in some real networks compared with other algorithms. In the artificial network, the algorithm also has advantages. Therefore, the proposed algorithm is more stable than other algorithms.

Keywords Community detection, Overlapping community, Density peak, Topological structure, Data mining

---

到稿日期:2018-09-04 返修日期:2018-11-14 本文受国家自然科学基金(61572406)资助。
封云飞(1995—),男,硕士生,主要研究方向为数据挖掘、机器学习等;**陈红梅**(1971—),女,教授,博士生导师,主要研究方向为粒计算与粗糙集、智能信息处理、数据挖掘等,E-mail:hmchen@swjtu.edu.cn(通信作者)。

1 引言

现实世界中的许多问题都可以被抽象为复杂网络，如生物网络、社交网络、神经网络等。复杂网络中的个体间存在着联系，某些个体之间联系紧密，而某些个体间联系稀疏甚至没有联系，联系紧密的个体构成群体，称之为社区[1]。以往对复杂网络的研究表明，复杂网络中蕴含着更深层次的结构和关系，因此研究复杂网络最重要的一步就是发掘节点间的结构规律，而社区是研究复杂网络的重要结构[2]。社区发现是将复杂的网络划分为若干社区的过程，同一社区中的节点之间连接紧密，不同社区的节点之间连接稀疏。社区发现有助于找到网络节点间所蕴含的结构关系，因此它是研究复杂网络的重要方法之一。社区发现能够检测复杂网络中具有紧密联系或者某种共性的群体，被应用于很多领域，如好友推荐[3]、精准营销[4]、个性化服务[5]等。对社区发现进行研究具有重要意义，吸引了众多国内外学者的关注。

许多学者为社区发现领域做出了重要贡献，Girvan 等[6]提出了经典的社区划分方法——GN 算法，这是一种基于分裂思想的社区发现算法，且无需预先设定社区的个数。Newman[7]在 GN 算法的基础上提出了模块度，用于度量所划分社区结构的模块性，为社区发现领域的发展奠定了重要基础。针对社区发现，学者们提出了不同的方法，例如聚合或分裂的方法、标签传播算法、演化算法、谱聚类算法等。Derenyi 等[8]提出了派系过滤(CPM)算法，该算法基于团渗透理论，通过合并完全子图来检测网络中的社区结构。Clauset 等[9]提出了节点凝聚(CNM)算法，这是一种基于聚合思想的社区发现算法。不同于上述聚合方法，Danon 等[10]提出了一种通过分裂来发现社区的方法。标签传播算法是社区发现常用的方法之一，该方法最早由 Zhu 等[11]提出，但传统的标签传播算法无法划分重叠社区，且存在标签震荡、逆流等缺点。刘世超等[12]和 Zhang 等[13]分别对标签传播算法做出了改进，提出了新的标签传播规则，使得标签传播算法更加稳定。近年来，一些学者将群体智能算法应用于社区发现，如 Li 等[14]将粒子群算法应用于社区发现，Liu 等[15]将果蝇算法应用于社区发现。但基于演化算法的社区发现算法存在初始化对算法结果影响较大、陷入局部优化、计算复杂度高等问题。蒋盛益等[16]提出了谱聚类动态社区划分算法，该算法是一种增量式社区发现算法，能够对动态的网络进行社区划分。Zhou 等[17]根据边的相似度，提出了基于边密度的社区发现算法，该算法表明基于密度的社区发现算法能得到质量较高的社区，并且其具有较低的时间复杂度。

密度峰值聚类算法[18]由 Alex 等于 2014 年提出。密度峰值聚类算法是一种新颖的基于密度的聚类算法，对于非球形的数据集聚类具有较明显的优势[19]。复杂网络中，节点间的连接复杂，且社区结构不规则，因此用密度峰值算法来探测复杂网络中的社区结构有望得到较好的划分结果，故一些学者将密度峰值算法应用于社区发现。网络中各节点间的距离计算，是密度峰值社区发现算法的重要步骤。Shi 等[20]提出了基于密度峰值的重叠社区发现算法，该算法中的距离是通过节点间的连接强度来衡量的，但当网络规模增大且结构复杂时，该计算方式的时间复杂度较大。黄岚等[21]提出了基于点距离的密度峰值社区发现算法，该算法采用 Jaccard 相似度[22]来计算节点间的距离，但这种方式计算出的距离值较紧凑，不利于社区中心的选择。上述密度峰值社区发现算法不仅时间代价较高，而且划分的社区质量较低。

本文提出的基于拓扑结构的密度峰值(TSDP)重叠社区发现算法结合了密度峰值算法的基本思想，并对其进行了相关改进。TSDP 算法根据网络拓扑结构中节点的拓扑关系来计算节点的局部密度，这种方式能较好地衡量节点的局部密度，且时间代价较低；利用节点的相似度计算节点间的距离，并将距离值进行离散化处理，以便选取社区中心。聚类中心的选择也是密度峰值算法的关键，TSDP 算法定义了核心跳变值来帮助选取社区中心，能够较好地避免大社区吞并小社区。在人工网络和真实网络数据集上的实验结果表明，本文算法具有可行性和有效性。

2 密度峰值聚类算法

密度峰值聚类算法是一种基于密度的聚类算法，无需预先给定聚类中心的个数。该算法能够根据数据的分布自动地选取一些局部密度较大且相互间距离较远的点作为聚类中心，然后将其他的样本点划分到距离最近且局部密度较大的点所属的簇中。

密度峰值聚类算法为节点定义了两个属性 ρ 和 δ，ρ 表示当前节点的局部密度，δ 表示局部密度大于当前节点且距离当前节点最近的节点到当前节点的距离。局部密度最大的节点的 δ 为距离其最远的节点到它的距离。节点的局部密度 ρ 的计算式如下：

$$\rho_i=\sum_j \chi(d_{ij}-d_c) \tag{1}$$

$$\chi(x)=\begin{cases}1, & x<0\\ 0, & \text{其他}\end{cases} \tag{2}$$

其中，ρ_i 表示第 i 个节点的局部密度，d_{ij} 表示节点 i 与节点 j 之间的距离，d_c 为截断距离。

节点的最小距离 δ 的计算式如下：

$$\delta_i=\begin{cases}\max\limits_j(d_{ij}), & \rho_i=\max\limits_k(\rho_k)\\ \min\limits_{j:\rho_j>\rho_i}(d_{ij}), & \text{其他}\end{cases} \tag{3}$$

选取同时具有较大 ρ 和 δ 的节点作为聚类中心点，其他节点则依次被划分到局部密度大于当前节点且距离当前节点最近的节点所对应的簇中。

3 TSDP 算法

TSDP 算法引入了网络拓扑中节点的度以及其邻居节点的度来衡量各节点的局部密度，利用节点间的相似度来计算节点间的距离，定义了核心跳变值以有效地选取社区中心，引入了重叠参数来划分重叠社区。TSDP 算法的流程如图 1 所示。

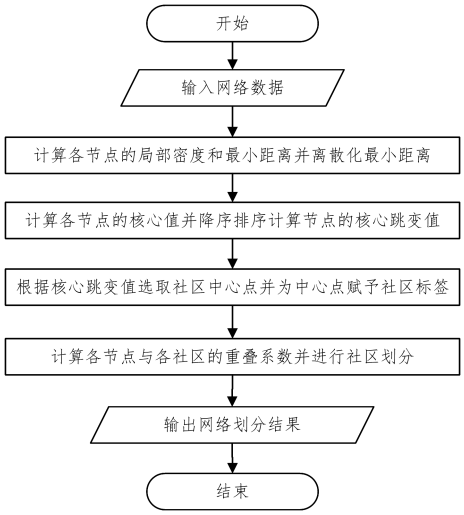


图 1　TSDP 算法的流程图

Fig. 1　Flow chart of TSDP algorithm

3.1　局部密度和最小距离的计算

在网络拓扑中,当两个节点存在着某种密切的联系或者具有某种共性时,两个节点间会存在一条边将两个节点连接起来,因此可以理解为:如果两个节点间存在边,使得两个节点直接相连,那么这两个节点具有紧密的联系和较高的相似性,即这两个节点之间相隔较"近"。在网络中,如果一个节点存在较多的邻居且有较多的节点与它直接相连,则可以理解为有较多的节点分布在该节点附近,使得当前节点具有较高的局部密度。因此,TSDP 算法采用节点的直接邻居的个数即当前节点的度来衡量节点的局部密度,直接邻居越多,节点的局部密度就越大,同时还要考虑间接邻居对当前节点局部密度的影响。节点的局部密度计算式如下:

$$\rho_i = k_i + \sum_{j \in n_i} k_j \tag{4}$$

其中,ρ_i 表示节点 i 的局部密度,k_i 表示节点 i 的度,n_i 表示与节点 i 直接相连的点构成的集合(节点 i 的邻居节点构成的集合),k_j 表示节点 j 的度。

图 2 给出了两个不同结构的社区。根据式(4)分别计算图 2 所示两个社区中黑色节点的局部密度,其中图 2(a)社区中黑色节点的局部密度为 3+2+2+2=9,图 2(b)社区中黑色节点的局部密度为 1+8=9,可以看出两个黑色节点具有相同的局部密度。但图 2(b)社区中黑色节点只有一个邻居节点(灰色节点),由于这个邻居节点的度较大,因此黑色节点具有了较大的局部密度。

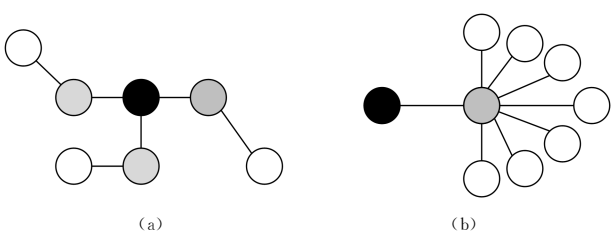


图 2　两个结构不同的社区

Fig. 2　Two communities with different structures

为了解决上述问题,要求邻居节点的度对当前节点局部密度的贡献度存在一定的衰减,因此引入衰减系数 ζ,$\zeta \in (0,1)$。ρ_i 的最终计算式如下:

$$\rho_i = k_i + \zeta \cdot \sum_{j \in n_i} k_j \tag{5}$$

利用式(5)计算图 2 中两个社区中黑色节点的局部密度,设 $\zeta=0.5$,则图 2(a)社区中黑色节点的局部密度为 3+0.5×(2+2+2)=6,图 2(b)社区中黑色节点的局部密度为 1+0.5×8=5。

δ_i 为局部密度大于节点 i 且距离节点 i 最近的节点到节点 i 的距离。距离的计算转换为相似度的计算,两节点的相似度越大,它们间的距离越小,反之距离越大。节点间的相似度及距离计算式如下:

$$S_{ij} = \frac{|\tau(i) \bigcap \tau(j)|}{\sqrt{|\tau(i)| \times |\tau(j)|}} \tag{6}$$

$$D_{ij} = 1 - S_{ij} \tag{7}$$

$$\delta_i = \begin{cases} \max\limits_j (D_{ij}), & \rho_i = \max\limits_k (\rho_k) \\ \min\limits_{j:\rho_j > \rho_i} (D_{ij}), & \text{其他} \end{cases} \tag{8}$$

其中,S_{ij} 表示节点 i 和节点 j 间的相似度,$\tau(i)$表示节点 i 本身及其邻居节点所构成的集合,D_{ij} 表示节点 i 和节点 j 之间的距离。相似度 $S_{ij} \in [0,1]$,因此求得 $\delta_i \in [0,1]$。δ 值过于紧凑时不利于中心点的选择,因此使用离散公式对 δ 值进行离散化处理。δ 的离散化公式如下:

$$\delta_i = \exp\left(\frac{\delta_i}{\Delta}\right)^2, \Delta \in (0,1) \tag{9}$$

3.2　社区中心的选择

社区中心的选择是密度峰值聚类算法的重要步骤之一。密度峰值算法在选取聚类中心(社区中心)时,要求选取的聚类中心同时具有较大的 ρ 和 δ。定义节点的核心值计算式如下:

$$C_i = \rho_i \times \delta_i \tag{10}$$

核心值(C)的大小可以衡量节点被选为社区中心点的潜力,核心值越大的节点越有可能成为社区中心点。但对于不同的数据集,各节点的核心值大小也不同,我们无法给出一个固定的值来衡量节点的核心值是否足够大以被选作社区中心。为了更好地选取社区中心,将所有节点按照核心值(C_i)的大小降序排列,并定义了核心跳变值(CJ)来衡量节点核心值的变化情况,进而帮助算法更精确地选取社区中心。CJ 的计算式如下:

$$CJ_i = \begin{cases} 0, & C_i = \max\limits_k (C_k) \\ \dfrac{C_{i-1} - C_i}{C_i - C_{i+\frac{n}{2}}} + \mu, & \text{其他} \end{cases} \tag{11}$$

$$\mu = 0.1 \times (C_{i-1} - C_i) \tag{12}$$

其中,i 为当前节点序号,$C_{i-1} - C_i$ 表示上一个节点核心值与当前节点核心值的差值;$C_i - C_{i+\frac{n}{2}}$ 表示当前节点核心值与余下节点的中间节点核心值的差值;μ 为平滑参数,防止分母为 0。

将节点按核心值降序排列后,可以得到节点核心值从大到小的变化情况。被选为社区中心点的节点应具有较大的核心值;而非社区中心节点的核心值较小,且变化趋于平缓。中心节点与非中心节点的核心值大小存在较大差异,因此需要找到中心节点与非中心节点的分界点,分界点具有最大的核

心跳变值。因为分界点之后的节点均为非中心点，而非中心点普遍具有较小的核心值，这会导致 $C_i-C_{\frac{i+n}{2}}$ 的取值很小；而分界点之前的节点为中心点，中心点普遍具有较大的核心值，这会导致 $C_{i-1}-C_i$ 与 $C_i-C_{\frac{i+n}{2}}$ 出现一个较大的差值，从而出现明显的核心跳变值。$C_i-C_{\frac{i+n}{2}}$ 中选择余下节点的中间节点 $C_{\frac{i+n}{2}}$ 而不是核心值最小的节点 C_n，以防止核心值过小的点影响核心跳变值的计算。同时，核心值过大的个别节点会导致过早地产生最大核心跳变值，这可能导致一个规模较大的网络只选出了很少的社区中心，甚至只选出了一个社区中心。为了防止这种情况，保证社区划分的质量，要求所选择的社区中心个数不少于网络中节点总数的 2%。社区中心的选取步骤如下：

Step1 将所有节点以核心值(C_i)降序排列，并初始化所有节点的核心跳变值 $CJ_i=0$，初始化 $i=|0.02\times n|$，n 为节点总数；

Step2 计算第 i 个节点的核心跳变值(CJ_i)，$i=i+1$；

Step3 如果 $i\leqslant\frac{m}{3}$，跳转 Step2，否则执行 Step4(m 为节点总数，为了保证社区的划分质量，划分的社区数量不超过节点总数的 $\frac{1}{3}$)；

Step4 遍历所有节点，找到最大的核心跳变值 CJ_k；

Step5 节点 k 为中心节点与普通节点的分界点，k 之前的节点被选为社区中心点。

3.3 重叠社区的划分

在现实生活的许多问题中，个体可能属于多个社区，因此重叠社区的划分更切合现实。TSDP 算法在选取了社区中心后，对网络中的其他节点进行划分时，采用与密度峰值聚类算法类似的策略，但密度峰值聚类算法只赋予每个节点一个标签，导致不能实现重叠社区的划分。因此，本文为 TSDP 算法引入重叠参数以实现重叠社区的划分，具体划分步骤如下：

Step1 分别为选取的社区中心点赋予不同的标签，并初始化重叠参数 $\gamma\in(0,1)$。

Step2 将所有节点按照局部密度大小进行降序排序，得到节点队列 $List_1$，再初始化一个空队列 $List_2$ 用于存储社区的划分结果。

Step3 如果 $List_1$ 不为空，$List_1$ 队首节点出队，并执行 Step4，否则算法结束。

Step4 如果该节点已经有社区标签，则表示该节点为社区中心(中心点)，将该节点入队 $List_2$，并跳转 Step3，否则执行 Step5。

Step5 按照式(6)和式(7)计算当前节点与 $List_2$ 中所有节点的距离(此时 $List_2$ 中节点的局部密度均大于当前节点的局部密度)，得到一个与 $List_2$ 等长的距离数组 $dis[\ ]$。$dis[\min]$为数组中的最小值，当前节点与 $dis[\min]$所对应的节点属于同一社区，遍历 dis 数组。若 $\frac{dis[i]-dis[\min]}{dis[\min]}<\gamma$，则当前节点也属于 $dis[i]$所对应的社区。如果当前节点的社区标签不止一个，则当前节点为重叠节点。

Step6 将当前节点入队 $List_2$，跳转 Step3。

经过上述步骤后，队列 $List_1$ 为空，而队 $List_2$ 列中存放了网络中的所有节点，并且这些节点都已被赋予了社区标签。具有相同社区标签的节点属于同一个社区，具有两个或两个以上社区标签的节点为重叠节点，该节点同时属于多个社区。

4 评价指标

为评价 TSDP 算法的有效性，本文采用了扩展的模块度(Extension of Modularity，EQ)、调整兰德系数(Adjust Rand Index，ARI)和归一化互信息(Normalized Mutual Information，NMI)3 个社区发现中常用的评价指标来衡量社区划分结果的模块性及准确性。

(1)EQ 评价指标

EQ 评价指标由 Shen 等[23]提出，是对模块度的扩展，用于评价重叠社区的模块性。EQ 的定义公式如下：

$$EQ=\frac{1}{2m}\sum_{l=1}^{C_l}\sum_{i\in l,j\in l}\frac{1}{O(i)O(j)}\left(A(i,j)-\frac{k(i)k(j)}{2m}\right) \tag{13}$$

其中，m 为网络中边的总数，C_l 表示社区 l，$O(i)$表示节点 i 所属社区中的节点数量，$k(i)$表示节点 i 的度。若存在边使得节点 i 与节点 j 直接相连，则 $A(i,j)$取值为 1，否则为 0。EQ 值越高，所划分社区的模块性就越好。

(2)ARI 评价指标

ARI 评价指标是在兰德系数(Rand Index，RI)的基础上由 Santos 等提出的，其增大了对划分错误点的惩罚力度[24]，因此 ARI 具有比 RI 更高的区分度。ARI 的计算公式如下：

$$ARI=\frac{\binom{n}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2-[(a+b)(a+c)+(c+d)(b+d)]} \tag{14}$$

$$\binom{n}{2}=a+b+c+d \tag{15}$$

其中，a 表示在真实社区结构中属于同一社区且在实验得到的社区结构中仍属于同一个社区的点对数；b 表示在真实社区结构中属于同一社区但在实验得到的社区结构中被划分到不同社区的点对数；c 表示在真实社区结构中不属于同一社区但在实验得到的社区结构中却被划分到同一个社区的点对数；d 表示在真实社区结构中不属于同一社区且在实验得到的社区结构中仍不属于同一个社区的点对数。

(3)NMI 评价指标

NMI 用于评价划分的社区结构与网络中真实的社区结构之间的相似性[25]。NMI 的计算式如下：

$$NMI(A,B)=\frac{-2\sum_{C_A}^{i=1}\sum_{C_B}^{j=1}\boldsymbol{C}_{ij}\cdot\log\left(\frac{\boldsymbol{C}_{ij}\cdot N}{\boldsymbol{C}_i\cdot\boldsymbol{C}_j}\right)}{\sum_{C_A}^{i=1}\boldsymbol{C}_i\cdot\log\left(\frac{\boldsymbol{C}_i}{N}\right)+\sum_{C_B}^{j=1}C_j\cdot\log\left(\frac{\boldsymbol{C}_j}{N}\right)} \tag{16}$$

其中，A 与 B 分别代表两种不同的社区划分结构；$\boldsymbol{C}$ 为混淆矩阵，$\boldsymbol{C}_{ij}$ 表示 A 中第 i 个社区与 B 中第 j 个社区两者之间公共节点的个数；N 为网络中节点的总个数；$\boldsymbol{C}_i$ 表示 A 中第 i 个社区中节点的个数；$\boldsymbol{C}_j$ 表示 B 中第 j 个社区中节点的个数。NMI 值介于 0 到 1 之间，值越大，算法划分的社区结构与真实社区结构的差别就越小。

5 实验结果与分析

本文算法 TSDP 用 Java 编程实现。程序运行的计算机软硬件环境为：Window10 操作系统，Intel Core i7-7500U

2.70GHz CPU，8GB 内存。基于人工网络数据集和真实网络数据集来验证 TSDP 算法的有效性。

(1)人工网络数据集

本文实验所使用的人工网络数据集为 LFR 基准网络。LFR 基准网络是由 Lancichinetti 等[26]提出的，可用于生成具有良好定义结构的人工网络，并且能够很好地模拟真实网络。LFR 提供了生成网络的不同参数，用户可以通过调整这些参数来指定生成的网络具有某些属性，如网络规模 N、节点的平均度 k、节点的最大度 $maxk$、最小社区规模 $minc$、最大社区规模 $maxc$、混合参数 μ 等。混合参数 μ 介于 0 到 1 之间，μ 值越大，网络的结构就越复杂，网络中的社区也就越难以被发现。调整参数 N，k，$maxk$，$minc$，$maxc$ 和 μ，可以生成不同规模和混合度的人工网络。

(2)真实网络数据集

一些学者和机构收集了真实网络数据以进行社区发现研究，我们从网站上下载了 6 个公共数据集[1)]来测试算法的性能，其信息如表 1 所列。

表 1　真实网络数据集信息

Table 1　Information of real network data sets

数据集	节点数	边数
Dolphins	62	159
Football	115	613
Karate	34	78
Lesmis	77	254
Polbooks	105	441
Power	4941	6594

Dolphins 网络描述的是新西兰 62 只宽吻海豚的生活习性，网络中的边表示对应的两只海豚经常一起活动。Football 网络描述的是美国某次足球联赛中 115 支大学生代表队的比赛情况，115 支大学生代表队被分为 12 个联盟，联盟内部先进行小组赛，之后联盟之间进行比赛，联盟内部的比赛多于联盟之间的比赛。Karate 网络描述的是美国某大学空手道俱乐部中成员之间的友谊关系。Lesmis 网络描述的是雨果的小说《悲惨世界》中的人物关系。Polbooks 网络描述的是 2004 年美国总统选举期间，亚马逊网站在线出售的关于美国政治书籍的销售网络。Power 网络是美国西部各州电网的拓扑结构。

5.1　人工网络实验结果及分析

本文从时间性能、划分社区结构的模块度以及划分社区结果的准确性 3 个方面来衡量 TSDP 算法在人工网络中的有效性，并将其与其他算法进行比较。社区的模块度采用 EQ 指标，社区结果的准确性采用 ARI 和 NMI 指标。

(1)时间性能分析

利用 TSDP 算法检测不同规模的人工网络中的社区结构，记录划分不同规模的人工网络所需的时间，并将其与其他算法进行比较。OCDDP 是 Bai 等提出的一种密度峰值的社区发现算法[27]，该算法通过迭代邻接矩阵的方式得到节点的距离矩阵。LDP 是 Huang 等提出的一种基于边的密度峰值的社区发现算法[28]，其利用边的相似度来计算距离。TSDP，OCDDP 和 LDP 3 种算法划分不同规模的人工网络所需要的时间如表 2 所列。

表 2　各算法划分不同规模网络所消耗时间的比较

Table 2　Time consumption of different algorithms *w.r.t.* partition different networks

(单位：ms)

网络规模	TSDP	OCDDP	LDP
50	30	10	1950
100	80	320	7097
200	200	853	33279
300	470	1650	71103
400	850	2851	163758
500	1371	4291	253024

算法消耗时间的对数函数($\log_{10}$)折线图如图 3 所示。图中的横轴为网络规模，纵轴为不同网络规模下相应算法计算时间的对数值。

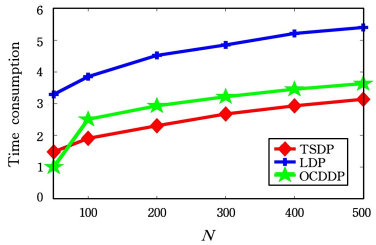


图 3　算法消耗时间的对数图

Fig. 3　Time consumption of different algorithms for different network sizes

由表 2 和图 3 可以看出，在划分同等规模的人工网络时，TSDP 算法所消耗的时间最短，LDP 算法所消耗的时间最长。在复杂网络中，边的数量一般远多于节点数量，而 LDP 算法是一种基于边的社区发现算法，其需要计算网络中任意两条边之间的相似度，且边之间的相似度计算比较复杂，这导致 LDP 算法具有较大的时间复杂度。OCDDP 算法通过迭代邻接矩阵的方式得到距离矩阵，相比 LDP 算法，其时间复杂度小得多；但随着网络中节点数量的增多，邻接矩阵的规模迅速增大，OCDDP 算法同样需要消耗较多的时间。TSDP 算法则通过节点的度来计算节点的局部密度，且节点间相似度的计算也比较简单，因此 TSDP 算法具有相对较小的时间复杂度，即使网络规模不断增大，其依旧具有较好的时间性能。但当网络规模很小时，网络的邻接矩阵规模也很小，OCDDP 算法的时间性能可能优于 TSDP 算法，例如图 3 中，当网络规模为 50 时，OCDDP 算法所消耗的时间最短。总体而言，TSDP 算法的时间性能最优。

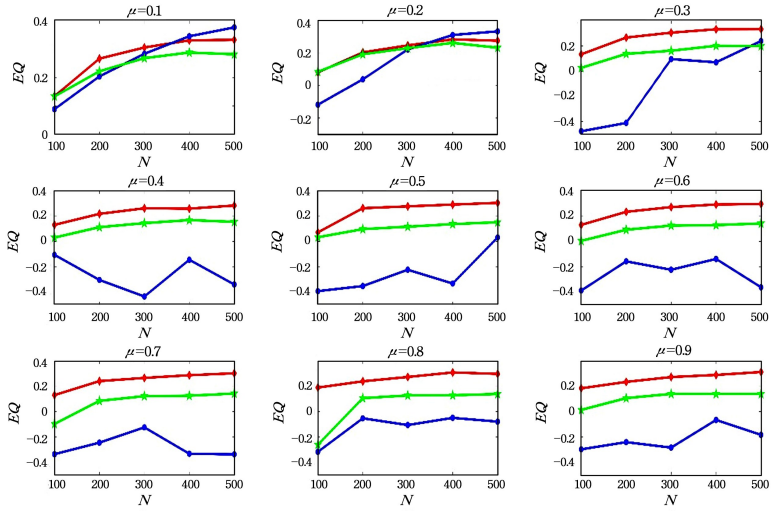
(2)EQ 指标的比较与分析

给定参数 $k=5$，$maxk=15$，$minc=20$，$maxc=50$，通过调整 N 值和 μ 值生成不同规模和混合度的人工网络，并比较 TSDP，LDP 和 OCDDP 在人工网络中的划分效果。各算法在不同规模和混合度的人工网络中划分社区结果的模块度如图 4 所示。

由图 4 可知，当 μ 取 0.1～0.2 时，生成的人工网络结构简单，TSDP 算法比 OCDDP 算法所划分的社区具有更高的模块度；LDP 算法在结构简单的人工网络中能够得到模块度较高的社区结构，例如在节点规模为 500 时 LDP 算法优于 TSDP 算法。当 μ 取 0.3～0.9 时，TSDP 算法划分得到的社区

1) http://www-personal.umich.edu/~mejn/netdata

结构具备的模块性明显优于其他算法。同时也可以看出，相对于其他算法，TSDP 算法在混合度 μ 值不断增大时，依旧可以表现出较为稳定的效果。因此，TSDP 算法不仅性能稳定，而且划分的结果更优。



注：表示 TSDP 算法，表示 LDP 算法，表示 OCDDP 算法

图 4 人工网络的 EQ 值比较

Fig. 4 Comparison of EQ value in artificial networks

为了更加直观地比较各算法在划分人工网络时，社区结构所具备的 EQ 值与混合度参数 μ 之间的关系，图 5 给出了各算法在不同混合度参数下划分人工网络的 EQ 均值。

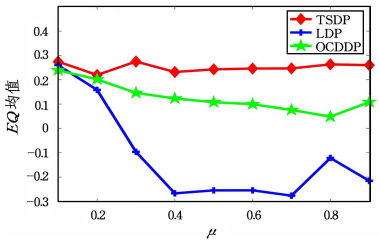


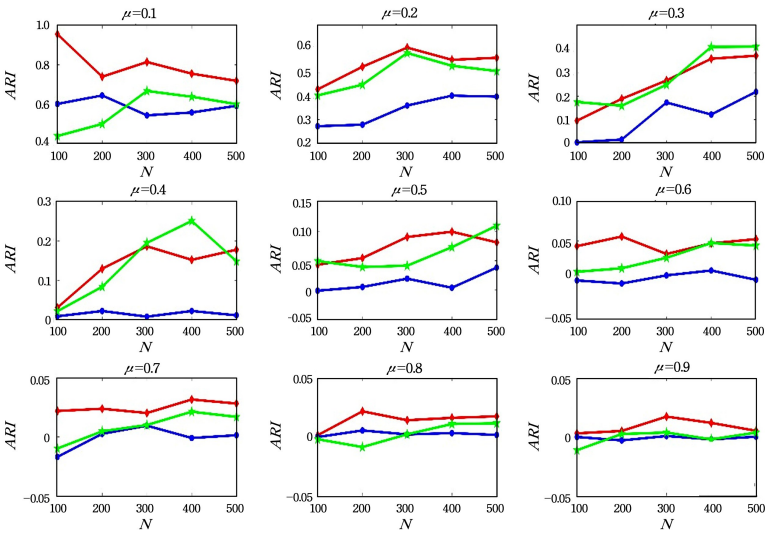
图 5 人工网络的 EQ 均值比较

Fig. 5 Average EQ values comparison in artificial networks

由图 5 可以看出，随着 μ 值的不断增大，TSDP 算法划分社区结构的 EQ 值只有小幅波动。可见在 EQ 评价指标下，TSDP 算法的划分效果基本不受 μ 值的影响。OCDDP 算法得到的社区结构的 EQ 均值随着 μ 值的增大而呈现出较为明显的下降趋势，相对于 TSDP 算法来说，其对 μ 值更加敏感。但总的来说，OCDDP 算法受 μ 值的影响较小。而 LDP 算法是基于边的密度峰值社区发现算法，μ 值增大时网络结构会变得复杂，而网络结构的复杂主要体现在边结构的复杂，这就导致基于边结构的 LDP 算法对 μ 值十分敏感，随着 μ 值的增大，LDP 算法划分人工网络所得的社区结构的 EQ 均值迅速下降。因此，以 EQ 作为评价指标时，TSDP 算法不仅具有较好的稳定性，同时划分得到的社区结构也最优。

(3)ARI 指标的比较与分析

LFR 生成的人工网络数据会给出该网络的真实社区划分结构，因此可以衡量算法划分的社区与真实划分结构之间的相似度，使用 ARI 评价指标来衡量算法划分社区结构的准确性。各算法在不同规模和混合度参数下的人工网络中划分社区结构的 ARI 值如图 6 所示。



注：表示 TSDP 算法，表示 LDP 算法，表示 OCDDP 算法

图 6 人工网络的 ARI 值比较

Fig. 6 Comparison of ARI value in artificial networks

图 6 表明，当 μ 取 0.1～0.2 或 0.6～0.9 时，TSDP 算法划分得到的社区 ARI 值高于 OCDDP 和 LDP 得到的社区 ARI 值。当 μ 取 0.3～0.5 时，多数情况下 TSDP 算法划分的社区具有最高的 ARI 值，但有时也会出现 OCDDP 算法划分的社区具有的 ARI 值高于 TSDP 算法的情况，这是由于生成的人工网络具有随机性，TSDP 算法无法保证在所有的网络中表现均最优。同时，从图 6 中还可以看出，LDP 算法划分所得到的社区具有的 ARI 值在多数情况下是较差的，这是由于 LDP 算法本身是一种基于边的密度峰值算法，边的复杂度远远大于节点的复杂度，从而导致 LDP 算法划分社区具有的 ARI 值不高。总的来说，在使用 ARI 作为评价指标时，TSDP 算法能够表现出较为明显的优势。

为了更直观地观察使用 ARI 作为评价指标时混合度参数对各算法划分结构的影响，图 7 给出了各算法在不同混合度参数下划分人工网络的 ARI 均值。

图 7 表明，当 μ 值小于或等于 0.2 时，TSDP 算法所划分的不同规模的人工网络得到的社区具有明显高于 OCDDP 和 LDP 算法的 ARI 均值。当 μ 值大于 0.2 且小于 0.9 时，LDP 算法划分得到的社区具有最低的 ARI 值，表明 LDP 算法对 μ 的敏感度大于 TSDP 算法和 OCDDP 算法，而 TSDP 算法和 OCDDP 算法划分的社区具有的 ARI 值差别不大，因此当 μ 值大于0.2 时，ARI 指标对于 TSDP 算法和 OCDDP 算法并没有很好的区分度，但可以看出 TSDP 算法相对于 OCDDP 算法来说具有一定优势。

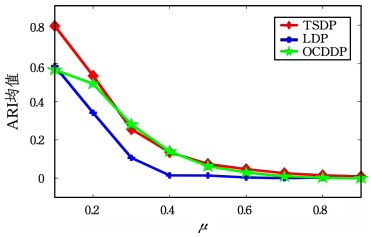
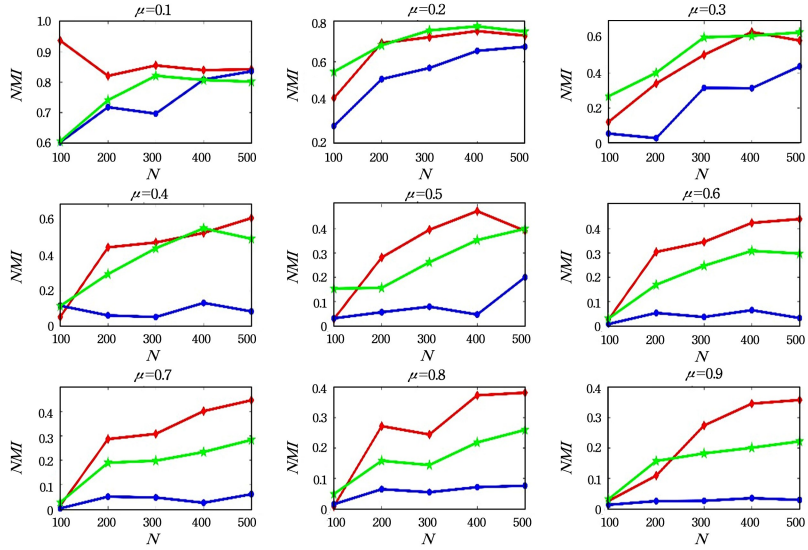


图 7　人工网络的 ARI 均值比较

Fig. 7　Average ARI values comparison in artificial networks

(4)NMI 指标的比较与分析

由于 ARI 指标对于 TSDP 算法和 OCDDP 算法并没有很好的区分度，因此引入 NMI 评价指标来进一步比较各算法划分结果的准确性。各算法在不同规模和混合度参数下的人工网络中划分社区结构的 NMI 值如图 8 所示。



注：表示 TSDP 算法，表示 LDP 算法，表示 OCDDP 算法

图 8　人工网络的 NMI 值比较

Fig. 8　Comparison of NMI values in artificial networks

图 8 表明，当 μ 取 0.1～0.5 时，由于生成人工网络具有随机性，有时 OCDDP 划分结构具有的 NMI 高，有时 TSDP 划分结构具有的 NMI 高，但随着 μ 的增大，TSDP 算法表现较优的情况更多，而 LDP 算法划分的 NMI 值较低。当 μ 取 0.6～0.9 时，基本 TSDP 算法划分结果的 NMI 值高于 OCDDP；并且当网络规模增大时，TSDP 算法具有明显的优势。LDP 算法由于受 μ 值影响较大，随着 μ 值的增大，其得到的划分结构的 NMI 值最低。

为了更加直观地观察在 NMI 评价指标下混合度参数对 TSDP 算法及其他算法的影响，图 9 给出了各算法在不同混合度参数下划分人工网络的 NMI 均值。

图 9 表明，μ 取 0.1～0.9 时，TSDP 算法划分不同规模所得到的社区具有的 NMI 均值一般高于 OCDDP 算法和 LDP 算法。但当 μ 取 0.2 和 0.3 时，OCDDP 算法优于 TSDP 算法，这是由于生成的人工网络存在随机性，TSDP 算法无法保证在所有的网络中均表现最优。但总体来说，TSDP 算法划分结构的准确性更高。同时，也可以看出 TSDP 算法受 μ 值影响最小，LDP 算法受 μ 值影响最大，NMI 指标可以很好地区分 3 种算法。

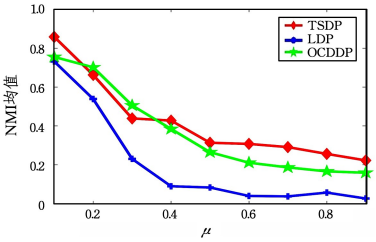


图 9　人工网络的 NMI 均值比较

Fig. 9　Average NMI values comparison in artificial networks

人工网络数据实验表明，TSDP 算法能够快速得到网络的社区划分结构，同时对于不同规模的人工网络，TSDP 算法能划分得到具有较高 EQ 值和更接近于真实划分结构的社区结构。随着 μ 值的增大，TSDP 算法相对于其他算法能得到更好的社区划分结构。

5.2 真实网络实验结果及分析

目前已有许多真实的公开网络数据集用于社区发现研究，为了验证 TSDP 算法在真实网络中的效果，我们在 Dolphins 和 Football 等真实网络中验证 TSDP 算法的有效性。由于一些真实网络数据并没有真实社区结构，因此采用 EQ 评价指标来衡量 TSDP 算法划分真实网络结果的质量，并将其与其他算法进行比较。

(1)TSDP 算法的划分结果

TSDP 算法在真实网络中划分社区结构的模块度以及在各网络中划分社区的数量如表 3 所列。

表 3 真实网络数据集的划分结果

Table 3 Partition results of real network data sets

数据集	*EQ*	社区数量
Dolphins	0.5126	4
Football	0.6139	12
Karate	0.4161	4
Lesmis	0.5556	6
Polbooks	0.5034	10
Power	0.9282	217

(2)与其他算法的比较

本文将 TSDP 算法分别与 OCDDP，SpeakEasy 和 SLPA 算法进行了比较。SpeakEasy 是 Chris 等提出的一种稳健的标签传播算法[29]，不同于传统标签的随机更新策略，其根据邻居节点标签和这些标签在整个网络中的受欢迎度来更新当前节点标签。SLPA 算法是由 Xie 等提出的重叠标签传播算法[30]，该算法的节点根据动态交互规则更新标签，能够划分重叠社区。TSDP 算法与上述算法的比较结果如表 4 所列。

表 4 在真实网络数据集上与其他算法的 EQ 值比较

Table 4 Comparison of EQ on real network data sets

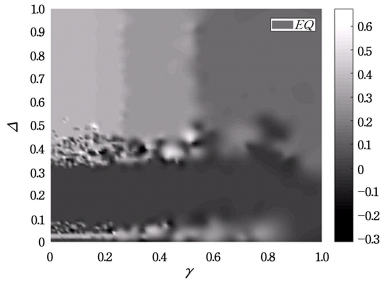
数据集	TSDP	OCDDP	SpeakEasy	SLPA
Dolphins	0.5126	**0.5202**	0.5017	0.4833
Football	**0.6139**	0.5958	0.5811	0.5984
Karate	0.4161	0.4063	**0.4198**	0.3924
Lesmis	**0.5556**	0.5310	0.5479	0.5270
Polbooks	0.5034	**0.5095**	0.4973	0.4831
Power	**0.9282**	0.8848	0.6745	0.6563
AVG	**0.5883**	0.5746	0.5371	0.5234

表 4 中，粗体数据表示算法在当前真实网络中所划分得到的社区结构具有的 EQ 值最高。从表中可以看出，与 OCDDP，SpeakEasy 和 SLPA 在 Dolphins 等 6 个真实网络中划分的社区结果相比，TSDP 算法在 Football 网络、Lesmis 网络和 Power 网络中具有明显优势。虽然在 Karate 网络中，SpeakEasy 算法划分的社区的模块度优于 TSDP 算法，在 Dolphins 网络和 Polbooks 网络中，OCDDP 算法划分的社区模块度优于 TSDP 算法，但差值并不大；并且，TSDP 算法划分的 Dolphins 等 6 个真实网络所得到社区结构的模块度均值(AVG)高于 OCDDP，SpeakEasy 和 SLPA。综上所述，TSDP 算法在真实网络中能够划分得到具有较高模块度的社区结构。

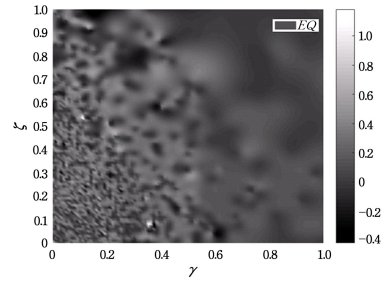
5.3 参数对算法的影响

TSDP 算法中包含 3 个参数：密度衰减系数 ζ、距离离散参数 Δ 以及重叠度参数 γ。ζ 决定了网络中各个节点的局部密度值；Δ 用于离散化距离值，其对社区中心的选择有较大影响；γ 决定了算法划分社区结果的重叠程度。3 个参数对算法划分的社区结构有较大影响，且对于不同的数据集，3 个参数的取值也有较大差异。因此，对于不同的网络，需要调整参数的取值以取得具有较高模块度的社区划分结构。为了检测参数对算法划分结构的影响，通过实验给出了在两个真实网络即 Karate 和 Power 中参数与社区划分结构 EQ 之间的关系图。

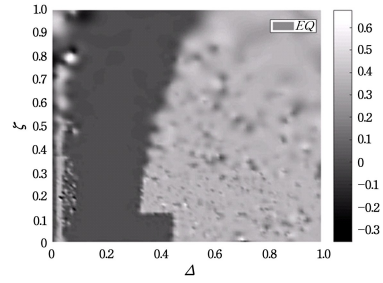
在 Karate 网络中，各参数与社区划分结构 EQ 的关系如图 10 所示。



(a)Δ 和 γ 与 EQ 的关系



(b)ζ 和 γ 与 EQ 的关系



(c)ζ 和 Δ 与 EQ 的关系

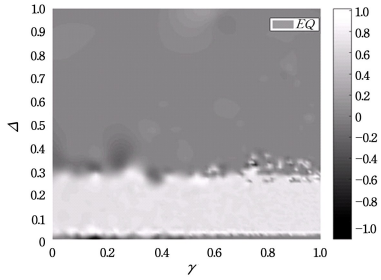
图 10 Karate 网络中各参数与 EQ 的关系图

Fig. 10 Relations between parameters and EQ in Karate network

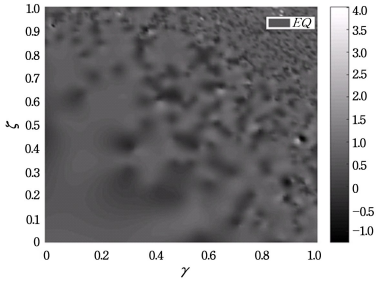
图 10 中，颜色越浅或越亮的区域表示参数取值在此范围内所划分的社区结构具有的 EQ 值越高，颜色越深或越暗的区域表示参数取值在此范围内所划分的社区结构具有的 EQ 值越低。图 10(a)表明，在 Karate 网络中，当 Δ 取值为 0.45 左右且 γ 取值为 0.5 左右，或者 Δ 取值为 0.05 左右且 γ 取值为 0.4 左右时，TSDP 算法能划分得到较高的 EQ 值；而当 Δ 取值在 0.1 到 0.3 之间时，TSDP 算法划分的社区结构的 EQ 值较差。从图 10(b)中可以看出，ζ 取值为 0.1 左右且 γ 取值为 0.4 左右时，TSDP 算法取得了极大的 EQ 值。从图 10(c)

中可以看出，Δ 的取值与 EQ 的关系存在一条明显的分界线，Δ 取值大于 0.4 时能得到较高的 EQ 值，同时当 Δ 取值为 0.05 左右且 ζ 取值为 0.8 左右时，也能得到具有较高 EQ 值的社区结构。因此，在 Karate 网络中，重叠度参数 γ 取值为 0.4 左右、距离离散参数 Δ 取值为 0.05 左右、密度衰减系数 ζ 取值为 0.1 左右时，能够取得具有较高模块度的社区划分结构。

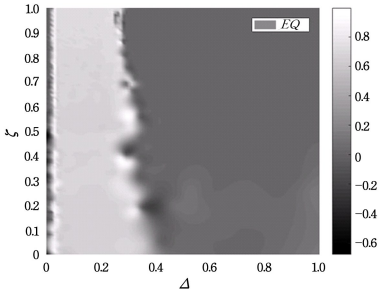
在 Power 网络中，各参数与社区划分结构 EQ 的关系如图 11 所示。



(a)Δ 和 γ 与 EQ 的关系



(b)ζ 和 γ 与 EQ 的关系



(c)ζ 和 Δ 与 EQ 的关系

图 11　Power 网络中各参数与 EQ 的关系图

Fig. 11　Relations between parameters and EQ in Power network

由图 11(a)可以看出，在 Power 网络中，Δ 的取值为 0 到 0.3 之间时划分的社区结构具有较高的 EQ 值，且 Δ 取值为 0.05 左右或 0.25 左右时取得 EQ 极大值。从图 11(b)可以看出，γ 取值为 0.9 左右时，EQ 取得最大值。从图 11(c)可以看出，Δ 取值为 0.3 左右且 ζ 取值为 0.6 左右或 0.4 左右时，算法在 Power 网络中能划分得到具有较大 EQ 值的社区结构。另外，图 11(a)和图 11(c)均存在明显的分界线，这表明社区结构的 EQ 值与 Δ 取值有较大关联，当 Δ 取值在 0.05 到 0.3 时，所划分的社区结构普遍具有较高的 EQ 值。总的来说，在 Power 网络中，重叠度参数 γ 取值为 0.9 左右、距离离散参数 Δ 取值为 0.05 左右或 0.3 左右、密度衰减系数 ζ 取值为 0.4 左右或 0.6 左右时，能够取得具有较高模块度的社区划分结构。

综上所述，参数的不同取值对社区的划分结构具有较大影响，因此对于不同的网络，需要适当调整参数以得到较优的社区划分结构。

结束语　本文基于网络拓扑结构提出了一种新的密度峰值社区发现算法——TSDP 算法。TSDP 算法不需要通过节点间的距离来计算局部密度，而是通过节点的度来衡量节点的局部密度，这种方式不仅能较好地衡量各个节点的局部密度，而且计算的复杂度较低；同时，通过节点间的相似度来计算节点间的距离，并通过离散化函数将节点的值进行离散化处理，以便选取社区中心点。为了更精确地选取社区中心，防止网络中规模较小的社区结构被大规模社区吞并，TSDP 算法定义了核心跳变值来帮助选取社区中心。实验采用了人工网络数据和真实网络数据来验证算法的有效性，应用 EQ 指标、ARI 指标和 NMI 指标来衡量 TSDP 算法划分结果的质量，并将其与多个近年来提出的社区发现算法进行比较。实验结果表明：TSDP 算法划分得到的社区结构不仅具有较高的模块度，而且更接近于网络的真实划分结构；并且在划分同等规模的网络时，TSDP 算法所消耗的时间较短，具有较好的时间性能。

已有的大多社区发现算法都是将复杂网络划分为非重叠社区或者重叠社区，但随着对复杂网络研究的不断深入，人们发现社区中还可能包含“小世界”，并且社区中存在着更为复杂的层次结构，层次社区发现算法成为研究的热点之一。另外，在现实世界中，个体往往包含内部属性，而传统的社区发现算法仅仅考虑节点间的连接关系，忽略了节点内部属性在社区划分中的重要性，这导致算法所划分的社区可能与真实情况存在较大差异。因此，可以在社区发现算法中加入节点的内部属性，以使算法的划分结果更符合实际。未来将结合网络中各节点的内部属性进一步研究 TSDP 算法，并将其扩展为层次结构的社区划分算法。

参 考 文 献

[1] FORTUNATO S. Community detection in graphs[J]. Physics Reports-Review Section of Physics Letters, 2010, 486(3): 75-174.

[2] NEWMAN M E J. The structure and function of complex networks[J]. Siam Review, 2003, 45(2): 167-256.

[3] YU Z D, YU H Q. Micro-Blog user recommendation based on community discovery and topic analysis[J]. Journal of East China University of Science and Technology(Natural Science Edition), 2014, 40(6): 763-768. (in Chinese)
余紫丹，虞慧群. 基于社区发现及主题分析的微博用户推荐[J]. 华东理工大学学报(自然科学版)，2014，40(6)：763-768.

[4] MA F M, WANG G. Method for commodity recommendation based on user community[J]. Computer & Digital Engineering, 2013, 41(8): 1354-1356. (in Chinese)
麻风梅，王刚. 基于用户社区的商品推荐方法[J]. 计算机与数字工程，2013，41(8)：1354-1356.

[5] PAN W F, LI B, SHAO B, et al. Service classification and recommendation based on software networks[J]. Chinese Journal

of Computers,2011,34(12):2355-2369. (in Chinese)
潘伟丰,李兵,邵波,等. 基于软件网络的服务自动分类和推荐方法研究[J]. 计算机学报,2011,34(12):2355-2369.

[6] GIRVAN M,NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of National Academy of Science,2002,99(12):7821-7826.

[7] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E Statistical Nonlinear and Soft Matter Physics,2004,69(6):066133.

[8] DERENYI I,PALLA G,VICSEK T. Clique percolation in random networks[J]. Physical review letters, 2005, 94(16): 160202.

[9] CLAUSET A,NEWMAN M E J,MOORE C. Finding community structure in very large networks[J]. Physical Review E Statistical Nonlinear and Soft Matter Physics,2004,70(6):066111.

[10] DANON L,DIAZGUILERA A,ARENAS A. Effect of size heterogeneity on community identification in complex networks[J]. Journal of Statistical Mechanics:Theory and Experiment,2006,2006(11):P11010.

[11] ZHU X,GHAHRAMANI Z. Learning from labeled and unlabeled data[J]. Technology Report,2002,3175(2004):237-244.

[12] LIU S C,ZHU F X,GAN L. A label-propagation-probability-based algorithm for overlapping community detection[J]. Chinese Journal of Computers,2016,39(4):717-729. (in Chinese)
刘世超,朱福喜,甘琳. 基于标签传播概率的重叠社区发现算法[J]. 计算机学报,2016,39(4):717-729.

[13] ZHANG X K,REN J,SONG C,et al. Label propagation algorithm for community detection based on node importance and label influence[J]. Physics Letters A,2017,381(33):2691-2698.

[14] LI L,JIAO L,ZHAO J,et al. Quantum-behaved discrete multi-objective particle swarm optimization for complex network clustering[J]. Pattern Recognition,2017,63:1-14.

[15] LIU Q,ZHOU B,LI S,et al. Community detection utilizing a novel multi-swarm fruit fly optimization algorithm with hill-climbing strategy[J]. Arabian Journal for Science & Engineering,2016,41(3):807-828.

[16] JIANG S Y,YANG B H,WANG L X. An adaptive dynamic community detection algorithm based on incremental spectral clustering[J]. Acta Automatica Sinica, 2015, 41(12): 2017-2025. (in Chinese)
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[J]. 自动化学报,2015,41(12):2017-2025.

[17] ZHOU X,LIU Y,WANG J,et al. A density based link clustering algorithm for overlapping community detection in networks [J]. Physica A Statistical Mechanics & Its Applications,2017,486(2017):65-78.

[18] RODRIGUEZ A,LAIO A. Clustering by fast search and find of density peaks[J]. Science,2014,344(6191):1492-1496.

[19] YANG J,WANG G Y,PANG Z L. Relative researches of clustering by fast search and find of density peaks[J]. Journal of Nanjing University(Natural Science), 2017, 53(4): 791-801. (in Chinese)
杨洁,王国胤,庞紫玲. 密度峰值聚类相关问题的研究[J]. 南京大学学报(自然科学版),2017,53(4):791-801.

[20] SHI X H,FENG G X,LI M,et al. Overlapping community detection method based on density peaks[J]. Journal of Jilin University(Engineering and Technology Edition),2017,47(1):242-248. (in Chinese)
时小虎,冯国香,李牧,等. 基于密度峰值的重叠社区发现算法[J]. 吉林大学(工学版),2017,47(1):242-248.

[21] HUANG L,LI Y,WANG G S,et al. Community detection method based on vertex distance and clustering of density peaks [J]. Journal of Jilin University(Engineering and Technology Edition),2016,46(6):2042-2051. (in Chinese)
黄岚,李玉,王贵参,等. 基于点距离和密度峰值聚类的社区发现方法[J]. 吉林大学学报(工学版),2016,46(6):2042-2051.

[22] HENNIG C,HAUSDORF B. Design of dissimilarity measures: A new dissimilarity between species distribution areas[M]// Data Science And Classification. Berlin Heidelberg:Springer,2006: 29-37.

[23] SHEN H,CHENG X,CAI K,et al. Detect overlapping and hierarchical community structure in networks[J]. Physica A Statistical Mechanics & Its Applications,2009,388(8):1706-1712.

[24] SANTOS J M,EMBRECHTS M. On the use of the adjusted rand index as a metric for evaluating supervised classification[C]// Proceeding of the 19th International Conference on Artificial Neural Networks. Berlin Heidelberg:Springer,2009:175-184.

[25] DANON L,DIAZGUILERA A,DUCH J,et al. Comparing community structure identification[J]. Journal of Statistical Mechanics:Theory and Experiment,2005,2005(9):P09008.

[26] LANCICHINETTI A,FORTUNATO S. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(1): 016118.

[27] BAI X,YANG P,SHI X. An overlapping community detection algorithm based on density peaks[J]. Neurocomputing, 2017, 226:7-15.

[28] HUANG L,WANG G,WANG Y,et al. A link density clustering algorithm based on automatically selecting density peaks for overlapping community detection[J]. International Journal of Modern Physics B,2016,30(24):1650167.

[29] GAITERI C,CHEN M M,SZYMANSKI K,et al. Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering[J]. Scientific Reports,2015,5:16361.

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