View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

DOI:10.11784/tdxbz201809051

基于边信任度的混合参数自适应重叠社区发现算法

汪 清[1]，顾春妹[1]，赵建军[1]，崔 鑫[1]，洪文兴[2]，徐文静[2]
（1. 天津大学自动化与信息工程学院，天津 300072；2. 厦门大学航空航天学院，厦门 361005）

摘 要：网络中的社区结构有助于简化网络拓扑结构分析，揭示系统内部的规律，能够为信息推荐和信息传播控制提供有力的支撑．网络重叠社区结构与真实生活更加接近，但其分析较非重叠社区结构更加困难．因此，针对重叠社区发现问题，在对网络的边进行峰值聚类的基础上提出了一种基于边信任度的混合参数的自适应重叠社区发现算法．定义了网络边的邻居边集合及与其邻居边之间的信任度函数，通过信息传递获取边的总信息量，并且基于此引入混合参数的概念．基于 k-means 算法使用混合参数对网络中的边进行聚类，即将网络中的边划分为核心边集与非核心边集，每个核心边作为一个聚类中心．根据非核心边到核心边的距离将所有非核心边划分至距离其最近的聚类中心所在社区．再根据网络中边与节点的关系实现重叠节点发现，最终实现重叠社区的发现．该算法的优点是每条边通过独立地完成信息扩散找到社区的结构，相比于传统的峰值聚类算法，不需要人为设置相关参数，实现重叠社区的自适应发现．为验证算法的可行性，对算法复杂度进行了分析，并且使用两种社区划分评价指标——标准化互信息和模块度，分别在人工数据集及 6 种真实数据集上进行实验，通过与其他算法进行对比分析，实验结果表明该算法更具可行性和有效性．

关键词：峰值聚类；边信任度；混合参数；重叠社区发现；自适应算法
中图分类号：TN915.01 文献标志码：A 文章编号：0493-2137(2019)06-0618-07

Adaptive Overlapping Community Detection Algorithm Based on Mixing Parameter with the Trust Degree of Edge

Wang Qing[1]，Gu Chunmei[1]，Zhao Jianjun[1]，Cui Xin[1]，Hong Wenxing[2]，Xu Wenjing[2]
（1. School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China；
2. School of Aerospace，Xiamen University，Xiamen 361005，China）

Abstract：The community structure in a network simplifies the analysis of the network topology，reveals the internal rules of the system，and provides strong support for information recommendation and information dissemination control. The overlapping community structure of the network is closer to real-life scenario，but its analysis is more difficult than the non-overlapping community. Therefore，to solve the overlapping community detection，based on the peak clustering，an adaptive overlapping community detection algorithm based on the mixing parameter with the trust degree of edge is proposed. In this study，the neighbor edge set of the network and the trust function between the edge and its neighbors are defined，and the total information of the edge is obtained through information transfer. Based on this concept，the concept of mixing parameters is introduced. Then，based on the k-means algorithm，clustering is performed using the mixed parameter，i.e.，the edges in the network are divided into a core edge set and a non-core edge set，and each core edge acts as a clustering center. According to the distance from the non-core edge to the core edge，the non-core edges are divided into the community of the nearest cluster center. According to the relation between edges and nodes in the network，overlapping node discovery is achieved. Ultimately the overlapping communi-

收稿日期：2018-09-17；修回日期：2018-10-15.
作者简介：汪 清（1982— ），博士，副教授，wangq@tju.edu.cn.
通信作者：洪文兴，hwx@xmu.edu.cn.
基金项目：国家自然科学基金资助项目(61871282)；福建省科技计划资助项目(2018H0035)；厦门市科技计划资助项目(3502Z20183011).
Supported by the National Natural Science Foundation of China(No. 61871282)，the Science and Technology Program of Fujian，China (No. 2018H0035)，the Science and Technology Program of Xiamen，China(No. 3502Z20183011).

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ties are detected. The advantage of this algorithm is that each edge finds the structure of the community by independently completing information transfer. Moreover，compared to the traditional peak clustering algorithm，the proposed algorithm does not need to set parameters；therefore，adaptive detection of overlapping communities is achieved. To verify the feasibility of our algorithm，the complexity of the algorithm is analyzed. The two evaluation indices of the community detection，normalized mutual information and modularity，are used to experiment on the artificial dataset and the six real datasets respectively. In comparison to other algorithms，the experimental results show that the proposed algorithm is more feasible and effective.

Keywords：peak clustering；trust degree of edge；mixing parameter；overlapping community detection；adaptive algorithm

网络与我们的日常生活息息相关，如交通系统、通信系统和社交网络等. 众所周知，社区在揭示网络隐藏结构方面发挥着重要作用，并且现实问题中节点可能属于多个社区，这些节点称为重叠节点. 如果社区包含重叠节点，则称该社区为重叠社区. 重叠社区在现实世界中很常见，如一个人可能对几个主题感兴趣并加入多个社交圈，即重叠的社交网络. 网络的重叠社区结构很好地反映了真实的网络结构. 研究重叠社区结构具有重要的现实意义，不仅有助于揭示复杂系统的内部规则，理解拓扑结构，而且能简化网络分析. 但是，由于网络中的节点具有多重身份，使得获得准确的社区结构的难度加大. 因此，重叠社区发现已成为该领域中广泛研究的问题，其主要目的是识别由内部密集连接的节点和外部稀疏连接的节点组成的社区.

Chen 等[1]于 2016 年基于密度峰值聚类和节点局部信息提出一种线性复杂度的社区发现算法. Huang 等[2]提出了一种基于密度峰值法的重叠检测算法，并获得了良好的性能. Zhou 等[3]利用蚁群算法和标签传播检测重叠社区. 该算法首先初始化节点标签和蚂蚁的位置，再按照预先设置的概率进行随机游走并更新节点的标签. 当满足条件后，对网络节点所含的标签序列进行处理并为节点分配标签，从而完成网络的重叠社区发现. 上述社区发现算法本质上是对网络中的节点进行分类，但是仅仅对节点集进行划分并不能获得网络的重叠社区. 因此，相关学者将网络中的边当作数据对象进行聚类研究，即将边分配到不同的社区完成社区发现. Ahn 等[4]提出的算法计算了每对边缘之间的相似度，然后使用具有相似性的层次聚类来确定边缘属性由于是层次聚类，因此算法可得到不同分辨率下的社区结构. Shi 等[5]基于遗传算法对网络中的边进行聚类，针对网络边定义网络社区结构的基因表达以及交叉和变异算子，从而得到网络的重叠社区. Zhang 等[6]在标签传播算法(label propagation algorithm，LPA)的基础上引入边缘聚类系数用于更新标签，而不是随机地进行邻居节点标签的更新，有效地抑制了标签的随机传播，但该算法仅能用于非重叠社区结构及小规模重叠社区的网络，并且不能达到自适应.

本文从边的角度出发，利用网络比边与点之间的关系，提出了一种基于边信任度的混合参数(mixing parameter with trust degree of edge clustering，MPTD-EC)自适应重叠社区发现算法. 该算法自动选择核心边，引入边之间信息传递，利用边的总信息量代替峰值聚类中密度，不需要人为地设置截断距离.

本算法不需要人为设定参数，实现了自适应. 为了评估提出的方法，将其应用于合成和真实网络. 实验结果证明了该算法的有效性.

1　基于边信任度的混合参数自适应重叠社区发现算法

首先，定义邻居边之间的信任度函数，在网络中进行边信息传递获得边信息矩阵，在此基础上计算出边距离矩阵. 然后进行核心边的选取，根据边距离矩阵将非核心边进行分配，获得边社区，再将其转换为网络重叠社区.

1.1　信息扩散

网络表示为图 $G=(V,E)$，其中，$V=\{v_1,v_2,\cdots,v_n\}$ 是图 G 的节点集，$E=\{e_1,e_2,\cdots,e_n\}$ 是图 G 的边集.

利用信息扩散，可以获得每条边的信息量，利用边的信息量去标识边在网络中的重要程度.

由于本文使用边聚类完成重叠社区发现，故定义边 e_{ij} (表示节点 i 与节点 j 之间的连边)的邻居边集 $N(e_{ij})$ 为

$$N(e_{ij})=\{e_{ik}\in E\,|\,1\le k\le n,k\ne j\}\cup \{e_{vj}\in E\,|\,1\le k\le n,v\ne i\} \tag{1}$$

定义网络边信息矩阵 $\boldsymbol{S}^e_{m\times m}$，其中 s_i^e 表示边 e_i 的初始信息量，s_{ij}^e 表示以边 e_i 为信息源传递到边 e_j 的信息量.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

边之间的信息传递采用广度优先算法[7]，把作为信息源的边 e_i 信息量扩散到网络其他边，其所传递信息量的大小依据边之间的信任函数定义，信任度越大所传递信息量越大．邻居边之间的信任度函数定义为

$$C(i,j)=|N(e_i)\cap N(e_j)| \tag{2}$$

$$\alpha(i,j)=\frac{|C(i,j)|+1}{|N(e_i)|} \tag{3}$$

$$\beta(i,j)=\begin{cases}\dfrac{2|E(C(i,j))|}{(|C(i,j)|)(|C(i,j)|-1)} & |C(i,j)|\geq 2\\ 0 & |C(i,j)|<2\end{cases} \tag{4}$$

$$T^e(i,j)=\alpha(i,j)(\beta(i,j)+1) \tag{5}$$

式中：边 e_i 与边 e_j 的公共邻居边数量由 $C(i,j)$ 表示；$|E(C(i,j))|$ 表示边 e_i 与边 e_j 公共邻居边之间的节点数；$\beta(i,j)$ 是基于公共邻居边之间的连接密度定义的相似度因子．基于此定义邻居边之间的信任度函数 $T^e(i,j)$．边信息传递步骤如下．

步骤 1 初始化所有边的信息量为 1，即网络信息量矩阵 $\boldsymbol{S}^e_{m\times m}$ 为单位矩阵．

步骤 2 遍历网络中的边，每条边依次作为信息源，将其原始信息 1 用广度优先算法扩散到网络所有边，且当以该边为信息源时，并不考虑其他边所含信息量的大小．

步骤 3 以边 e_i 信息源，其传递到邻居边 e_j 的信息量为 $s^e_{ij}=1\times T^e(i,j)$，邻居边 e_j 把从信息源获得的信息量传递到其邻居边 e_k，边 e_k 所获得的信息量为 $s^e_{ik}=s^e_{ij}T^e(j,k)$，按上述规则直到网络中所有边都含有信息源的信息量时，该边信息扩散结束，获得信息矩阵 $\boldsymbol{S}^e_{m\times m}$．

步骤 4 信息传递结束后，$s^e_{ii}=1$ 表示边 e_i 的原始信息量，$s^e_{ij}(j=1,2,\cdots,n,j\neq i)$ 表示源边 e_i 传递到其他边的信息量．$T^e_j=\sum_{i=1}^{n}s^e_{ij}$ 表示边 e_j 的信息总量．

在本算法中，用边的总信息替换峰值簇中每个节点的密度，以避免选择截止距离．

此外，需要根据获得信息矩阵 $\boldsymbol{S}^e_{m\times m}$ 确定网络中边的距离矩阵 $\boldsymbol{D}^e_{m\times m}$．

(1) 由于边到自身的距离为 0，故设 $\boldsymbol{D}^e_{m\times m}$ 对角线元素为 0．

(2) 令 $D^e_{ij}=S^e_{ij}(i\neq j)$，获得 $\boldsymbol{D}^{e'}_{m\times m}$．

(3) 根据获得的 $\boldsymbol{D}^{e'}_{m\times m}$ 及该矩阵每行的最大值将其逐行进行归一化，获得距离矩阵 $\boldsymbol{D}^e_{m\times m}$．

矩阵 $\boldsymbol{D}^e_{m\times m}$ 中 d^e_{ij} 表示边 i 到边 j 的距离，参数 δ^e_i 表示边 i 到比它信息量大的边的最短距离，当边 i 的密度最大时，令 $\delta^e_i=\max\{\delta^e_j(j=1,2,\cdots,m,j\neq i)\}$．

至此，网络中的每一条边都可以用该边的总信息量 T^e_i 和距离 δ^e_i 标记，记 $M^e_i=(T^e_i,\delta^e_i)$．

1.2 核心边获取

本文利用改进的 k-means 算法进行核心边选取．通过获取核心边可以获取整个网络社区的核心集及网络所包含的重叠社区的数目．

k-means 算法利用质心(不同聚类的中心)来表示不同类别，具体步骤如下．

步骤 1 随机选取 k 个初始聚类中心．

步骤 2 计算出剩余边到聚类中心的距离，将每个数据点分配到距离其最近的中心所在类别．

步骤 3 重新计算 k 个聚类的中心．

步骤 4 重复步骤 2、步骤 3，直到聚类中心不再改变．

为使 k-means 算法能更好地区分核心边和非核心边，引入混合参数 $\gamma^e_i=T^e_i\delta^e_i,i=1,2,\cdots,m$，$\gamma^e$ 值越大越有可能是核心边．把 k-means 算法应用于一维变量 γ^e 上，进行网络核心边的选取，具体步骤如下．

步骤 1 分别选取 γ^e 的最大值与最小值作为初始聚类中心，分别记 γ^p_{c}、γ^p_{nc} (p 代表第 p 次迭代，c 代表核心边集，nc 代表非核心边集)．

步骤 2 计算剩余边的 γ_i 值到聚类中心的距离：$d^i_{\mathrm{c}}=|\gamma^e_i-\gamma^p_{\mathrm{c}}|$ (γ^e_i 到核心边的距离)，$d^i_{\mathrm{nc}}=|\gamma^e_i-\gamma^p_{\mathrm{nc}}|$ (γ^e_{i} 到非核心边的距离)，若 $d^i_{\mathrm{c}}<d^i_{\mathrm{nc}}$ 则将该边划分至核心边集，否则将该边划分至非核心边集．

步骤 3 根据如下公式重新计算 2 个聚类的中心．

$$\gamma_1^{p+1}=\frac{1}{N_1}\sum_{j\in\mathrm{clu1}}\gamma^e_j,\ \gamma_1^{p+1}=\frac{1}{N_2}\sum_{j\in\mathrm{clu2}}\gamma^e_j$$

步骤 4 重复步骤 2、步骤 3，直到聚类中心不再改变，即 $\gamma_1^p=\gamma_1^{p+1}$，$\gamma_2^p=\gamma_2^{p+1}$．即可得到核心边集和非核心边集，核心边集聚类中心 $\overline{\gamma}^e_{\mathrm{c}}$ 大于非核心边集聚类中心 $\overline{\gamma}^e_{\mathrm{nc}}$，并且核心边集中边的个数即为网络中的社区数．

通过上述方法即可获得核心边集．

1.3 社区划分

通过第 1.2 节获取了核心边集，根据第 1.2 节对核心边的定义以及 k-means 算法的意义，可知核心边集中的每一个核心边是其所在社区的核心．因此可将每一个核心边单独分配一个社区．对于非核心边，由距离矩阵 $\boldsymbol{D}^e_{m\times m}$ 可以确定非核心边到核心边的距离(即矩阵中的元素值)，将其分配到与其距离最近的核心边所在的社区，即找出该行最小值所在列值．遍历

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

所有非核心边，即可将所有非核心边分配到相应核心边所在社区，从而获得边社区划分．根据边的定义可知，每条边都存在其所连接的节点，因此，当得到网络边社区结构后，按如下规则将其转换为重叠社区：网络节点所属社区为与其相连的边所属社区，即如果网络中的边属于多个社区，即不同边集中包含同一边，说明该边是跨社区连接的，则与该边相连的节点均为重叠节点，如边ε 所连接节点分别为 a、b，在进行边集划分时，多个边集合（如 E_1、E_2、E_3）中包含该边ε，则 a、b 两个节点同时属于这 3 个社区，此时 a、b 均为重叠节点．据此，可获得网络中节点的社区划分．MPTD-EC 算法步骤如下．

1）信息扩散

步骤 1 通过信息扩散获取网络边的 $\boldsymbol{S}^e_{m\times m}$ 矩阵．

步骤 2 据信息矩阵得归一化距离矩阵 $\boldsymbol{D}^e$．

$D^e_{ij}=S^e_i/S^e_{ij}(\boldsymbol{S}^e_i=\max\{S^e_{ij},j=1,2,\cdots,m\},i\neq j)$，$D_{ii}=0$．

步骤 3 使用边的总信息量替代边密度 ρ^e ：$\rho^e=T^e_j=\sum_{i=1}^{n}s^e_{ij}$．

步骤 4 使用 $M^e_i=\left(T^e_i,\delta^e_i\right)$ 表征每条边的属性．

2）核心边的划分

步骤1 引入混合参数 $\gamma^e_i=T^e_i\delta^e_i,i=1,2,\cdots,m$，选取 γ^e 的最大值与最小值作为初始聚类中心，记为 γ^p_{c}、γ^p_{nc}．

步骤 2 计算剩余边的 γ^e_i 值到两个聚类中心的距离：$d^i_{\mathrm{c}}=\left|\gamma^e_i-\gamma^p_{\mathrm{c}}\right|$，$d^i_{\mathrm{nc}}=\left|\gamma^e_i-\gamma^p_{\mathrm{nc}}\right|$，将该边划分至距其距离较近的一类．

步骤 3 根据如下公式重新计算两个聚类的中心：$\gamma^{p+1}_1=\frac{1}{N_1}\sum_{j\in\mathrm{clu}1}\gamma^e_j$，$\gamma^{p+1}_1=\frac{1}{N_2}\sum_{j\in\mathrm{clu}2}\gamma^e_j$．

步骤4 重复步骤 2、步骤 3 至聚类中心不再改变．

3）非核心边的划分

据距离矩阵 $\boldsymbol{D}^e_{m\times m}$ 确定非核心边到核心边的距离（即矩阵中的元素值），将其分配到与其距离最近的核心边所在社区．

2　算法的时间复杂度分析

记网络中边的数量为 m，第 1 步，网络中边的信息传输时间复杂度为 $O(m-1)$．此外，由于边都应该作为源将信息进行传播，其时间复杂度为 $O(m^2-m)$．计算距离矩阵的复杂度均为 $O(2m^2)$．第 2 步，核心边选取使用了 k-means 算法，由于只需将边分为两类，所以复杂度为 $O(2mT)$，其中 T 为迭代次数，因为核心边与非核心边区别明显，迭代次数相对于网络边数可忽略，复杂度可简化为 $O(2m)$．第 3 步，只需依次遍历网络边和节点，即可完成重叠社区划分，其时间复杂度为 $O(m+n)$，n 为网络中节点数．故算法总的时间复杂度为 $O(m^2)$．

3　评价指标

为了衡量社区发现算法所获社区结构的优劣，提出了模块度[8]、标准化互信息（NMI）等[9]评价指标．研究发现模块度存在分辨率的限制，即网络中较小规模社区结构的存在会造成模块度值很大，但划分的社区结构并非网络的最佳划分．相比模块度，NMI 更能评价社区划分的性能．同时因为人工产生的数据集网络本身即重叠社区，使用 NMI 进行评价．但对于真实数据集，人为将其划分为非重叠社区，因此进行重叠社区划分时，使用 NMI 并不能衡量算法的性能，因此在真实数据集中，使用改进的模块度进行算法性能的评价．

3.1　标准化互信息

本文通过标准化互信息（NMI）表征本算法在人工网络数据集的性能．NMI 通过信息熵来衡量社区发现算法所划分的社区结构和网络已知的社区结构的差异．其计算公式为

$$\mathrm{NMI}=\frac{-2\sum_{i,j}N_{ij}\ln\left(\frac{N_{ij}m}{N_{i.}N_{.j}}\right)}{\sum_i N_{i.}\ln\left(\frac{N_{i.}}{m}\right)+\sum_j N_{.j}\ln\left(\frac{N_{.j}}{m}\right)}\tag{6}$$

式中：混淆矩阵 $\boldsymbol{N}$ 是由已知的社区结构构成的；m 表示网络中的节点；$\boldsymbol{N}_{i.}$ 表示矩阵的第 i 行中元素的和；$\boldsymbol{N}_{.j}$ 表示矩阵的第 j 列中元素的和．

3.2　基于模块度的改进评价指标

为了测量本算法在实际网络上的性能，实验选择了由 Nicosia 等[10]于 2009 年提出的考虑节点所属的社区数量以及每个社区中的节点度的一个被广泛接受的模块化函数 Q_{ov}，以此衡量算法性能．其定义为

$$F(\alpha_{i,c},\alpha_{j,c})=\frac{1}{(1+e^{f(\alpha_{i,c})})(1+e^{-f(\alpha_{j,c})})}\tag{7}$$

$$\beta_{l(i,j),c}=F(\alpha_{i,c},\alpha_{j,c})\tag{8}$$

$$\beta^{\mathrm{out}}_{l(i,j),c}=\frac{\sum_{j\in V}F(\alpha_{i,c},\alpha_{j,c})}{|V|}\tag{9}$$

$$\beta^{\mathrm{in}}_{l(i,j),c}=\frac{\sum_{i\in V}F(\alpha_{i,c},\alpha_{j,c})}{|V|}\tag{10}$$

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

$$Q_{\mathrm{ov}}=\frac{1}{m}\sum_{c}\sum_{i,j\in V}\left[\beta_{l(i,j),c}A_{ij}-\beta^{\mathrm{out}}_{l(i,j),c}\beta^{\mathrm{in}}_{l(i,j),c}\frac{k_i^{\mathrm{out}}k_j^{\mathrm{in}}}{m}\right] \tag{11}$$

式中：$\alpha_{i,c}$ 表示任意节点 i 属于社区 c 的隶属系数；$\beta_{l(i,j),c}$ 表示节点 i 和节点 j 之间的边缘属于社区 c 的隶属系数；$\beta^{\mathrm{out}}_{l(i,j),c}$ 表示连接到社区 c 中节点 i 的出边缘的隶属系数的平均值；$\beta^{\mathrm{in}}_{l(i,j),c}$ 表示连接到社区 c 中节点 i 的入边缘的隶属系数的平均值. Q_{ov} 越大，其相应的社区结构越接近网络的最佳划分.

4 仿真实验

为了验证本算法的可行性与有效性，分别在合成网络和真实网络上进行了实验，同时将本算法与一些已提出的算法(COPRA[11]和 CMP[12])进行了对比分析.

4.1 人工网络数据集

本文使用 LFR 模型[13]生成网络，在此基础上进行了分析. 由于该模型可通过调整参数控制整个网络和社区属性，如大小、节点度分布、重叠节点数等，合成的网络结构更逼近于真实的社交网络拓扑.

本文使用该模型生成 3 种网络. 网络参数如表 1 所示. 其中，网络中的节点数量为 10 000，社区大小为 20～100 不等. 混合参数 μ 分别设为 0.1、0.2 和 0.3，理论表明该值越接近 1，社区结构越难发现. 参数 O_{m} 为重叠节点所属的最大社区数，O_{n} 为重叠节点数. 研究发现，现实中的大多数情况下，重叠节点属于不超过 6 个社区，故本实验将其值分别设置为{2，3，4，5，6}.

表 1 网络参数

Tab.1 Parameters of network

网络	N	μ	社区规模	O_{n}	O_{m}
LFR1	10 000	0.1	20,100	1 000	2,3,4,5,6
LFR2	10 000	0.2	20,100	1 000	2,3,4,5,6
LFR3	10 000	0.3	20,100	1 000	2,3,4,5,6

4.2 真实网络数据集

为了进一步测试所提算法，同时在几个经典的经常被用作测试网络的真实社交网络(空手道数据集[14]、Dolphins 数据集[15]、Football 数据集[16]、Polbooks 数据集[17]、Email 数据集[18]和 PGP 数据集[19])上进行实验. 这些社交网络的社区结构如表 2 所示. 其中，Email 数据集和 PGP 数据集社区数均为 0.

表 2 6种真实网络的参数

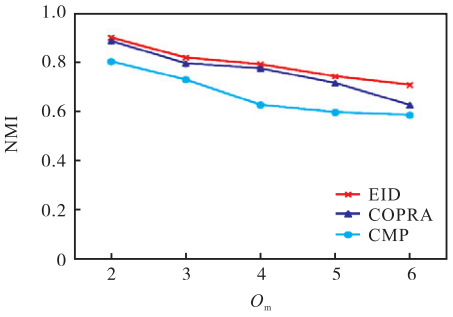
Tab.2 Parameter of six real networks

数据集	节点数	边数	社区数
Karate	34	78	2
Dolphins	62	158	2
Football	115	612	12
Polbooks	105	441	3
Email	1 133	5 254	0
PGP	10 680	24 316	0

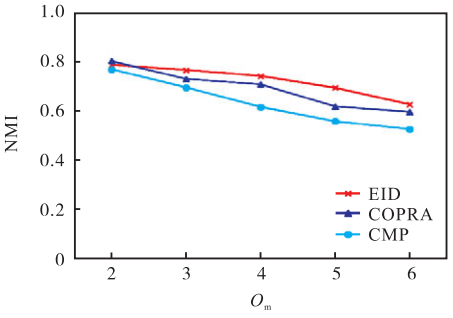
4.3 实验结果及分析

4.3.1 人工网络数据集算法性能

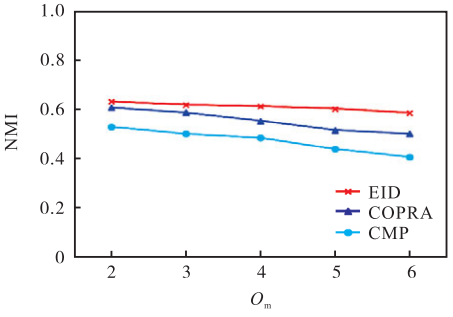
在本节中，分别在 LFR1、LFR2 和 LFR3 上进行了本算法的性能测试. 同时，为了进行比较，在相同的网络上进行 COPRA 和 CMP 的测试. 实验结果如图 1 所示.



(a) LFR1 网络



(b) LFR2 网络



(c) LFR3 网络

图 1 MPTD-EC 算法性能

Fig.1 Performance of the MPTD-EC algorithm

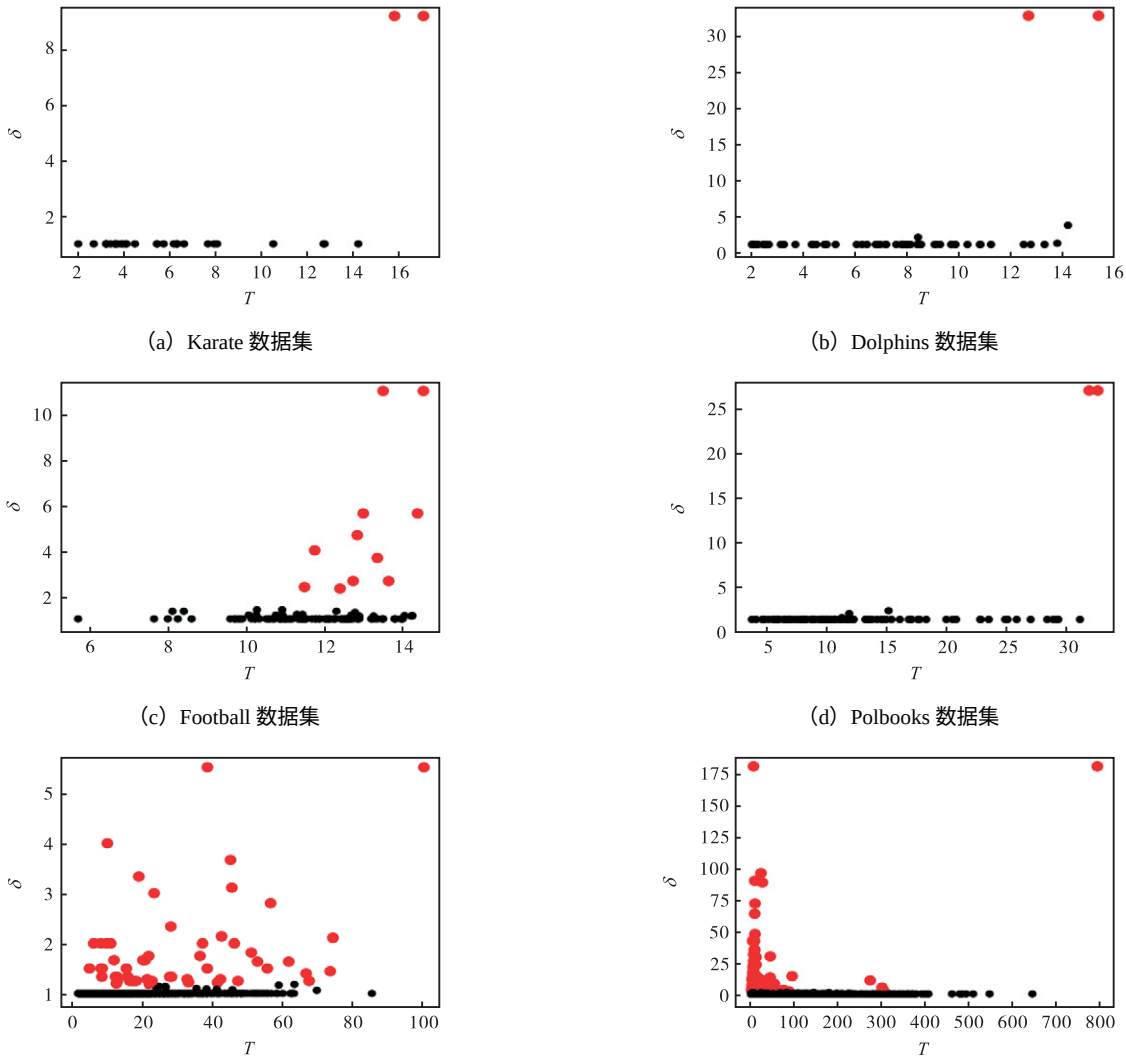
结果表明 3 种算法的性能随着 O_{m} 增长而减

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

少. 即重叠节点属于更多社区和整个网络结构更大时，网络更加复杂，更难分析，这与实际情况是相符合的. 对 3 个结果进行纵向对比发现随着 μ 的增长，3 种算法的性能也会降低. 结果也符合预期，即混合参数越大，网络就越复杂，算法性能有所下降. 结果表明，虽然 MPTD-EC 在某些情况下与 COPRA 及 COPRA 性能接近，但在其他情况下，MPTD-EC 优于其他两种方法. 并且与已提出的算法相比，本算法更适用于重叠社区较多的复杂网络.

4.3.2　真实网络数据集算法性能

为了评价本算法性能，实验采用基于模块度的改进评价指标，分别将 5 种算法应用在 6 种真实数据集上，数据集相关参数参见表 3，实验结果获得 6 种网络决策图(根据边的总信息量 T 及距离 δ 绘制)，如图 2 所示.



(a) Karate 数据集　(b) Dolphins 数据集

(c) Football 数据集　(d) Polbooks 数据集

(e) Email 数据集　(f) PGP 数据集

图 2　MPTD-EC 算法在 6 种真实数据集上获得的决策图

Fig.2　Decision graph obtained by MPTD-EC algorithm on six real data sets

5 种算法分别在 6 个真实网络中获得的 Q_{ov} 如表 3 所示.

COPRA 是随机算法，图 3 中决策图红点个数表示核心边的个数，即表征了网络中社区的个数. 表 3 中相应的实验结果是 50 个独立结果的平均值. 表 3 表明本算法在这些社交网络中具有最佳效果，说明本算法的有效性.

表 3　5 种算法所得的 $\boldsymbol{Q}_{ov}$ 值

Tab.3　Values of $\boldsymbol{Q}_{ov}$ obtained by five algorithms

网络	MPTD-EC	COPRA	CPM	LFM	CFinder
Karate	0.68	0.52	0.52	0.42	0.52
Dolphins	0.74	0.69	0.66	0.28	0.66
Football	0.69	0.68	0.64	0.45	0.64
Polbooks	0.84	0.82	0.79	0.74	0.79
Email	0.56	0.51	0.46	0.25	0.46
PGP	0.81	0.78	0.57	0.44	0.57

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

5 结 语

本文在密度峰值聚类的基础上，提出了一种基于边信任度的混合参数自适应重叠社区发现算法．在本算法中，峰值簇中的密度和距离由网络边与边之间的信息传递决定，避免截断距离选取，并且引入混合参数，使用 k-means 算法进行核心边的选取，从而确定网络社区个数，并且通过边到点的转换完成重叠节点和重叠社区的发现，不需要人为设定相关参数，使算法达到自适应．综合实验结果，可知本算法在人工数据集上较已提出的基于边聚类的社区发现算法更适用于重叠社区数较多的社区发现问题，并且在真实数据集上性能也有较大提升，验证了该算法的可行性和有效性．

参考文献：

[1] Chen Y，Zhao P，Li P，et al. Finding communities by their centers[J]. Scientific Reports，2016，6：24017-1-8.

[2] Huang L，Wang G，Wang Y，et al. A link density cluster ing algorithm based on automatically selecting density peaks for overlapping community detection[J]. International Journal of Modern Physics B，2016，30(24)：165-167.

[3] Zhou X，Liu Y，Zhang J，et al. An ant colony based algorithm for overlapping community detection in complex networks[J]. Physica A Statistical Mechanics & Its Applications，2015，427：289-301.

[4] Ahn Y Y，Bagrow J P，Lehmann S. Link communities reveal multiscale complexity in networks[J]. Nature，2010，466(7307)：761-764.

[5] Shi C，Cai Y，Fu D，et al. A link clustering based over lapping community detection algorithm[J]. Data & Knowledge Engineering，2013，87(9)：394-404.

[6] Zhang X K，Tian X，Li Y N，et al. Label propagation algorithm based on edge clustering coefficient for community detection in complex networks[J]. International Journal of Modern Physics B，2014，28(30)：1450216-1-15.

[7] Hu Y，Li M，Zhang P，et al. Community detection by signaling on complex networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2008，78(2)：016115.

[8] Clauset A，Newman M E，Moore C. Finding community structure in very large networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2004，70：066111-1-6.

[9] Danon L，Díazguilera A，Duch J，et al. Comparing community structure identification[J]. Journal of Statistical Mechanics Theory & Experiment，2005，2005(9)：09008-1-10.

[10] Nicosia V，Mangioni G，Carchiolo V，et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of Statistical Mechanics：Theory & Experiment，2009(3)：3166-3168.

[11] Gregory S. Finding overlapping communities in networks by label propagation[J]. New Journal of Physics，2010，12(10)：2011-2024.

[12] 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.

[13] Lancichinetti A，Fortunato S. Benchmarks for testing com munity detection algorithms on directed and weighted graphs with overlapping communities[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2009，80：016118-1-8.

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

[15] Lusseau D. The emergent properties of a dolphin social network[J]. Proceedings Biological Sciences，2003，270(Suppl 2)：186-188.

[16] Girvan M，Newman M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences，2002，99(12)：7821-7826.

[17] Newman M E J，Girvan M. Finding and evaluating community structure in networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2004，69(2)：026113-1-15.

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

[19] Boguñá M，Pastor-Satorras R，Díaz-Guilera A，et al. Models of social networks based on social distance attacment[J]. Physical Review E，2004，70(5)：056122-1-8.

(责任编辑：王晓燕)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net