doi:10.3969/j.issn.1000-436x.2014.03.002

基于 Laplace 矩阵 Jordan 型的复杂网络聚类算法

牛建伟，戴彬，童超，霍冠英，彭井
（北京航空航天大学 虚拟现实技术与系统国家重点实验室，北京 100191）

摘 要：在目前复杂网络聚类算法中，基于 Laplace 特征值的谱聚类方法具有严密的数学理论和较高的精度，但受限于该方法对簇结构数量、规模等先验知识的依赖，难以实际应用。针对这一问题，基于 Laplace 矩阵的 Jordan 型变换，提出了一种先验知识的自动获取方法，实现了基于 Jordan 矩阵特征向量的初始划分。基于 Jordan 型特征值定义了簇结构的模块化密度函数，并使用该函数和初始划分结果完成了高精度聚类算法。该算法在多个数据集中的实验结果表明，与目前主流的 Fast-Newman 算法、Girvan-Newman 算法相比，基于 Laplace 矩阵 Jordan 型聚类算法在不依赖先验知识的情况下，实现了更高的聚类精度，验证了先验知识获取方法的有效性和合理性。
关键词：复杂网络；聚类算法；Laplace 矩阵；Jordan 型；先验知识获取
中图分类号：TP301.6 **文献标识码**：A **文章编号**：1000-436X(2014)03-0011-11

Complex network clustering algorithm based on Jordan-form of Laplace-matrix

NIU Jian-wei, DAI Bin, TONG Chao, HUO Guan-ying, PENG Jing
(State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191)

Abstract: Among existing clustering algorithms, the graph-Laplacian-based spectrum clustering algorithm has rigorous theoretical basis and high accuracy. However, the application of this algorithm is limited by its dependence on the prior knowledge, such as the number and the size of clusters. Based on the Jordan form of graph Laplacian, an algorithm was proposed which can obtain the prior knowledge, and perform the primary clustering based on the eigenvalues of the Jordan form. The modularity density function of clusters was defined, and an improved spectrum clustering algorithm with the help of the function and the primary clustering was proposed. The experiments were conducted on diverse datasets showing that, compared with the classic algorithms such as Fast-Newman and Girvan-Newman, the algorithm can reach a high clustering accuracy and a fast convergence rate.
Key words: complex network; clustering algorithm; Laplace-matrix; Jordan-form; prior knowledge

1 引言

现实世界中的许多系统，例如，因特网、移动电话网、蛋白质交互网、神经元网等都是一种复杂网络[1]。由于这类网络中节点的异构性，簇结构成为复杂网络最普遍和最重要的拓扑结构属性之一，具有簇内节点连接紧密、簇间节点连接稀疏的特点。研究复杂网络聚类算法并准确揭示真实的网络簇结构是分析复杂网络中节点关系随时间的演化过程、信号或信息在网络中的传播速度与范围，以及预测网络中节点的行为[2,3]等众多问题的理论基础。同时，聚类算法已被应用于社会网络分析与组织管理[4]、未

收稿日期：2012-10-17；**修回日期**：2013-04-11
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2013CB035503）；国家自然科学基金资助项目（61170296，61190125）；国家高技术研究发展计划（“863”计划）基金资助项目（2012BAH07B01, 2013BAH35F01）；北京市自然科学基金资助项目（4123101）
Foundation Items: The National Basic Research Program of China (973 Program) (2013CB035503); The National Natural Science Foundation of China (61170296, 61190125); The National High Technology Research and Development Program of China (863 Program) (2012BAH07B01, 2013BAH35F01); The Natural Science Foundation of Beijing (4123101)

知蛋白质功能预测[5]、主控基因识别以及 Web 社区挖掘和搜索引擎等众多领域，具有广阔的应用前景。

针对复杂网络中分簇问题，国内外的研究人员设计并实现了多种聚类算法。2002 年，Flake 等人基于最大流—最小截定理提出了启发式聚类算法——MFC (maximum flow community)算法[6]。MFC 算法通过计算最小截集，发现网络“瓶颈”，删除簇间连接，将网络逐渐分割成一个簇集合。同年，Girvan 和 Newman 提出了 GN 算法[7]，该算法通过反复计算网络中的边介数，识别并删除簇间连接，生成一棵自顶向下的层次聚类树。2 种算法基于某些启发式规则设计，虽然能够快速地找到近似最优解，但都缺少严格的理论保证。

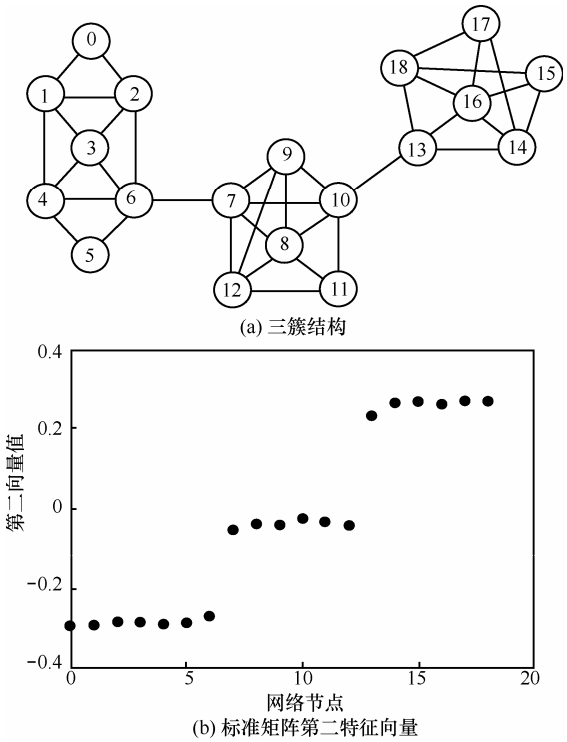
2004 年，Newman 提出了的 FN 算法[8]，该算法是一种优化算法，优化目标为网络模块性评价函数[9]（或称 Q 函数）。初始状态下，FN 算法将每一个节点看作一个簇，通过在迭代过程中最大化 Q 函数的合并操作，计算出自底向上的簇结构关系树。基于 Q 函数，Guimera 和 Amaral 提出了融合模拟退火算法——GA(guimera-amaral)算法[10]，该算法通过计算候选解对应的 Q 函数值来评价其优劣，并通过模拟退火策略 Metropolis 准则决定是否接受候选解。然而，近年的 Q 函数已被证明是有偏的[11]，有偏的目标函数必然导致有偏的聚类结果。有学者针对 Q 函数进行了改进[12]，但聚类效果均没有显著提高。

复杂网络一般使用图来描述和分析，而 Laplace 矩阵能准确地表征图的拓扑特性。因此，许多学者采用 Laplace 矩阵来研究复杂网络的聚类过程，目前，这类算法主要包括谱平分法和谱分解法[13~16]。谱平分法使用 Laplace 矩阵第二小特征值所对应的特征向量对网络进行切分，在每次二分过程中，网络被分割成 2 个近似平衡的子网络。当网络中含有多个簇时，谱平分法递归地分割现存的子网络，直到满足预先定义的停止条件为止。谱分解法则基于 Laplace 矩阵前 n 个接近 0 的特征向量组建一个 n 维子空间，通过全网节点向该空间的投影完成一次簇结构划分。2 种算法虽然都基于 Laplace 矩阵实现了复杂网络的聚类，但都具有以下不足。

1) 谱平分法需要借助先验知识定义递归终止条件，即谱方法不具备自动识别网络簇总数的能力。此外，谱平分法的递归二分策略在研究多簇结构的网络或结构不对称的网络时，不能保证得到正确的网络划分。

2) 谱分解法可以看作谱平分法的一种改进，虽然消除了递归二分策略产生的不利影响，但在无先验知识情况下不能确定簇结构个数，即无法获得投影子空间的维度。

为了克服谱聚类方法对先验知识的依赖，Capocci 等人在传统谱方法的基础上提出了一种新的聚类算法[17]，该算法利用网络标准矩阵第二特征向量呈现出的阶梯状来提取先验知识（如图 1 所示）。但该方法仅对簇结构明显的网络有效，当网络的簇结构不明显时，第二特征向量不具备明显的阶梯状，而接近于一条连续曲线，此时，无法依赖第二特征向量中的元素来对网络进行划分了。



(a) 三簇结构

(b) 标准矩阵第二特征向量

图 1 三簇结构及其对应的标准矩阵第二特征向量

针对以上问题，本文分析发现，现有的基于 Laplace 矩阵的谱聚类算法都是将高维空间投影至低维空间，其核心算法均使用了 Laplace 矩阵的特征值及其对应的特征向量。然而，特征值和特征向量并不能完全代表 Laplace 矩阵的所有信息，这种“有损”的信息提取使得 2 种算法都依赖先验知识来判断对特征值或特征向量的取舍，同时也降低了聚类算法对真实簇结构的还原精度。

基于以上分析，本文首次将矩阵相似的概念引入到基于 Laplace 矩阵的聚类算法中，实现对 Laplace 矩阵信息的“无损”利用。相似是一种特殊

的线性变换，经过相似变换，原矩阵的特征值及其代数重数不会发生变化，并且变换矩阵中包含了所有特征向量和其结构信息等。其形式如式(1)所示。

$$\boldsymbol{A}=\boldsymbol{P}^{-1}\boldsymbol{B}\boldsymbol{P} \tag{1}$$

Laplace 矩阵的 Jordan 型是 Laplace 矩阵所在相似类 $\boldsymbol{A}\in\boldsymbol{F}^{(mn)}$ 中最简单的方阵，因此 Jordan 型能够完整体现 Laplace 矩阵中包含的拓扑结构各项属性，故本文通过 Laplace 矩阵 Jordan 型来讨论原网的某些聚类性质。

2　基于 Jordan 型先验知识获取

复杂网络可以建模为一个图 $G=(V,E)$，其中 V 表示网络的节点集合，E 表示连接的结合。例如，对于复杂社会网络而言，每个节点代表一个人，边则表示人与人之间存在的关系。复杂网络的拓扑信息都可以由图的 Laplace 矩阵代表。Laplace 矩阵的一种计算形式为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，其中 $\boldsymbol{L}$ 为图的 Laplace 矩阵，$\boldsymbol{D}$ 为该图的度矩阵，$\boldsymbol{W}$ 为该图的邻接矩阵。Laplace 矩阵具有以下的性质。

1) 任何行列之和均为 0，因此(1, 1,⋯, 1)总为其特征向量。

2) 如果该图能完全分成 g 个簇，则 $\boldsymbol{L}$ 可以写成分块对角阵。

3) $\boldsymbol{L}$ 的所有特征值非负且为实数。

4) $\boldsymbol{L}$ 为实对称矩阵。

5)实对称阵不相同的特征向量正交，因此除(1,1,⋯,1)外其余特征向量均含有正负元素。

2.1　Jordan 型的图论特征分析

基于对 Laplace 矩阵性质的分析发现，Jordan 型矩阵作为复杂网络对应的 Laplace 矩阵的相似矩阵，具有下面的特征。

定理 1　若实对称方阵 $\boldsymbol{S}$ 正交相似于对角矩阵 $\boldsymbol{D}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$，则对角元 $(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 是 $\boldsymbol{S}$ 的全体特征值。

定理 2　实对称方阵 $\boldsymbol{S}$ 的特征值全部都是实数。

根据定理 1 和定理 2，Laplace 矩阵相似于对角阵 $\boldsymbol{D}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$，其中，$(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 是 Laplace 矩阵的全体特征值。这里的 $(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 均为不小于 0 的实数。为方便讨论，本文对 $(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 按数值升序排列，且有 λ_1 恒等于 0，下文中的 Jordan 型对角阵均默认为排序后的形式。基于以上定理，可以得出假设 1。

假设 1　若拓扑结构图能够完全划分成 g 个簇，则有 $(\lambda_1,\lambda_2,\cdots,\lambda_g)$ 均为 0，即该拓扑结构图对应的 Laplace 矩阵 Jordan 型对角线上有 g 个 0。

这一假设是基于 Jordan 型获取先验知识的重要理论基础。本文通过 2 个引理来证明这一假设。

引理 1　若拓扑结构图不能完全划分成若干个簇，则 Laplace 矩阵只有一个特征值为 0 的特征向量。

证明　若拓扑结构图不能完全划分为若干个簇结构，则其为连通图[18]；图 G 是连通的当且仅当 G 的第二个最小的 Laplace 特征值 $\lambda_{n-1}>0$。

证毕

由引理 1 可知，Laplace 矩阵中只有一个特征值为 0 的特征向量，说明特征值 0 在 Jordan 型中只对应一个分块。

引理 2　Laplace 矩阵 Jordan 型中的节点排序不影响矩阵的相似标准型。

证明　矩阵节点顺序的变化可通过公式 $\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}$ 实现，其中，$\boldsymbol{P}$ 为初等行变换或列变换矩阵，即节点顺序的安排不会影响到 Laplace 矩阵的 Jordan 型。

证毕

由引理 2 可知，在分块对角阵形式的图 Laplace 矩阵 Jordan 型中，每一分块矩阵都不能再被完全分开。

假设 1 证明　Laplace 矩阵 Jordan 型为对角阵，由“相似保证矩阵秩不变”的性质可知，分块对角阵（其每一分块均为方阵）中每一块方阵都奇异，秩为 $m-1$ (方阵的规模为 mm)。又因(1, 1,⋯, 1)为 Laplace 矩阵特征向量，其对应的分块有且仅有这一个特征向量。依此规律，(0, ⋯, 0, 1, ⋯, 1), (0, ⋯, 0, 1, ⋯, 1, 0, ⋯, 0),⋯,(1, ⋯, 1, 0, ⋯, 0)形式的向量构成了特征值 0 所对应的特征子空间。若原图能分成 g 个簇，则此类形式的向量有 g 个。又因 Laplace 矩阵 Jordan 型为对角阵，所以，Jordan 型中 $(\lambda_1,\lambda_2,\cdots,\lambda_g)$ 均为 0。

证毕

通过对 Jordan 型图论特征的分析和证明，本文得出以下结论：Laplace 矩阵 Jordan 型中对角线特征值“0”的个数等于拓扑结构图中可完全划分开的簇的个数。

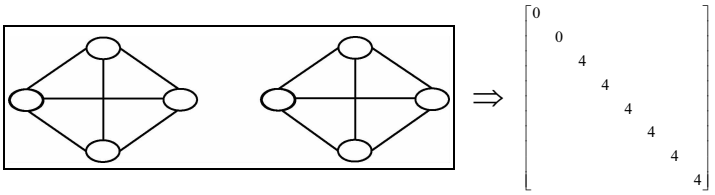
2.2　Jordan 型的聚类特征分析

基于对 Jordan 型图论特征的结论，本文进一步对 Jordan 型聚类特征进行了分析。

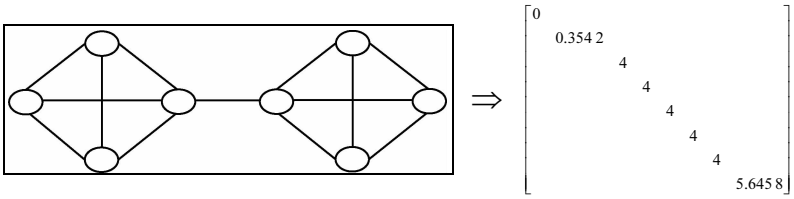
复杂网络的拓扑结构与其 Laplace 矩阵具有一一对应的关系，而 Laplace 矩阵 Jordan 型是对

Laplace 矩阵的“无损”变换，与原复杂网络拓扑结构也具有一一对应的关系。因此，拓扑结构中的聚类属性可以准确映射到其 Laplace 矩阵 Jordan 型中，如图 2 所示。

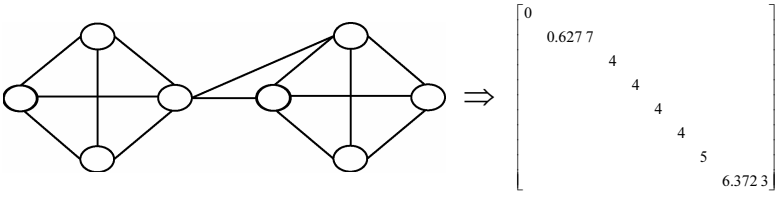
通过图 2 可以发现，初始状态（如图 2(a)所示），Laplace 矩阵 Jordan 型中 2 个数值为 0 的特征值，对应拓扑结构中 2 个可以完全划分的簇结构；随着拓扑结构图中 2 个全连通分支之间连边的不断增加，Laplace 矩阵 Jordan 型中的第二小特征值对应增大；当特征值增大到一定程度时（如图 2(e)和图 2(f)所示），原拓扑结构中明显的簇结构属性消失。因此，Laplace 矩阵 Jordan 型准确描述了原拓扑结构图的聚类属性，其 Jordan 型中特征值越小，对应原图的簇结构越明显。



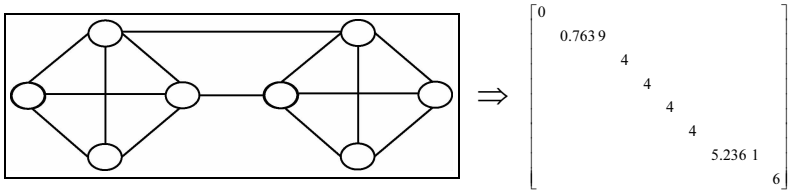
(a) 拓扑结构 1 及其 Laplace 矩阵 Jordan 型



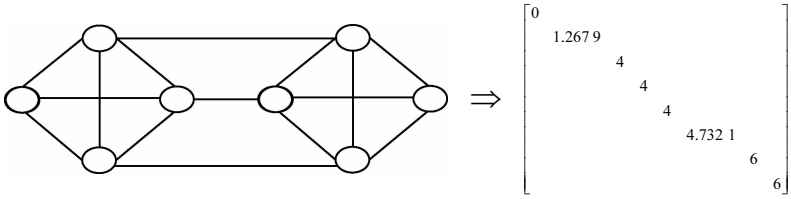
(b) 拓扑结构 2 及其 Laplace 矩阵 Jordan 型



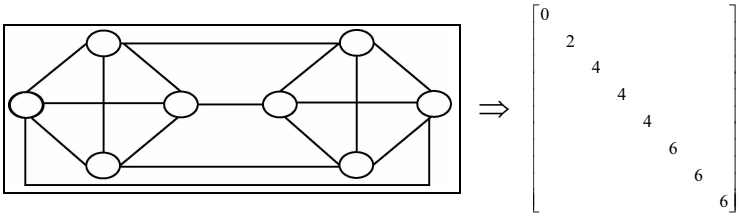
(c) 拓扑结构 3 及其 Laplace 矩阵 Jordan 型



(d) 拓扑结构 4 及其 Laplace 矩阵 Jordan 型



(e) 拓扑结构 5 及其 Laplace 矩阵 Jordan 型



(f) 拓扑结构 6 及其 Laplace 矩阵 Jordan 型

图 2　不同拓扑结构对应 Laplace 矩阵 Jordan 型

2.3　基于 Jordan 型的先验知识获取模型

基于对 Laplace 矩阵 Jordan 型聚类特征的分析，本文提出了基于 Jordan 型的先验知识获取模型，通过该模型，可以实现对复杂网络簇结构数量的自动获取和对拓扑结构的初始划分，为聚类算法提供先验知识和初始解。其主要步骤如图 3 所示。

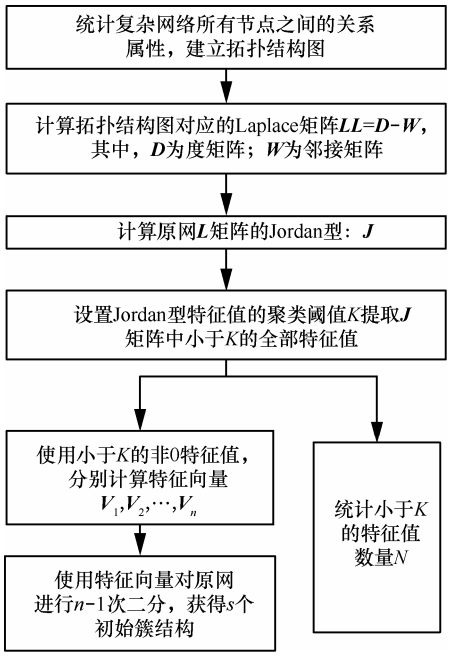


图 3　基于 Jordan 型的先验知识获取模型

模型中，聚类阈值 K 是判断簇结构数量并对复杂网络进行初步划分的关键参数。本文针对这一参数的确定进行了分析和实验。

首先，从图 1 中可以发现，Laplace 矩阵 Jordan 型按行或列均为对角占优阵（即对角线上元素大于等于该行或该列其余元素的绝对值之和）。对于严格的对角占优阵，矩阵是满秩的，并且所有的特征值在区间[0,1]之中；对于非严格对角占优阵，它的秩至多为 $n-1$，此时部分特征值大于 1。

其次，大量实验表明，Laplace 矩阵 Jordan 型中的特征值“1”对应了拓扑结构图中簇结构的临界状态。如图 4 中拓扑结构所对应的 Laplace 矩阵 Jordan 型为 diag(0,1,6,6,6,6,7)，特征值“1”所对应的特征向量，既可以将单独节点分开，划分为 2 个独立的簇结构；也可以将单独的节点放入由其他 4 个节点构成的全连通分支，整体划分为一个簇。

因此，本文设置“1”作为聚类阈值 K，并使用开区间，对应图 4 网络结构为一个整体簇结构，后续实验也都使用该设置。同时，针对不同性质的复杂网络，该参数的设置可以发生变化。

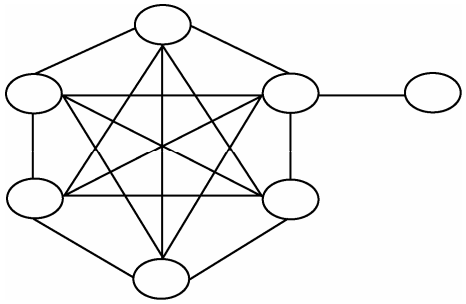


图 4　临界状态下的簇结构

与 Capocci 提出的先验知识提取方法相比，基于 Laplace 矩阵 Jordan 型的先验知识自动获取模型不受网络簇结构明显程度的限制，可以应用于所有拓扑结构。同时，该方法能够通过调整阈值 K，满足不同粗细粒度要求的聚类先验知识提取。

在确定 K 值后，假设有 n 个小于 K 的特征值，根据谱分解的方法对拓扑结构网络实行 $n-1$ 次划分（假设原拓扑结构网络不存在独立的连通分量，因此只使用非 0 的特征值进行 $n-1$ 次划分）。每次划分对应着一个二分簇结果，对所有划分结果进行交集运算，便可得到 s 个初始划分的簇结构。

划分过程中，交集运算共进行 2^{n-1} 次，可以证明得到的不为空的初始簇结构至多为 $n-1$ 个，即 $s \leqslant n-1$。

证明　设第 i 次划分对应的结果是簇 S_{i1}、S_{i2}，须证明 $\bigcup_{i=1}^{m} S_{ij} \mid j=1,2$ ，其中 $S_{i1} \cap S_{i2} = \phi$。

当 2 种交集只有一项取不同时，$\exists k,s,t, A = (\bigcap_{i=1,i\neq k}^{m} S_{ij}) \cap S_{k1}, B = (\bigcap_{i=1,i\neq k}^{m} S_{ij}) \cap S_{k2}$ 故 $A \cap B = 0$，即 A、B 没有公共的元素；当每 2 个交集中至少有一项取不同时，设交集的集合为 S，对于 $\forall A,B \in S$，$A \cap B = 0$，因此，对于有 n 个节点的拓扑结构图，采用以上方法划分簇的个数不大于 n，即得到的不为空的初始簇结构至多为 $n-1$ 个。

证毕

3　基于 Jordan 型的聚类方法

基于 Laplace 矩阵 Jordan 型的先验知识获取，本文定义了簇结构的模块化密度函数，设计并实现了基于 Jordan 型特征值的聚类算法。

3.1　簇的模块化密度

本文定义簇的模块化密度 P，旨在通过 P 函数

从连接密度的角度描述簇结构的相对模块性和独立性。P 函数的定义为簇内相对连接密度与簇间相对连接密度的差，其计算形式如式（2）所示。

$$P=\left[\frac{2m_s}{d_s}-\frac{d_s-2m_s}{\min(d_s,d-d_s)}\right] \tag{2}$$

其中，m_s 表示簇内连接边数，d_s 表示簇内节点度之和，d 表示全网节点数，$\min(d_s,d-d_s)$ 表示簇内节点度之和与簇外节点度之和之间的较小值。

当簇结构为独立的全连通分量时，P 函数取极大值 1，因此，$(1-P)$ 的取值在区间 $[0,\infty)$ 之间，与 Laplace 矩阵 Jordan 型特征值取值空间保持一致，并在一定程度上描述了 Laplace 矩阵 Jordan 型特征值的物理含义。

3.2 基于 Jordan 型的聚类算法

基于先验知识的自动获取和上文定义模块化密度函数 P，本文实现了基于 Jordan 型的聚类算法，算法的核心步骤如下。

step1 计算 s 个初始簇的模块化密度 P。

step2 利用优化思想，对初步划分结果进行合并操作。设置初始簇结构合并过程中的优化目标函数 A 和用于保存最小的 A 值的变量 $\min A$，A 函数的计算如式（3）所示。

$$A=\sum_{i=1}^{s}k_i-(1-\varepsilon)(1-P_i) \tag{3}$$

其中，当 $i\leqslant n$ 时，k_i 即为对应的 Jordan 型特征值；当 $i>n$ 时，$k_{n+1},k_{n+2},\cdots,k_s=1$；$\varepsilon$ 为特征值与密度函数之间的拟合修正系数，针对不同规模复杂网络为不同的常数。

step3 判断当前簇个数 s 与先验知识中簇结构个数 n 是否相等，若相等，当前 s 个簇结构即为原网聚类结果，若 s 大于 n，则列举当前网络中所有的簇结构对 i、j，然后转 step4 执行。

step4 判断当前网络中所有的簇结构对是否都已经被取过，若没有，任取一对没有取过的簇结构对 i、j，若全部被取过，转 step3 执行。

step5 判断簇结构 i 和簇结构 j 之间是否存在连接的边，若存在，执行 step6，若不存在，转 step4。

step6 假定将簇结构 i 和簇结构 j 合并得到新的簇结构 i'，计算合并后全网的目标函数 A，此时 $s'=s-1$。

step7 比较 step6 中得到的目标函数值 A 是否小于当前最小 A 值的变量 $\min A$，若否，不执行合并操作，转 step4，若是，更新 $\min A$ 的值为 step6 中得到的目标函数的 A 值，将 step6 中的簇结构对合并，然后转 step4。

4 实验结果及分析

为客观全面评价算法聚类效果，本文使用基于 Laplace 矩阵 Jordan 型的聚类算法和 FN 算法、GN 算法同时对 4 个不同的数据集 Zachary's Karate Club[19]、Les Miserables[20]、USAir[21]以及 Neural Network[22]进行聚类运算。聚类效果通过模块化函数值 Q[9]和 NCP(network community profile)[23]系统中的评价函数进行综合评价。

4.1 Zachary's Karate Club 数据集实验

空手道俱乐部（zachary's karate club）网络是社会网络分析的一个经典问题。该网络共有 34 个节点，78 条边，分别代表 34 名俱乐部成员和成员之间的社交关系。其具体拓扑结构参数如表 1 所示。

表 1 Zachary's Karate Club 数据集属性

属性	数值
节点数	34
聚类系数	0.316
边数	78
直径	6
三角闭分组数	49
平均最短路径数	2.415

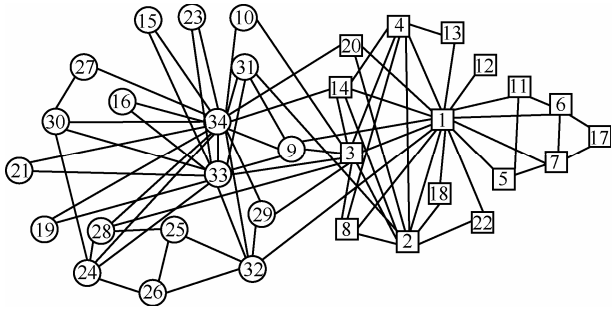
为比较不同算法的聚类效果，本文使用 Q 函数对聚类结果进行量化评价。Q 函数于 2004 年由 Newman 和 Girvan 共同提出，旨在定量刻画网络分簇后的模块化程度。对于由人组成的网络，Q 函数描述了其“人以群分”的程度，而该程度具有一定稳定性，不会因聚类算法的改变而产生大幅度变化。该函数定义为簇内实际连接数目与随机连接情况下簇内期望连接数目之差。计算形式如式(4)所示。

$$Q=\sum_{S=1}^{K}\left[\frac{m_S}{m}-\left(\frac{d_S}{2m}\right)^2\right] \tag{4}$$

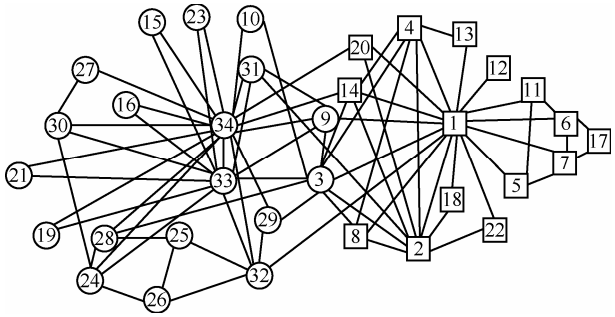
其中，K 表示全网簇结构的个数，m 表示全网连边总数，m_s 表示网络簇 s 中的连边总数，d_s 表示网络簇 s 中节点度之和。

空手道俱乐部网络中，FN 算法将其划分成 2 个社区（如图 5(a)所示），其划分对应的 Q 值为

0.381。GN 算法也将其分成 2 个社区（如图 5(b)所示），对应的 Q 值为 0.411。



(a) 使用 FN 算法对 Zachary's Karate Club 数据集聚类结果



(b) 使用 GN 算法对 Zachary's Karate Club 数据集聚类结果

图 5　FN 算法、GN 算法对 Zachary's Karate Club 数据集聚类结果

应用本文聚类算法，该网络对应的 Laplace 矩阵 Jordan 型中共有 3 个小于 1 的特征值，故判断该俱乐部成员可以分成 3 个社区。根据特征值对应的特征向量，网络成员被初始划分为 5 个社区，分别为

$$\begin{cases} S_1 = \{3,10\} \\ S_2 = \{9,31\} \\ S_3 = \{15,16,19,21,23,24,25,26,27,28,29,30,32,33,34\} \\ S_4 = \{1,2,4,8,12,13,14,18,20,22\} \\ S_5 = \{5,6,7,11,17\} \end{cases}$$

从图 6 可以明显看出，S_3、S_4 和 S_5 对应着 3 个核心社区，具有簇内连接紧密而簇间连接稀疏的特点，符合本文算法的分析和证明。基于本文提出的末端优化聚类方法，初始簇 S_1 与 S_5 合并；簇 S_2 与 S_3 合并，此时优化目标函数获得最小值，网络最终被划分为 3 个社区。

这一划分结果对应的 Q 值为 0.402，处于 FN 算法和 GN 算法分簇结果构成的 Q 值区间[0.381, 0.411]中，这说明：在不依赖先验知识的情况下，基于 Laplace 矩阵 Jordan 型的谱聚类算法能够与 FN、GN 算法给出相近的聚类结果，并准确地描绘出空手道俱乐部“人以群分”的程度。这证明了本文提出了先验知识自动获取模型和末端聚类算法的有效性。

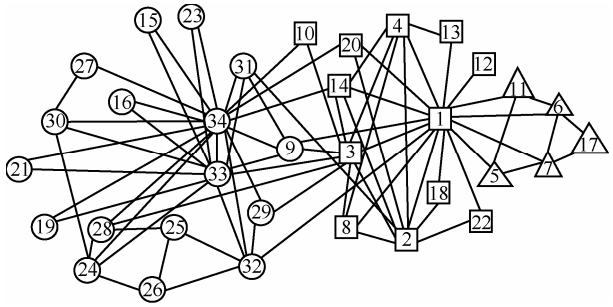


图 6　基于 Laplace 矩阵 Jordan 型的聚类算法对 Zachary's Karate Club 数据集的聚类结果

4.2　Les Miserables 数据集实验

Les Miserables 网络是《悲惨世界》小说中的人物共性网络，全网共有 77 个节点，254 条边，代表了人物之间的关系。其具体拓扑结构参数如表 2 所示，此网络相较空手道俱乐部，社区个数明显增多，划分情况更为复杂。

表 2　　Les Miserables 数据集属性

属性	数值
节点数	77
聚类系数	0.209
边数	254
直径	6
三角闭分组	36
平均最短路径数	3.456

FN 算法将其分为 5 个社区（如图 7 中的 B 所示），对应划分的 Q 值为 0.501；GN 算法将其划成了 11 个社区（如图 7 中的 C 所示），Q 值达到了 0.54，但 GN 算法将 2 个与其他社区有连边的节点分别独立成了 2 个社区。

利用基于 Laplace 矩阵 Jordan 型的聚类算法，该网络对应的 Laplace 矩阵中小于 1 的特征值共有 9 个，故该网可以分成 9 个社区，处于 FN 算法和 GN 算法给出的社区数量之间。聚类算法最终给出的划分结果如图 7 中的 A 所示，其对应的 Q 值为 0.493，与 FN 算法表现出了相似的聚类性能。与 GN 算法相比，该划分结果将 GN 算法聚类结果中的 2 个单独节点合入了与之有连边的社区中，因此 Q 值小于 GN 算法的 0.54。事实上，这 2 个单独的节点分别

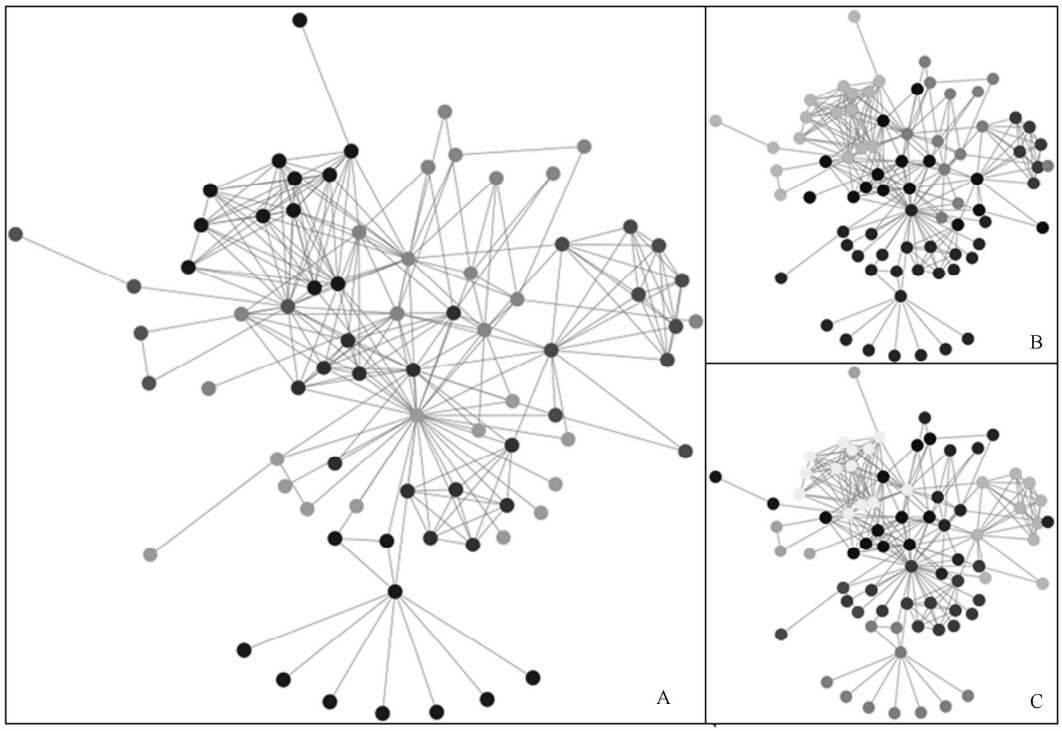


图 7　基于 Laplace 矩阵 Jordan 型的聚类算法、FN 算法和 GN 算法对 Les Miserables 数据集的聚类结果

对应小说中 “Gavroche”和“Bishop Kyriel”2 个人物，通过阅读小说，发现把这 2 个人物合并到有连边的社区中，更加能够反映小说中人物之间的社会关系。因此，提出的基于 Laplace 矩阵 Jordan 型的聚类算法比其他 2 个算法更能反映真实的社区情况。

4.3　USAir 数据集实验

USAir 数据集属于社会系统中的复杂交通网络，网络中共有 332 个节点，每个节点代表了一个美国境内的机场，边则代表了其连接的 2 个机场之间存在实际运行的航线，全图航线总数为 2 126，其具体拓扑结构参数如表 3 所示。FN 算法和 GN 算法对该网络的聚类分析依赖机场在地理位置上的区域划分情况。

表 3　　USAir 数据集属性

属性	数值
节点数	332
聚类系数	0.301
边数	2 126
直径	3
三角闭分组数	2 453
平均最短路径数	2.333

因数据集过于庞大，无法通过直观的分析来评价聚类结果，因此，本文使用斯坦福大学提出的 NCP 来评价聚类结果，NCP 提供了一系列聚类精度评价函数，本文使用其中的 4 个函数，从不同的维度评价本文算法的聚类精度，评价函数分别为 Conductance、Expansion、Cut Ratio 和 Normalized Cut (N-Cut)，其定义如下。

$$\text{Conductance：}\quad f(S)=\frac{c_s}{2m_s+c_s} \tag{5}$$

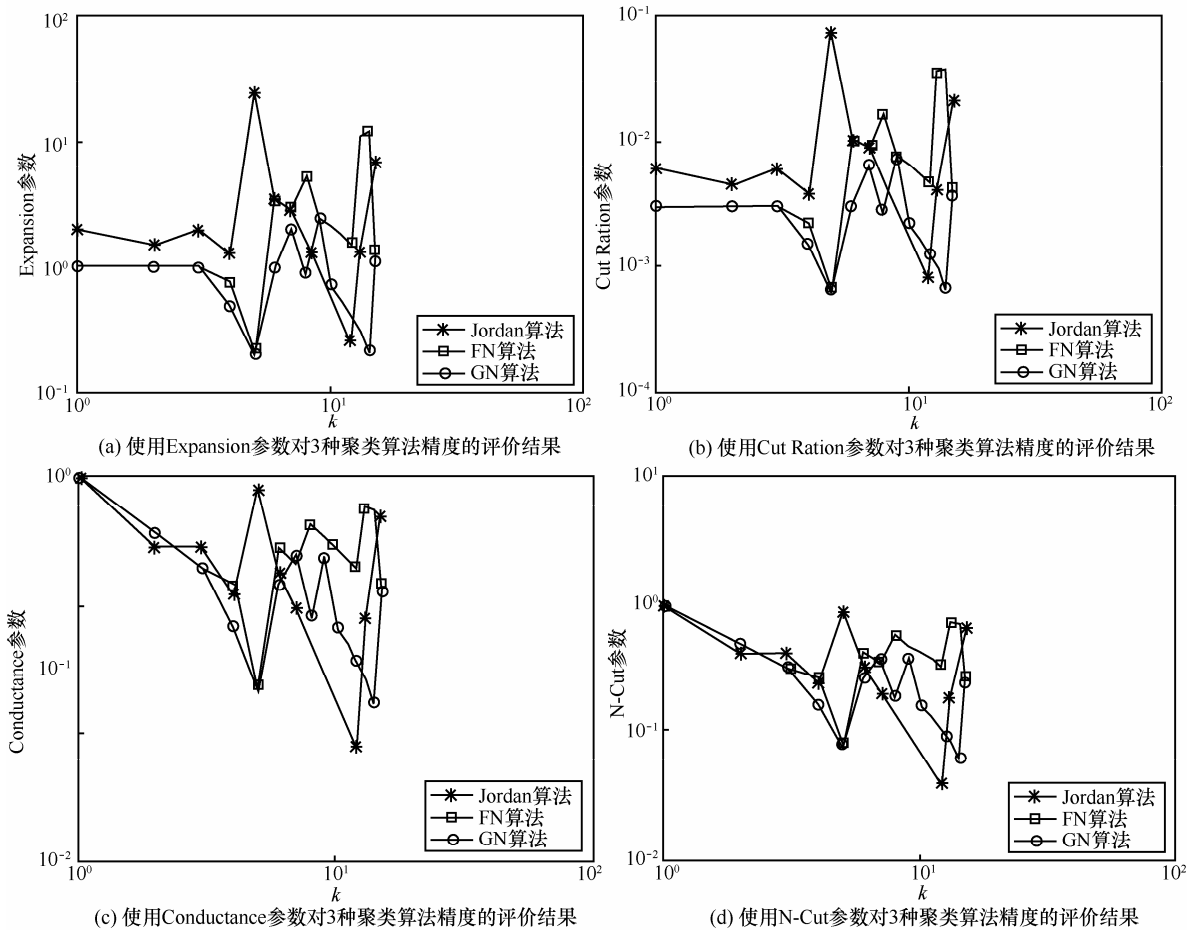
$$\text{Expansion：}\quad f(S)=\frac{c_s}{n_s} \tag{6}$$

$$\text{Cut Ratio：}\quad f(S)=\frac{c_s}{n_s(n-n_s)} \tag{7}$$

$$\text{N-Cut：}\quad f(S)=\frac{c_s}{2m_s+c_s}+\frac{c_s}{2(m-m_s)+c_s} \tag{8}$$

其中，c_s 表示簇 s 内节点与簇 s 外节点连边的总数，m_s 表示簇 s 内的连边总数，n_s 表示簇 s 内的节点总数，m 表示全网总边数，n 表示全网节点总数。评价函数的数值越低，说明聚类精度越高、效果越好。

图 8 显示了 4 个评价函数对基于 Laplace 矩阵 Jordan 型的聚类算法以及 FN 算法、GN 算法在 USAir 数据集聚类结果的评价。4 个评价函数中的平均值最小值显示，基于 Laplace 矩阵 Jordan 型的



(a) 使用Expansion参数对3种聚类算法精度的评价结果　(b) 使用Cut Ration参数对3种聚类算法精度的评价结果

(c) 使用Conductance参数对3种聚类算法精度的评价结果　(d) 使用N-Cut参数对3种聚类算法精度的评价结果

图 8　基于 Laplace 矩阵 Jordan 型的聚类算法、FN 算法和 GN 算法在 USAir 数据集中的聚类精度比较

聚类算法精度比 FN 算法提高了 32%，比 GN 算法精度提高了 2%。此外，在相同的运算平台上，通过对算法连续 5 次的运行计时显示，基于 Laplace 矩阵 Jordan 型的聚类算法的聚类速度（平均 3.7 秒/次）比 GN 算法（平均 8.9 秒/次）提高了一倍以上。

基于 Laplace 矩阵 Jordan 型的聚类算法对于 USAir 数据集的高精度聚类，显示了 332 个机场之间隐含的交通密度关系。这一聚类结果，一方面可以帮助航空管理部门调整机场所属的航管局，以提高管理效率，降低管理风险；另一方面，可以帮助分析航线效率及其合理性，为增加、取消和调整航线提供支持。

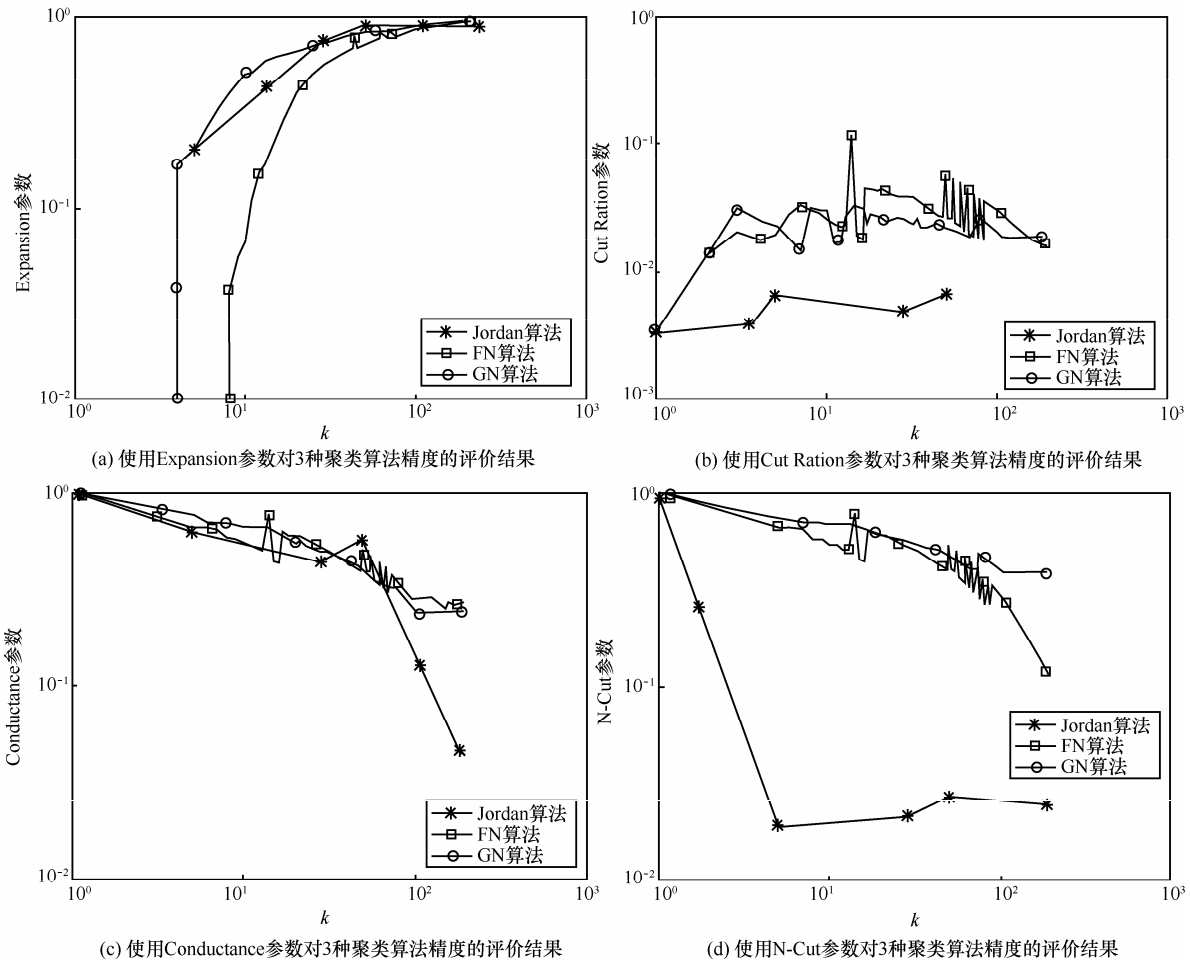
4.4　Neural Network 数据集实验

Neural Network 数据集属于生命系统中的神经元复杂网络，其网络中的节点和边对应真实的医学意义为每个节点代表了一个完整并独立的神经元，边则代表了神经元之间的突触连接。因医学发展水平限制，该数据集在聚类前无法获得任何先验知识，其具体拓扑结构参数如表 4 所示。

表 4　Neural Network 数据集属性

属性	数值
节点数	297
聚类系数	0.292 4
边数	2 359
直径	5
三角闭分组数	3 241
平均最短路径数	2.455 3

图 9 显示了在数据集 Neural Network 中，使用 NCP 函数对 3 种算法聚类效果进行评价的结果。其中，FN 算法聚类结果的 Conductance 平均值为 0.763 3，GN 算法聚类结果的 Conductance 平均值为 0.521，基于 Laplace 矩阵 Jordan 型的聚类算法运行结果的 Conductance 平均值为 0.213。Normalized Cut 函数评价图显示基于 Laplace 矩阵 Jordan 型的聚类算法结果精度比 FN 算法和 GN 算法提高了一个数



(a) 使用Expansion参数对3种聚类算法精度的评价结果

(b) 使用Cut Ration参数对3种聚类算法精度的评价结果

(c) 使用Conductance参数对3种聚类算法精度的评价结果

(d) 使用N-Cut参数对3种聚类算法精度的评价结果

图 9　基于 Laplace 矩阵 Jordan 型的聚类算法、FN 算法和 GN 算法在 Neural Network 数据集中的聚类精度比较

量级。在 Expansion 函数中 3 种算法的评价结果相近。而在 Cut Ratio 评价函数中基于 Laplace 矩阵 Jordan 型的聚类算法精度比 FN、GN 算法明显提高。这一系列结果说明基于 Laplace 矩阵 Jordan 型的聚类算法在无先验知识的情况下，对复杂网络聚类精度的提升效果极为显著。

基于 Laplace 矩阵 Jordan 型的聚类算法对于 Neural Network 数据集的聚类分析，更加精准地发掘出了具有相似功能属性的神经单元，并给出了相似功能神经元之间和相异功能神经簇之间结构关系。这一聚类结果可以帮助医学研究人员更好地了解神经系统工作机理，分析神经类疾病产生的原因，并为神经疾病治愈方法的研究提供理论支撑。

5　结束语

本文提出了一种基于 Laplace 矩阵的 Jordan 型自动获取先验知识的获取模型，实现了对无先验知识复杂网络簇结构数量和密度的自动获取。随后，本文定义了簇结构的模块化密度函数，并利用这一函数和基于 Jordan 型特征值的簇结构初始划分完成了完整的聚类运算。实验证明，基于 Laplace 矩阵 Jordan 型的聚类算法在不同类别和规模的无先验知识网络数据集中均能实现较高的聚类精度，表现出了良好的聚类性能；同时，也验证了本文先验知识自动获取方法的合理性和准确性。

参考文献：

[1] ALESSANDRO V. Complex networks: the fragility of interdependency[J]. Nature, 2010, 464(7291): 984-985.

[2] SONG C M, QU Z H, NICHOLAS B. Limits of predictability in human mobility[J]. Science, 2010, 327:1018-1021.

[3] ALESSANDRO V. Predicting the behavior of techno-social system[J]. Science, 2009, 325:425-428.

[4] DOU B L, LI S S, ZHANG S Y. Social network analysis based on structure[J]. Chinese Journal of Computers, 2012, 35(4):741-753.

[5] YU L, GAO L, SUN P G. Research on algorithms for complexes and functional modules prediction in protein-protein interaction networks[J].

Chinese Journal of Computers, 2011, 34(7):1239-1251.

[6] FLAKE G W, LAWRENCE S, GILES C L. Self-organization and identification of Web communities[J]. IEEE Computer, 2002, 35(3): 66-71.

[7] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proc of the National Academy of Science, 2002, 9(12):7821-7826.

[8] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical Review E, 2004, 69(6):066133.

[9] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical Review E, 2004, 69(2):026113.

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

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

[12] YANG B, LIU D Y, LIU J M, *et al*. Complex network clustering algorithms[J]. Journal of Software, 2009, 20(1):54-66.

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

[14] SHIGA M, TAKIGAWA I, MAMITSUKA H. A spectral clustering approach to optimally combining numerical vectors with a modular network[A]. Proc of the 13th ACM SIGKDD Int'l Conf on Knowledge Discovery and Data Mining[C]. New York, 2007.647-656.

[15] WHITE S, SMYTH P. A spectral clustering approach to finding communities in graphs[A]. Proc of the 5th SIAM Int'l Conf on Data Mining[C]. Philadelphia: SIAM, 2005.76-84.

[16] DONETTI L, MUNOZ M A. Improved spectral algorithm for the detection of network communities[A]. Proc of the 8th Int'l Conf. on Modeling Cooperative Behavior in the Social Sciences[C]. New York, 2005. 104-107.

[17] CAPOCCI A, SERVEDIO V D P, CALDARELLI G, *et al*. Detection communities in large networks[J]. Computer Science, 2004, 3243: 181-187.

[18] FIEDLER M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(98): 298-305.

[19] Supporting Website[EB/OL]. http://www-personal.umich.edu/~mejn/netdata/ karate.zip. 2012.

[20] Supporting Website[EB/OL]. http://www-personal.umich.edu/~mejn/netdata/lesmis.zip. 2012.

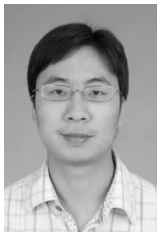
[21] Supporting Website[EB/OL]. http://vlado.fmf.uni-lj.si/pub/networks/data/mix/USAir97.net, 2011.

[22] HANSEN L K, SALAMON P. Neural network ensembles[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1990, 162(10): 993-1001.

[23] JURE L, KEVIN J L, MICHAEL W M. Empirical comparison of algorithms for network community detection[A]. Proc of the 19th International Conference on World Wide Web[C]. North Carolina, 2010. 26-30.

作者简介：

牛建伟（1969-），男，河南郑州人，博士，北京航空航天大学教授，主要研究方向为嵌入式与移动计算。



戴彬[通信作者]（1987-），男，天津人，硕士，北京航空航天大学教师，主要研究方向为机会网络和社会网络计算。E-mail:daibin_buaa@hotmail.com。

童超（1978-），男，四川成都人，博士，北京航空航天大学讲师，主要研究方向为移动与社会网络分析。

霍冠英（1990-），男，重庆人，北京航空航天大学硕士生，主要研究方向为数据分析。

彭井（1988-），男，重庆人，北京航空航天大学硕士生，主要研究方向为社会网络分析。