doi: 10.7690/bgzdh.2018.09.006

# 一种检测重叠和非重叠社区的方法

王一萍
（齐齐哈尔大学计算机与控制工程学院，黑龙江 齐齐哈尔 161006）

**摘要：**为发现有重叠和非重叠连接结构的社区，针对现有方法的不足提出一种新型社区检测方法。基于主节点准则，为每个节点计算内部和外部关联度检测社区，设计社区检测综合算法，在人工网络上进行仿真并与已知算法对比。结果表明，该方法具有较为优越的性能和可行性。

**关键词：**社区检测；社会网络；重叠连接结构；规范互信息
**中图分类号：**TP393 **文献标志码：**A

# A Method for Overlapping and Non-overlapping Community Detection

Wang Yiping
(*College of Computer & Control Engineering*, *Qiqihar University*, *Qiqihar* 161006, *China*)

**Abstract:** In order to find the community with overlapping and non-overlapping connection structures, a new community detection method is proposed for the deficiency of existing methods. Based on the principle of master node, calculate the internal and external correlation detection communities for each node, design the IEDC algorithm, simulate on the artificial network and compare with the known algorithm. The results show that the method has superior performance and feasibility.

**Keywords:** community detection; social networks; overlapping connectivity structures; normalized mutual information

## 0 引言

检测社区结构能揭示网络实体关系背后潜在机制。由于社区的重要性，社区检测被广泛应用，如场景检测[1]、流行传播模型[2]和基因癌症驱动因素检测[3]等。社区检测与聚类相似，但数据网络表示导致典型聚类技术社区检测结果较差。一方面用网络科学增强检测算法准确性；另一方面社区发现复杂性引起很多问题。检测算法有重叠和非重叠结构[4-5]。

一般说，网络中个体趋于属多个社区，如社会网络中一个人与同学、同事联系，研究人员与不同领域合作者关系。基于重叠结构算法如寻找相邻派系、边划分和标签传播算法[6]。早期社区检测从初始化敏感性进行，精度低且有单独个体社区未被发现。多数方法试图根据社区结构重叠或不重叠假设发现社区。尽管在预先确定重叠假设数据集上取得良好效果，但未设计出一个通用框架。事实上社区可由重叠和不重叠元素形成。笔者基于内部和外部关联度组合，提出一种基于主关联节点准则社区检测新方法。该方法通过外部关联度考虑网络社区重叠结构，且通过计算每个网络成员内部关联度应用社区结构非重叠模式。

## 1 相关工作和基本想法概述

### 1.1 相关工作

文中算法是基于网络连通性结构，含重叠和不重叠[7]问题。一些重叠社区检测将网络分有外部连接密集和内部连接弱子网，非重叠社区检测将网络分内部密集和外部连接弱子网。多数算法重叠结构中检测非重叠难以实现。

### 1.2 基本思想

设计通用框架不仅考虑重叠连通性，且需考虑不重叠连通性。提出“将每个社区划分为非重叠部分和重叠部分”，基于网络社区形成通过考虑重叠和不重叠连通性来组成。所提方法考虑了提取社区重叠生成随机建模和提取内部社区结构特征方法，基于概率模型架建通用社区检测方法，以检测重叠和非重叠连接结构社区，比类似重叠或非重叠过程具有更高准确性和可接受复杂度。文中算法考虑每个节点的社区成员的非重叠和重叠部分 2 个参数。第 1 个参数“IA”用给定节点局部邻接连接测量其对各社区内部关联度。第 2 个参数通过生成随机块

**收稿日期：**2018-05-24；**修回日期：**2018-07-19
**作者简介：**王一萍(1971—)，女，黑龙江人，硕士，副教授，从事复杂网络与群智能研究。

模型“EA”来衡量网络中每个节点社区间的交互量。因此一个节点属于特定社区概率取决于它的邻接及生成随机块模型。节点属于基于以下关联度的社区： 1) 捕捉社区中非重叠内部联系；2) 揭示社区重叠部分外部联系。

## 2 所提方法的要素

用 $G(V,E)$ 表示网络，$V$ 为顶集，$E$ 为边集。社区检测为发现满足条件：$C=\{c_i \mid i=1,2,\cdots,k \wedge \cup_{i=1}^{k} c_i = V\}$ 的 $C$ 集合。

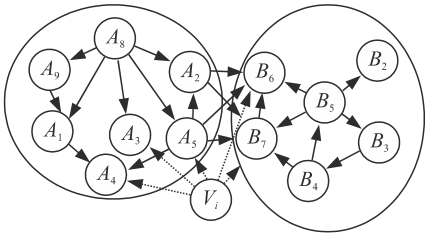
1) 非重叠区域的内部关联。

内部关联度表示每个节点与特定社区的亲和度，被定义为式(1)，其中 $N(v_i)$ 是节点 $v_i$ 的邻接点集，$p(v_j \mid c_i)$ 表示节点 $v_j$ 属于社区 $c_i$ 的社区传播概率。直观上，式(1)指出一个节点与通过给定社区的亲和度取决于其邻接点的亲和度。

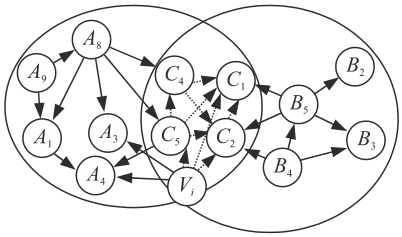
$$IAv_i(c_i)=\frac{\sum_{v_j \in N(v_i)} p(v_j \mid c_i)}{|N(v_i)|}。\tag{1}$$

2) 重叠区域的外部关联。

块模型是统计理论中社会网络中查找有相同属性节点组的工具，目标是找到每个节点社区标签和交互矩阵 $\beta_{k\times k}$ 计算社区交互级别。参数 $\beta_{c_1,c_2}$ 是社区 $c_1$ 和 $c_2$ 间链接数。根据块模型，2 个社区重叠度取决于交互边数。图 1 为社区间交互作用导致更密集重叠区域计算各节点外部关联度：1) 估计任何社区对间交互矩阵；2) 根据社区传播概率和交互矩阵算节点外部关联度。用提出方法检测社区，将节点内部与外部关联度对比。得到基于“内部和外部关联发现社区”(IEDC)整合算法。




(a) 2 个互动率高的非重叠社区




(b) 2 个重叠社区

**图 1　社区的外部关联程度**

## 3 IEDC 算法

网络由一组交互社区组成，通过对社区结构进行总结后的假设。网络中各节点都与社区有双重关联：一是通过一个社区与其他节点内部相关，另一个与其他社区节点外部关系相关。大多数算法已应用节点内部或外部关联发现网络中隐藏社区，文中称为非重叠或重叠技术。笔者提出一种综合概率方法，通过利用网络中每个节点的内部和外部关联度来检测社区结构。

### 3.1 综合概率方法

针对每个节点 $v_i$ 提出属于社区 $c_i$ 隶属度准则：

$$\begin{aligned} p(v_i \in c_i) &= p(v_i \in c_i \mid N(v_i) \in c_i) + \\ &\quad p(v_i \in c_i \mid N(v_i) \in c_{-i}) = \\ &\quad p_1(c_i)\bullet IA_{c_i} + p_2(c_i)\bullet EA_{c_i}。 \end{aligned}\tag{2}$$

$v_i$ 邻接点属于 $c_{-i}=\cup_{j=1}^{k} c_j \setminus c_i$，而 $p_1(c_i)$、$p_2(c_i)$ 是关联部分重要参数。式(2)中第一项计算 $v_i$ 在社区 $c_i$ 内部关联度。第二项揭示 $N(v_i)$ 社外部关系对当前社区影响。交互度越高产生重叠越多。用生成概率法得每个节点外部关联为式(2)第三项。$v_i$ 的外部关联 $EA_{c_i}(v_i)$ 定义为：

$$EA_{c_i}(v_i)=\frac{\sum_{v_j \in N(v_i)} \sum_{cj \in C} P(v_j \mid c_j)\bullet \beta(c_i,c_j)}{|N(v_i)|}。\tag{3}$$

其中 $\beta(c_i,c_j)$ 是社区之间的交互矩阵。任何一对社区之间的交互矩阵可以通过最大可能性方法统计来近似地被定义，$\rho$ 被称为稀疏正则化因子，计算每个节点的内部关联度和外部关联度后，每个节点的最终传播概率取决于重要程度 $p_1$ 和基于以下等式获得的 $p_2$。

$$\left.\begin{aligned} &\frac{p_1(c_i)}{p_2(c_i)}=\frac{\sum_{(v_i,v_j),v_i \in c_i,v_j \in c_i}(v_i,v_j)}{\sum_{(v_i,v_j),v_i \in c_i,v_j \notin c_i}(v_i,v_j)} \\ &p_1(c_i)+p_2(c_i)=1 \end{aligned}\right\}。\tag{4}$$

### 3.2 IEDC 算法的描述

算法初始化取决于网络结构，更新传播概率迭代次数为 Maxiyrt。先为每个节点初始化各社区传播概率，用边聚类法获取网络初始社区结构。InterMax 和 fFactor 是文献[7]中交互社区矩阵、内部和外部关联度函数。findIA 据式(1)计算节点与社区内部关联度。findEA 据式(3)估计节点外部关联度，Udpropagae

据式(2)更新传播概率。如 $v_i$ 对 $c_i$ 传播概率大于阈值，则 $v_i$ 将分配给 $c_i$。IEDC 算法描述为：

输入：$G(V,E)$,Maxiter；输出：$C=\{c_i|i=1,\cdots,k\}$。

1) 初始化：寻找初始社区传播概率 $p(v_i\,|\,c_i)$

2) 计算：$\beta(c_i,c_j)=\text{InterMax}\,(P,G)$

3) 计算：$p_1(c_i),p_2(c_i)=\text{fFactor}(P,G)$

4) $\text{iter}=0$

5) while (iter＜Maxiter)

6) $\text{iter}=\text{iter}+1$ for $i=1$ to $N$ do

7) for j = 1 to k do

8) $IA_{v_i}(c_j)=\text{findIA}(P,G)$，$EA_{v_i}(c_i)=\text{findEA}(P,G,\beta)$

9) $p^{\text{update}}(v_i|c_i)=\text{Udpropagate}\,(P,IA,EA,p_1,p_2)$

10) for $i=1$ to $N$ do

11) for $j=1$ to $k$ do

12) if $p(v_i|c_j)\geqslant$ threshold then Add: $F(c_j)\leftarrow vi$

### 3.3 所提出方法的时间复杂度分析

IEDC 法分 3 阶段：根据边聚类法找社区种子集。因大多数网络稀疏性，复杂度为 $O(m)$；计算各节点内部关联度，与边和社区数有关。完全连通网络最多 $O(n\times m\times k)$，（$n$ 节点数，$k$ 社区数），可减少到 $O(m)$。第三阶段，根据式(3)计算各社区节点外部关联度，最复杂情况下更新 $\beta$ 参数，复杂度是 $O(n^2)$。

## 4 实验仿真与分析

用人工生成网络和真实数据集分析 IEDC 算法。首先在人工网络上仿真；然后将 IEDC 在真实网络数据集上实验，并与其他算法进行比较。

1) 评价标准。

应用归一化互信息 NMI 和 F1 分数评估算法性能。NMI 是度量被检测社区与真实社区的相似性，而 F1 分数是基于真实信息对每个社区中正确分类的成员进行测量。

2) 人工网络生成。

为生成满足各种情况人工网络，采用 LFR 法。根据实验研究，模块度值大于 0.5 导致稀疏重叠社区形成，小于 0.4 导致密集重叠社区结构。用多种最先进算法检查 IEDC 方法，得 NMI 和 F1 分数。结果表明除有 500 个节点稀疏网络之外仿真场景，IEDC 优于最先进算法。

3) 真实网络数据集的实验。

笔者所提出方法通过各种基准真实网络与最先进的社区检测算法进行比较。使用 2 个评估标准 NMI 和 F1 分数来检查在真实网络上提出的方法。

在非重叠社区结构 Football, Polbooks, Polblogs, RiceCalTech 真实网络数据集上评估 IEDC，节点数和边数分别为(115, 616)，(102, 441)，(1 490, 167 236)，(769, 16 656)，(4 087, 184 828)。实验结果如图 2，可见基于 2 个评估标准，IEDC 在 Polbooks、CalTech 和 Football 上 F1 分数和 NMI 值比其他算法表现稍好或相同。实验结果表明：5 个不同数据集与其他算法相比，IEDC 方法能以适当准确度检测出非重叠结构。

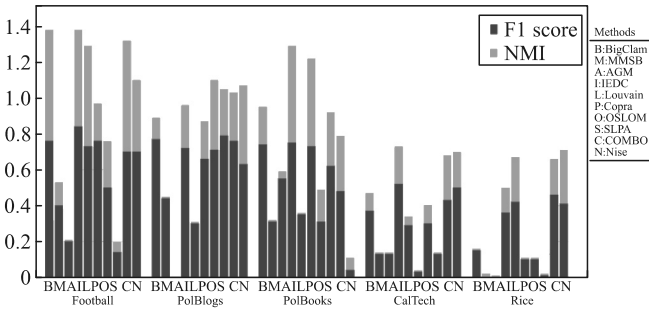



图 2 非重叠社区结构网络中 F1 分数和 NMI 标准的结果

在有重叠社区结构 6 个真实数据集上评估 IEDC，网络名及对比算法见图 3。该图为基于 NMI 和 F1 分数评估标准重叠社区划分实验结果。由于基于大型网络基准算法在社区结构重叠大网络，用基于抽样策略在每个网络中获得重叠子图社区。

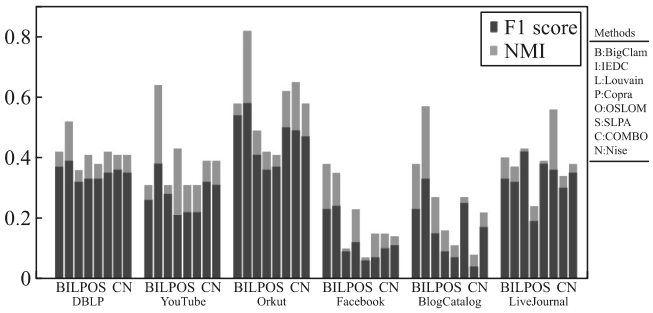



图 3 重叠社区结构网络中 F1 分数和 NMI 标准的结果

## 5 结论

笔者基于主节点准则，通过为每个节点计算内部和外部关联度检测社区，设计 IEDC，在人工网络上仿真并与已知算法对比，证明 IEDC 具有优越的性能和可行性。用具有不同重叠社区小规模真实数据集到基准网络。实验验证了 IEDC 方法的优点。

## 参考文献：

[1] HAMDAQA M, TAHVILDARI L, LACHAPELLE N, et

al. Cultural scene detection using reverse Louvain optimization[Z]. Science of Computer Programming 95，2014: 44–72.

[2] REN G, WANG X. Epidemic spreading in time-varying community net- works[J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2014, 24(2): 068701–4251.

[3] CANTINI L, MEDICO E, FORTUNATO S, et al. Detection of gene com- munities in multi-networks reveals cancer drivers[J]. Scientific Reports, 2015, 5: 17386.

[4] YANG J, LESKOVEC J. Community-affiliation graph model for overlapping network community detection[J]. Proc. IEEE Int. Conf. Data Min., 2013, 5(1): 1170–1175.

[5] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New J. Phys., 2008, 11(3): 19–44.

[6] 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.

[7] GIRVAN M, NEWMAN M E. Community structure in social and biological networks[J]. Proceedings of the national academy of sciences, 2002, 99(12): 7821–7826.

**********

（上接第 19 页）

## 参考文献：

[1] HE B L, MAO Z. A Track-to-track Association Algorithm with Chaotic Neural Network[C]. Asian-Pacific Conference on Synthetic Aperture Radar Proceedings. IEEE PRESS, 2009: 788–791.

[2] 金宏斌, 徐毓. 基于状态方程的雷达目标航迹模拟方法[J]. 航空计算技术, 2003, 33(2): 34–36.

[3] 郭锐, 崔爱国, 杨大志. 一种航迹仿真方法及其在雷达组网系统中的应用[J]. 雷达与对抗, 2006(3): 50–53.

[4] 李欣, 彭世蕤. 一种空间三维航迹建模新方法[J]. 雷达科学与技术, 2007, 5(5): 365–370.

[5] 陈敏, 王晓亮, 汪万维, 等. 不等间隔时间采样高精度雷达数据仿真方法[J]. 计算机仿真, 2015, 32(2): 111–114, 118.

[6] 杨俊强, 毛征, 张志, 等. 雷达航迹融合算法验证系统设计[J]. 国外电子测量技术, 2009, 28(11): 63–66.

[7] 郭吉成. 高炮火控系统射击问题算法研究[D]. 大连: 大连理工大学, 2005: 20–24.

[8] 李文才, 张翼. 近似计算求解高炮相遇问题[J]. 兵工自动化, 2011, 30(4): 8–9, 11.

[9] 高萌. 雷达航迹处理算法及仿真平台设计与实现[D]. 西安: 西安电子科技大学, 2015: 29–30.