基于网络加权机制的动态迭代聚类算法

汪自洁[1] 周雅静[2] 李慧嘉[2]
(中国政法大学司法文明协同创新中心 北京 100080)[1]
(中央财经大学管理科学与工程学院 北京 100081)[2]

摘 要 动态网络在分析功能属性与拓扑结构的相关性方面具有重要作用。文中提出了一个新的动态迭代聚类算法,通过引入包含拓扑信息的权重 W 和紧密度 T 来调整边权和节点紧密度,以提高网络聚类结构检测的速度与准确度。值得一提的是,为了估计最优的迭代停止时间,文中利用以时间 t 为分辨率参数的稳定性指标(stability)作为测度指标,可以自然地找到使聚类划分达到最优的时刻 t。该算法非常高效,而且不需要预先指定聚类的数目,因此可以方便地应用于各种模糊网络。最后在包括法律案例关联网络等数据上的实验结果表明,该算法能快速而准确地探测各种人工和现实网络的聚类结构。

关键词 动态循环算法,网络聚类检测,加权机制,紧密度,法律案例关联网络
中图法分类号 TP393 **文献标识码** A

Dynamical Network Clustering Algorithm Based on Weighting Strategy

WANG Zi-jie[1] ZHOU Ya-jing[2] LI Hui-jia[2]
(Collaborative Innovation Center of Judicial Civilization, China University of Political Science and Law, Beijing 100080, China)[1]
(Central University of Finance and Economics, School of Management Science and Engineering, Beijing 100081, China)[2]

Abstract Network dynamic plays an important role in analyzing the correlation between the function properties and the topological structure. This paper proposed a novel dynamical iteration algorithm incorporating the iterative process of membership vector with weighting scheme, i. e. weighting W and tightness T. These new elements can be used to adjust the link strength and the node compactness for improving the speed and accuracy of community structure detection. To estimate the optimal stop time of iteration, this paper utilized stability function defined as the Markov random walk auto-covariance. The algorithm is very efficient, and doesn't need to specify the number of communities in advance, so it naturally supports overlapping communities by associating each node with a membership vector describing node's involvement in each community. Theoretical analysis and experiments show that the algorithm can uncover communities effectively and efficiently.

Keywords Dynamical iteration algorithm, Network clustering, Weighting strategy, Tightness, Judicial case network

1 引言

从 Barabasi 和 Albert 的开创性研究[1]之后,人们从复杂网络的角度对众多现实世界复杂系统进行了深入的研究和探索。基于海量的现实世界数据,我们可以用节点和边分别描述复杂系统的成分单元和它们之间的相互作用[2-3]。在复杂网络的研究中,聚类结构的探测和分析已经成为一个非常重要的课题。一般来说,网络聚类是指在网络中的一组节点,相比于网络其它部分,其内部的相互关联更加紧密[4-6]。如何从大规模的网络数据中发现最优的聚类结构是一个开放的热点问题,一般可以通过优化特定的指标函数来实现,其中 Newman 等提出的模块度 Q 最为流行[6]。其函数形式为:

$$Q=\sum_{i=1}^{K}\left[\frac{l_i^{\text{in}}}{L}-\left(\frac{d_i}{2L}\right)^2\right]=1-\frac{L_{\text{inter}}}{L}-\frac{1}{K}-\frac{1}{K}\sum_{j=2}^{K}\sum_{k=1}^{j-1}\left(\frac{d_j-d_k}{2L}\right) \tag{1}$$

其中,K 表示聚类的数目,L 表示网络中边的总数,l_i^{in} 和 $d_i=l_i^{\text{in}}+l_i^{\text{inter}}$ 表示聚类 i 的团内边数和总的边数,L_{inter} 代表总的团间边数。模块度 Q 定义为:位于聚类内部的边数,减去不考虑聚类结构时落在随机网络中的相同数量边的期望值。Q 的数值可以指示聚类结构的优劣,Q 值越大表示网络中的聚类结构越明显。

在优化模块度 Q 的算法中,Girvan-Newman(GN)算法是应用最广泛的[6]。GN 算法是一种启发式算法,它通过反复识别并删除聚类之间的边来实现网络的划分,但是这种算法的时间复杂度为 $O(m^2n)$,其中 m 为边的个数,n 为点的个数,当网路规模较大时计算速度会受到很大限制。另外研究表明,优化模块度 Q 得到的聚类结构还会出现如分辨率限制(Resolution limit)和极端退化问题(extremly degeneration)[7]等缺陷。为了缓解优化模块度 Q 的限制,Alaireza 等[8]提出了一种加权方案,以边权的形式在网络中加入某些重要的拓扑信息。实验证明,加入权重 W 可以很大程度上缓解模块度

本文受国家自然科学基金项目(71871233,71401194),北京市自然科学基金(9182015)资助。
汪自洁(1987—),女,博士生,主要研究方向为司法文明、复杂网络和大数据分析等;**李慧嘉**(1985—),男,博士,副教授,CCF 会员,主要研究方向为数据挖掘、商务智能、人工智能等,E-mail:Hjli@amss.ac.cn。

Q 的这些限制。然而，大多数算法需要网络全局信息，这对于一些大且混杂的网络来说是极具难度的。

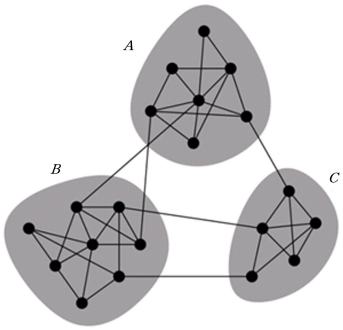


图 1 一个包含 3 个聚类的网络(不同聚类用灰色表示)

我们注意到利用动态过程来分析网络特性已经成为一个新的研究方向，但是现有的聚类算法大多只依赖于拓扑网络而忽略了内在的动态特性。相较于静态算法，从动态的角度能更加灵活而有效地进行聚类探测。本文提出一个新的动态迭代聚类算法，通过引入包含拓扑信息的权重 W 和紧密度 T 来调整边权和节点紧密度，以提高网络聚类的速度与准确度。值得一提的是，为了估计最优的迭代停止时间，我们利用以时间 t 为分辨率参数的稳定性指标(stability)作为测度指标，可以自然地寻找使划分达到最优的时刻 t。本文算法非常高效，而且不需要预先指定聚类的数目，因此可以方便地应用于各种模糊网络中。最后实验表明，本文算法能快速而准确地探测各种人工和现实网络中的聚类结构。

2 理论框架

2.1 基于加权机制的目标函数

给定网络 $G(V,E)$，V 为节点集合，E 为边的集合，$N=|V|$ 和 $M=|E|$ 分别表示节点和边的数目。假设 $G(V,E)$ 是没有自循环的无向网络，即 $A(i,j)=A(j,i)$ 且 $A(i,i)=0$。如果 $g(v,e)$ 是 $G(V,E)$ 的一个聚类，且 $n=|v|$，$m=|e|$，则团内边(即两端都位于聚类 $g(v,e)$ 内部的边)的比率可以写为：

$$\zeta_{\text{inside}}=\frac{m}{n(n-1)/2} \tag{2}$$

同理可得团间边(即连接聚类 $g(v,e)$ 与网络的其余部分的边)的比率为：

$$\zeta_{\text{outside}}=\frac{m_{ext}}{n(N-n)} \tag{3}$$

其中，m_{ext} 表示团内边数量。

根据聚类结构的定义，要求团内边的密度大于团间边的密度，即 $\zeta_{\text{inside}}>\zeta_{\text{outside}}$。另外网络 $G(V,E)$ 的平均边比率定义为 $\zeta_{\text{rand}}=\frac{M}{N(N-1)/2}$，那么团内边比率会大于平均边和团间边。这个关系可以用下式表示：

$$\zeta_{\text{inside}}>\zeta_{\text{rand}}>\zeta_{\text{outside}} \tag{4}$$

接下来，假设网络 $G(V,E)$ 可以划分为 c 个网络聚类，即 $g_1,\cdots,g_c$，其点和边的个数分别为 $n_1,\cdots,n_c$ 和 $m_1,\cdots,m_c$。设 $P=\zeta_{\text{rand}}$ 为空模型(null model)中团内部的比例，则式(4)可以整理为：

$$\begin{aligned}&\zeta_{\text{inside}}(q)=\frac{m_q}{n_q(n_q-1)/2}>P\\&\Rightarrow 2m_q>P_{n_q}(n_q-1)\end{aligned} \tag{5}$$

对于网络 $G(V,E)$ 来说，设计高效的目标函数对于衡量聚类结构至关重要。基于式(5)中聚类连接的非均匀性，理想的划分可以通过最大化式(6)得到：

$$F=\sum_{q=1}^{c}[2m_q-P_{n_q}(n_q-1)] \tag{6}$$

对于网络中的 c 个聚类 $g_1,\cdots,g_c$，每一个节点 i 都有一个归属程度 x_i^q，表示节点 i 属于聚类 q 的概率，其向量形式可以表示为 $\boldsymbol{X}_i=[x_i^1,\cdots,x_i^c]$。如果网络为硬划分，即节点只属于唯一的一个聚类，那么归属向量为 $\boldsymbol{X}_i=[0,\cdots,0,1,0,\cdots,0]$。将目标函数式(6)重写为基于归属向量 $\boldsymbol{x}_i$ 的加权形式：

$$F=\sum_{i\neq j}^{N}\sum_{q=1}^{c}(W_{ij}A_{ij}-\gamma P_{ij}x_i^q x_j^q) \tag{7}$$

其中，W_{ij} 表示边的权重。式(6)中加入了权重这一系数，其实际上是对模块度 Q 的改进。

2.2 加权

现有研究表明，模块度 Q 具有分辨率限制(Resolution limit)和极端退化问题(extremly degeneration)[7]这两个缺陷。但 Khadivi 等[8]研究发现，通过加入权重 W 可以很大程度缓解模块度 Q 的这些限制。受到文献[20]的启发，我们可以利用加权机制来改进网络聚类的效率。具体来说，由于网络拓扑结构对聚类有很大的影响，本文将一些重要拓扑性质整合进来，提出了一种新的权重机制 W 以利于聚类的探测。首先，由于聚类内部的节点拥有共同邻居节点的机率越大，它们的关联越紧密，因此假设权重 W 与共同邻居节点比率(用 R_{ij} 表示)相关：

$$R_{ij}=\frac{2\sum_k A_{ik}A_{jk}}{\sum_k A_{ik}+\sum_k A_{jk}} \tag{8}$$

其中，A_{ij} 为邻接矩阵 $\boldsymbol{A}$ 中的元素。

另一个重要拓扑指标是边介数 $\varphi(e_{ij})$，它代表网络中通过边 e_{ij} 的最短路径的数目：

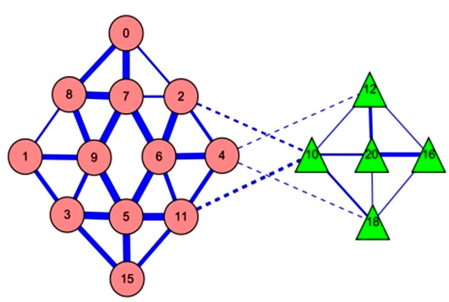
$$B_{ij}=\sum_{u\neq v\in V}\frac{\varphi_{uv}(e_{ij})}{\varphi_{uv}} \tag{9}$$

虽然不同聚类之间的连边相当少，但它们起到了保持网络连通和信息流通的作用。也就是说，团间边的边介数通常会大于团内边。因此，式(9)可以用来区分团间边与团内边。

利用式(8)和式(9)，我们提出以下权重方案 W_{ij}：

$$W_{ij}=\eta\frac{\left(\frac{B_{ij}}{B_{\max}}\right)^{-\alpha}\left(\frac{R_{ij}}{R_{\max}}\right)^{\beta}}{\sum_{m\neq n}\left(\frac{B_{mn}}{B_{\max}}\right)^{-\alpha}\left(\frac{R_{mn}}{R_{\max}}\right)^{\beta}} \tag{10}$$

其中，η 是用于平衡权重的系数；$B_{ij}/B_{\max}$ 和 $R_{ij}/R_{\max}$ 分别为 B_{ij} 和 R_{ij} 的标准化形式；指数参数 α 和 β 用来调整边权中 B_{ij} 和 R_{ij} 在 W 中的比重。图 2 展示了 W_{ij} 对特定网络的影响，其中网络边权用边的粗细来表示，可以明显看出团内边比团间边的边权比重大。



注：边越粗表示权重 W 越大

图 2 一个拥有两个聚类的网络

2.3 紧密度

在文献[9-10]中，节点之间的最短距离被用作网络聚类

划分的标准：当节点属于同一聚类时，节点之间的最短距离一般较小，反之则较大。然而，这些算法仅考虑了整个网络的全局最短距离，为了使探测更加准确，我们基于最短距离提出一种新的量度——紧密度。首先，定义节点 i 和 j 的结构相似度 Sim_{ij} 为：

$$Sim_{ij}=\frac{\sum\limits_{X\in\Gamma(i)\cap\Gamma(j)}W_{ix}W_{jx}}{\sqrt{\sum\limits_{X\in\Gamma(i)}W_{ix}^2}\sqrt{\sum\limits_{X\in\Gamma(j)}W_{jx}^2}} \tag{11}$$

其中，$\Gamma(i)$ 表示节点 i 和它邻居节点的并集。然后，定义紧密度 T 为：

$$T_{ij}=\frac{Sim_{ij}}{\sum\limits_{m,n\in V}Sim_{mn}}[\log(\langle Sim_{ij}\rangle)]^{\theta} \tag{12}$$

其中，$\langle Sim_{ij}\rangle$ 是 Sim_{ij} 的平均值，θ 是调整紧密度的指数。

2.4　动态迭代

为了实现聚类划分，我们设计了一种基于权重 W 和紧密度 T 的动态系统，用来有效地计算每个节点的归属向量。用 $\boldsymbol{X}_i(t)=[x_i^1(t),\cdots,x_i^c(t)]$ 表示节点 i 的归属向量，其元素 $x_i^q(t)$ 为 t 时刻节点 i 属于聚类 $q(q=1,2,\cdots,c)$ 的概率。对于每一个节点 i，x_i^q 随机分布在区间 $[0,1]$ 中。我们将 x_i 标准化，使得对于每个节点 i，x_i^q 之和总为 1。在每个时刻 t，我们将节点 i 的归属向量进行下列循环：

$$x_i^q(t+1)=\frac{\sum\limits_{j=1,i\neq j}^{N}A_{ij}(W_{ij}+1)(W_{ij}+1)[x_i^q(t)]^{\lambda}}{\sum\limits_{q=1}^{c}\sum\limits_{n=1,m\neq n}^{N}A_{mn}W_{mn}T_{mn}[x_m^q(t)]^{\lambda}} \tag{13}$$

其中，A_{ij} 是邻接矩阵 $\boldsymbol{A}$ 中的元素，W_{ij} 表示权重，T_{ij} 表示边的紧密度，$\lambda\in[1,\infty)$。

式(13)为典型的随机游走(Random Walk)动态系统[11-13]，我们注意到，在所有的时间 t 有 $\sum\limits_{q=1}^{c}x_i^q(t)=1$ 成立，确保 $x_i^q(t)$ 是概率的标准化形式。当 $t\to\infty$ 时，$x_i^q(t)$ 逐渐收敛到一个特定值(不一定为最优)，表示节点属于特定聚类的概率。从式(13)可以发现，每个节点的归属向量取决于边权 W，W 可加速 $x_i^q(t)$ 收敛的速度。此外，紧密度 T 可以很大程度上提高动态过程的运行效率；因子 λ 还可以控制归属向量动态循环过程的偏离度。

然而，最优聚类划分的停止时刻是很难估计的，因此需要考虑用有效的目标函数控制迭代时间。

3　动态迭代聚类算法

为了探测网络聚类结构，本文提出了一种基于归属向量 x_i^q 的动态迭代算法，并利用稳定性指标来估计迭代到最优聚类结构的时刻 t[14-15]。

3.1　算法框架

根据上述分析，本文提出一种新型的动态迭代算法，用来有效地探测网络聚类结构。其通过循环地迭代归属向量 x_i^q 来得到理想的划分。

算法 1　基于离散势能论的半监督聚类算法

输入：邻接矩阵 $\mathbf{A}$，最大迭代次数 R_{max}

输出：最优的聚类归属矩阵 $x_i^q(t)$

1. 对于每一个节点 i，设置一个初始状态 $x_i^q(0)=\frac{1}{c}$，$c>\frac{N}{2}$。

2. 根据式(10)和式(12)，计算网络中每一对节点的 W_{ij} 和 T_{ij}。

3 循环迭代动态系统 $x_i^q(t)$，直到目标函数 F 达到最大。

4. 当 $t\geqslant R_{max}$ 时，停止运算；否则 $x_i^q(0)\leftarrow x_i^q(t)$，返回第 1 步。

5. 结束。

在每一次迭代中，将权重 W 和紧密度 T 引入循环。权重 W 可以使得信息在选择路径时变得明确，也就是说在同一聚类内节点连接越紧密，则信息传输速度越快。另外基于相似性的分析，紧密度 T 越大，则节点的位置越接近，因此在聚类的过程中，T 也有助于加速归属向量的迭代过程。

3.2　稳定性指标

算法 1 的一个重要步骤是搜索使得目标函数到达最大值的时间 t。然而，并不是所有的指标函数都能完美地适合此项任务，比如模块度 Q 不包含时间刻度并会将导致分辨率限制和极端退化问题。为了解决这些问题，我们使用由 Delvenne 等[12-13]提出的稳定性指标(stability)作为量度标准。

定义 Markov 随机过程的稳态分布为 $\pi=\frac{d}{2m}$，相应的对角矩阵可以写为 $\boldsymbol{\Pi}=\mathrm{diag}(\pi)$。基于节点 i 的归属向量 $\boldsymbol{x}_i$，我们定义一个以时间 t 为分辨率参数的自协方差矩阵：

$$\boldsymbol{R}_t=\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{\Pi}(M^t-\pi^{\mathrm{T}}\pi)\boldsymbol{X} \tag{14}$$

其中，$\boldsymbol{X}=(\boldsymbol{x}_i)$ 是归属矩阵，正好与本文以归属矩阵为主要构成元素相吻合。$\boldsymbol{R}_t$ 使用时间 t 作为内在的分辨率参数，通过调整 t 的大小可以识别不同规模和数目的网络聚类。最优的聚类结构会使一个随机游走者(random walker)停留在同一聚类比在不同聚类之间转移花费更多的时间。因此可以定义稳定性指标：

$$\begin{aligned}F&=H(X,t)=\min_{0\leqslant s\leqslant t}\sum_{i=1}^{c}(\boldsymbol{R}_s)_{ij}\\&=\min_{0\leqslant s\leqslant t}trace[\boldsymbol{R}_s]\end{aligned} \tag{15}$$

对于一个特定的归属矩阵 $\boldsymbol{X}$，我们可以利用稳定性指标 F 评价 t 时刻网络聚类划分的质量，也就是说，可以通过 F 自然地评价任一时间 t 网络聚类的优劣。因此，我们将稳定性指标 F 用于算法 1 中，用来寻找使聚类划分达到最优的时刻 t。

3.3　计算复杂性

本文算法的计算复杂度主要集中在归属向量 $x_i^q(t)$ 的动态循环过程以及搜索稳定性指标 F 达到最大值的时刻 t 上。通过运行迭代算法，可以通过每一对节点相应的 W 和 T 来更新归属向量 $x_i^q(t)$。此过程中的复杂度是 $O(n\log n)$，其中 n 为节点数。接下来，通过遍历所有节点找到最大 F 值的时刻 t 至少需要 $O(n^2)$。事实上，加权迭代过程等价于广度优先搜索(BFS)，因此算法的总时间复杂度为 $O[n(n+\log n)]$。但是对于较为稀疏的网络来说，计算复杂度将会低于 $O(n^2)$。

4　实验

本节将本文算法应用到人工基准网络及现实社会网络中，实验目标为：1)检验算法的精确性；2)调整参数，使得算法能够探测不同规模的聚类结构。

4.1　LFR 基准网络

首先，通过在人工网络上与一些著名算法进行比较，来验证本文动态迭代算法(用 DI 来表示)的有效性。比较算法包括：GN 算法[6]、Infomap 算法[11]、RB 算法[16]、Louvain 算法[17]和 GA 算法[18]。我们在 Lancichinetti 等提出的 LFR 基准网络[9]上进行验证，该网络拥有无标度形式的度和聚类规模分布，这要比其他基准网络测试更具有代表性。LFR 网络由一些系数控制生成，包括节点数 N、平均节点度 $\langle k\rangle$、最大节点度　混合比率 μ　最小聚类的规模和最大聚类规模。μ 在 $[0,1]$ 范围内，用来决定网络聚类的模糊程度；μ 越大，网络聚

类越模糊。测试中，设定以下默认参数的配置：$N=5000$，$\langle k\rangle=15$，$max_k=50$，$\min_k=20$，$\max_c=50$。

图 3 给出了每个算法的运行情况，其中 x 轴表示网络规模，y 轴表示算法运行时间。从图 3 可知其他 5 个算法的运行时间近似呈指数增加，然而无论节点数目如何增加，本文算法的运行时间总是小于 20 s。接下来，对各个算法进行稳定性测试，实验结果如图 4 所示。可以看出，除了在小型网络上本文算法的稳定性比 GA 算法略低，在其他情况下本文算法的稳定性指标比其他算法都高。另外，我们还利用 NMI 指标来评估划分的优劣，结果如图 5 所示，其中 x 轴表示混合参数 $1-\mu$，用来控制网络中聚类结构的模糊程度，y 轴表示 NMI 值。可以直观地看出，当 $1-\mu$ 位于 $(0.4,0.674)$ 范围内时，GA 和 DI 的精度高于其他算法，当 $1-\mu$ 逐渐高于 0.7 时，本文算法的运行效果优于其他算法。图 3－图 5 中，每个点为 20 次运算结果的平均值。

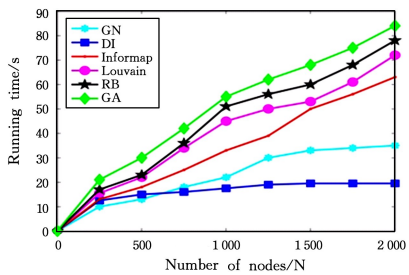


图 3 LFR 网络上 6 种算法运行时间的比较结果

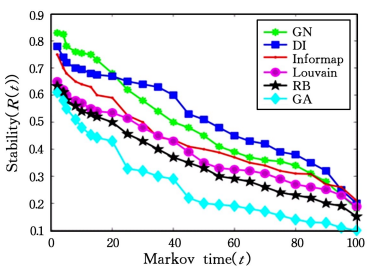


图 4 LFR 网络上 6 种算法的稳定性比较结果

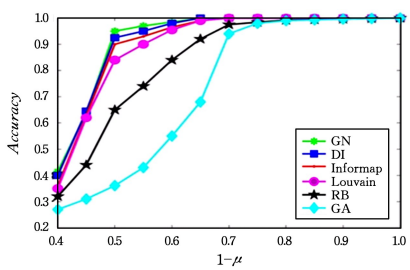
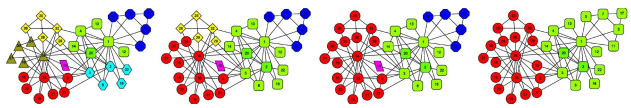


图 5 LFR 网络上 6 种算法的准确性比较结果

4.2 Zarchary 的空手道俱乐部网络

在 20 世纪 70 年代初，Wayne 观察了一家美国大学空手道俱乐部成员之间的社会关系变化情况[20]。基于俱乐部成员的社会活动，他构建了成员之间的关系网。由于俱乐部的管理者与首席空手道老师就是否提高俱乐部的费用产生纠纷，致使成员以管理者和老师为中心一分为二，即第一个是以老师为中心的 16 个节点（节点 1－8，11－14，17－18，20 和 22）的集合，第二个是以管理员为中心的剩余 18 个节点的集合。

为了确保结果的准确性，我们首先设定一个足够大的初始聚类个数 $c=8$，接下来迭代每个节点的归属向量 $\boldsymbol{x}_i$ 使得稳定性函数 F 达到最大值。令 $\alpha=0.381$，$\beta=0.513$，$\eta=1.33$，$\theta=1.5$，通过迭代动态系统(13)最大化 F 达到 0.444，结果如图 6(a)所示。可以观察到，该网络被划分成 7 个聚类，特别地，节点 10 被归类为一个单独的聚类，因为它很难被检测到归属于其他网络聚类。如图 6(b)所示，当 α 增大到 0.395，β 增大到 0.521，η 增加到 1.4，θ 增大到 2.0 时，所检测的网络聚类的数目减少到 5。从图中可以看出，三角形的节点被吸收到圆形的网络聚类。如图 6(c)所示，随着算法运行，除了节点 10 以外，一些小的聚类被吸收到标记为圆形和正方形的两个主要的聚类中，这两个类的核心为节点 1 和 33。最终，网络被分成两个聚类，节点 10 也被纳入右侧的圆形的网络聚类。如图 6(d)所示，此时 F 达到最佳值 0.75(与正确划分完全一致)。通过上例可以证明，本文算法能够以多尺度的方式[21-22]有效地进行不同规模的网络聚类划分。



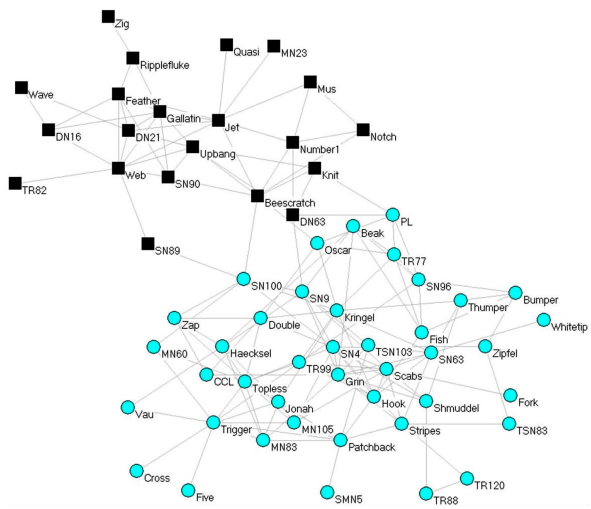
(a)$F=0.444$ (b)$F=0.511$ (c)$F=0.613$ (d)$F=0.75$

注：网络中不同聚类的节点用不同形状表示

图 6 不同参数下空手道俱乐部网络的聚类结果

4.3 海豚关系网络

海豚关系网络是 Lusseau 在 1994 年到 2001 年之间创建的，其中的 62 个节点代表宽吻海豚，159 条边代表海豚之间比随机更经常出现的关联[5-6]。由于其中一只海豚（记为 $SN100$）的暂时消失，海豚分成了两组，海豚网络两个聚类之间的团间边共有 6 条。我们将算法应用到海豚网络中，结果如图 7 所示。



注：不同聚类中的节点用不同形状和颜色表示

图 7 海豚关系网络的聚类结果(电子版为彩色)

图 7 中 $\alpha=0.395$，$\beta=0.521$，$\eta=1.4$，$\theta=2.0$，不同聚类的节点用不同的形状和灰度表示，其中第一个聚类用蓝色的圆形表示，包括 21 个节点，第二个聚类用黑色的正方形表示，共包括 21 个节点。可以发现除了节点 $SN89$，两组海豚都可以正确分类[23]。

4.4 法律案例关联网络

最后，我们将本文方法应用到一种典型的社会网络——法律案例关联网络中。为了满足分析验证，首先要收集案例并构造拥有一定规模的法律案例网络，使得分析具有显著的指导意义。选取最高人民法院 2012 年至 2017 年发布的 17 批共 92 件指导性案例和经过细致的收集和整理筛选出的 100 件最

高人民法院公报案例，其涵盖刑法、民法和行政法三大领域。其中法律案例都包含三大部分，每个部分包含若干属性，即背景属性（如自然地理、社会环境、人为属性）、事件属性（时间、空间、行为）和司法属性（司法主体、司法措施和司法客体）。为了保证网络的稀疏性，我们判定如果两个案例的相同属性超过 5 个，即为相似案例，其之间具有一条连边。我们构造出具有 192 个节点的网络，其中点代表法律案例（法律事件），连边代表法律事件之间的关联，其网络拓扑结构如图 8 所示。

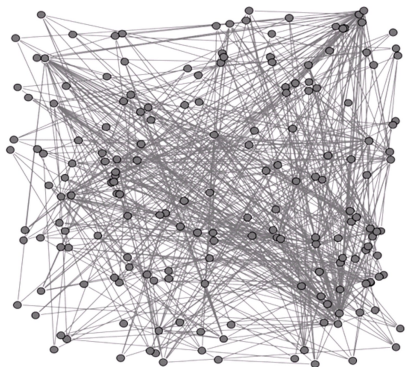


图 8　法律案例关联网络的拓扑结构

利用构造的法律案例关联网络，我们分析了网络中的基本参数，如平均加权度（节点与其他节点相关联的边权和）、聚类系数和平均加权路径长度，并利用本文算法进行划分，得到相应的模块度 Q 值，结果如表 1 所列。从参数分析结果来看，法律案例关联网络是一种典型的无标度（scale-free）网络[1]，与很多现实社会网络具有相似的结构和拓扑性质。通过模块度 Q 值可以看出，本文算法具有较高的效率。

表 1　网络参数

	平均加权度	聚类系数	平均路径长度	模块度 Q
本文算法	2.537	0.694	3.636	0.542

进一步可以发现，关联网络被划分为 4 个社团，具有非常强的社团特征。这些社团分别代表法律案例的 4 个研究领域：刑法、民商法、经济法和行政法，符合实际的法律研究领域。这说明本文算法能够划分功能社团，从而有助于找出其中的内在特性，对进一步的研究具有非常大的价值。

结束语　本文提出一个新的动态迭代聚类算法，通过引入包含拓扑信息的权重 W 和紧密度 T，从而调整边权和节点紧密度，以提高网络聚类结构检测的速度与准确度。为了估计最优的迭代停止时间，我们利用以时间 t 为分辨率参数的稳定性指标作为测度指标，来自然地寻找使聚类划分达到最优的时刻 t。值得一提的是，在实际的应用中，聚类数量这一信息必须提前给出，比如 K-means 算法，这是一个非常严格的约束；但是本文算法不需要预先指定聚类的数目，因此可以方便地应用于各种模糊网络中。

参 考 文 献

[1] ALBERT R, Albert-LÁszlÓ B I. Statistical mechanics of complex networks [J]. Reviews of Modern Physics, 2002, 74(1): 47.

[2] ROMUALDO P S, VÁZQUEZ A, VESPIGNANI A. Dynamical and correlation properties of the Internet [J]. Physical review letters, 2001, 87(25): 258701.

[3] PODANI J, OLTVAI Z N, JEONG H, et al. Comparable system-level organization of Archaea and Eukaryotes [J]. Nature Genetics, 2001, 29(1): 54-56.

[4] LEON D, DÍAZ-GUILERA A, ARENAS A. The effect of size heterogeneity on community identification in complex networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2006, 11: P11010.

[5] FORTUNATO S. Community detection in graphs [J]. Physics Reports, 2010, 486(3): 75-174.

[6] 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.

[7] SANTO F, BARTHÉLEMY M. Resolution limit in community detection [J]. Proceedings of the National Academy of Sciences, 2007, 104(1): 36-41.

[8] ALIREZA K, AJDARIRAD A, HASLER M. Network community-detection enhancement by proper weighting [J]. Physical Review E, 2011, 83(4): 046104.

[9] NEWMAN M E J. Fast Algorithm for Detecting Community Structure in Networks [J]. Physical Review E, 2004, 69: 066133.

[10] ZHOU H J. Distance, dissimilarity index, and network community structure [J]. Physical Review E, 2003, 67(6): 061901.

[11] ROSVALL M, BERGSTROM C T. Maps of Random Walks on Complex Networks Reveal Community Structure [J]. Proceedings of the National Academy of Sciences, 2008, 105(4): 1118-1123.

[12] DELVENNE J C, YALIRAKI S N, BARAHONA M. Stability of graph communities across time scales, eprint [J]. arXiv: 0812.1811.

[13] RENAUD L, DELVENNE J C, BARAHONA M. Laplacian dynamics and multiscale modular structure in networks [J]. arXiv, 2008, 0812.1770.

[14] 李慧嘉，李爱华，李慧颖. 社团结构迭代快速探测算法 [J]. 计算机学报，2017，40(4)：970-984.

[15] 李慧嘉，严冠，刘志东，等. 基于动态系统的网络社团线性探测算法 [J]. 中国科学：数学，2017，47：241-256.

[16] REICHARDT J, BORNHOLDT S. Detecting fuzzy community structures in complex networks with a Potts model [J]. Physical Review Letters, 2004, 93(21): 218701.

[17] VINCENT D, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2008(10): P10008.

[18] GUIMERA R O, NUNES A L A. Functional cartography of complex metabolic networks [J]. Nature, 2005, 433(7028): 895-900.

[19] LANCICHINETTI A, FORTUNATO S. Community detection algorithms: a comparative analysis [J]. Physical Review E, 2009, 80(5): 056117.

[20] ZACHARY W W. An information flow model for conflict and fission in small groups [J]. Journal of Anthropological Research, 1977: 452-473.

[21] 李慧嘉，李慧颖，李爱华. 多尺度的社团结构稳定性分析 [J]. 计算机学报，2015，38(2)：301-312.

[22] LI H J, ZHANG X S. Analysis of Stability of Community Structure Across Multiple Hierarchical Levels [J]. Europhysics Letters, 2013, 103: 58002.

[23] AARON C, NEWMAN M E J, MOORE C. Finding community structure in very large networks [J]. Physical Review E, 2004, 70(6): 066111.