第36卷 第14期 计 算 机 工 程 2010年7月
Vol.36 No.14 Computer Engineering July 2010
·开发研究与设计技术· 文章编号：1000—3428(2010)14—0227—03 文献标识码：A 中图分类号：TP393

一种新的评价社区结构的模块度研究

王 林[1]，戴冠中[2]，赵焕成[1]
(1. 西安理工大学自动化学院，西安 710048；2. 西北工业大学自动化学院，西安 710072)

摘 要：指出 Newman 和 Girvan 提出的模块度概念(Physical Review E, 2004, E69)不适用于社区大小差异较大的情形。为克服这一缺陷，提出与社区大小无关的连接密度和内聚系数概念，在此基础上，构造一种新的模块度。理论和实践证明，该模块度适用于社区大小相似以及社区大小差异较大的情形。

关键词：模块度；社区结构；连接密度；社区的内聚系数

Research on Modularity for Evaluating Community Structure

WANG Lin[1], DAI Guan-zhong[2], ZHAO Huan-cheng[1]
(1. School of Automation, Xi'an University of Technology, Xi'an 710048;
2. School of Automation, Northwestern Polytechnical University, Xi'an 710072)

【Abstract】 The approach of modularity, proposed by Newman and Girvan in order to measure the satisfaction with network decomposition, is found not suitable to the evaluation of community structure in networks when the number of links incident to each community differs too much. To resolve this problem, this paper presents a new modularity approach based on the concepts of linking density and cohesion of communities. Through both theoretical and empirical studies, the modularity is applicable to all cases including the networks with communities having nearly the same number of incident links and the networks when the number of links incident to each community differs too much.

【Key words】 modularity; community structure; linking density; cohesion of community

1 概述

社区结构是复杂网络继小世界和无标度特征之后发现的又一重要特征。研究发现，许多实际网络都呈现出这一特征，即整个网络可以被划分成若干个群。在群内部，节点之间的连接数比较紧密，而各个群之间的连接则相对稀疏，这些群被称作社区。在大型复杂网络中自动搜寻和发现社区具有重要的实用价值，比如在理解和可视化网络的结构方面，它能提供许多帮助[1-2]。

近年来，关于复杂网络社区结构的研究取得了一定成果，提出了许多社区结构的发现算法。比如传统的分级聚类方法，它是基于各个节点之间连接的相似性或强度，把网络自然地划分为各个子群。首先，计算每个节点对 i, j 的一个权值 W_{ij}，它表示节点对之间连接的紧密程度，然后从一个包括所有节点，但没有边的空网出发，将权值按由大到小的顺序排列，不断地在该网络的节点对之间加边。按照这种方法，节点被聚集成越来越大的社区。整个流程可以用树状图来表示，如图 1 所示。

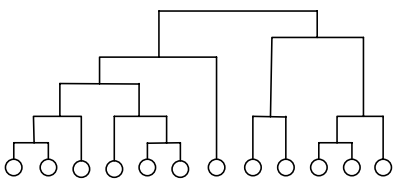


图 1 树状图

底部的各个圆代表网络中的各个节点。当水平虚线从树的底端逐步上移，各个节点也逐步聚合成为更大的社区。当虚线移至顶部，则整个网络就划分成一个社区。在该树状图的任何一个位置用虚线断开，就对应着一种社区结构。这类方法被称作凝聚方法[1,3]。

另一类算法被称作分裂方法。树状图的构造顺序与凝聚方法恰好相反。该方法是从所关注的网络出发，不断移除边，这样就逐步把整个网络划分成越来越小的彼此不相连的子图(社区)。在分裂方法中，一个最重要的步骤就是选取被移除的边，而该边则一定是连接不同社区之间的边[4-5]。

无论是凝聚算法还是分裂算法，都存在一个缺陷：即对于网络的社区结构没有一个量的定义。因此，不能直接从网络的拓扑结构判断它所求得的社区是否是实际网络中的社区结构，从而需要一些附加的关于网络意义的信息来判断所得到的社区结构是否具有实际意义。为了解决这个问题，Newman 和 Girvan 引进一个衡量网络划分质量的标准——模块度[1]。它是指网络中连接社区结构内部顶点的边所占的比例与另外一个随机网络中连接社区结构内部顶点的边所占比例的期望值相减得到的差值。这个随机网络的构造方法为：保持每个顶点的社区属性不变，根据顶点的度随机连接顶点间的边。如果社区结构划分得好，则社区内部连接的稠密程度应高于随机连接网络的期望水平。模块度定量的表示了社区结构的强弱，被许多科研工作者所接纳[3-5]。

因此，必须计算模块度来最终决定树状图的哪一个分解

基金项目：国家“863”计划基金资助项目(2005AA147030)；陕西省自然科学基础研究计划基金资助项目(2007F14)
作者简介：王 林(1963－)，男，教授、博士，主研方向：复杂系统及应用；戴冠中，教授、博士生导师；赵焕成，硕士研究生
收稿日期：2010-02-18 **E-mail：**wanglin@xaut.edu.cn

所对应的社区结构是合理的。但是，当社区大小差异较大时，Newman 和 Girvan 所提出的模块度(以下简称 NG 模块度)并不能得到所期望的值[6-8]。为了克服这种缺陷，基于社区的连接密度和社区的内聚系数，本文提出一种新的模块度。理论和实践研究证明，在度量网络中的社区结构时，这种新的模块度比 NG 模块度能产生更好的结果。

2 NG 模块度

假定网络 N 已经被划分成 n 个小社区 $S_1,S_2,\cdots,S_n$。定义一个 $n\times n$ 的对称矩阵 $\boldsymbol{e}=(e_{ij})$，其中，元素 e_{ij} 表示原网络中连接社区 i 和社区 j 中节点的边数所占的比例。当 $i\neq j$ 时，e_{ij} 为社区 S_i 和 S_j 之间连接边数的一半，这样矩阵 $\boldsymbol{e}$ 中所有元素的和是 1。这个矩阵的迹 $Tre=\sum_{i=1}^{n}e_{ii}$ 给出了网络中连接同一个社区中节点的边所占的比例。定义每行(或者列)中各元素之和为 $a_i=\sum_j e_{ij}$。它表示与第 i 个社区中的节点相连的边所占的比例。Newman 和 Girvan 指出，在一个网络中，如果不考虑节点属于哪个社区而在节点对之间增加边，则有 $e_{ij}=a_ia_j$。从而，模块度可以表达为

$$Q=\sum_{i=1}^{n}(e_{ii}-a_i^2)=Tre-\left\|\boldsymbol{e}^2\right\| \tag{1}$$

其中，$\left\|\boldsymbol{e}^2\right\|$ 表示矩阵 $\boldsymbol{e}^2$ 中所有元素之和。上式的物理意义是：网络中连接 2 个同种类型的节点的边(即社区内部边)的比例，减去在同样的社区结构下任意连接这 2 个节点的边的比例的期望值。如果社区内部边的比例不大于任意连接时的期望值，则有 $Q=0$。Q 的上限为 $Q=1$，Q 越接近这个值，就说明社区结构越明显。

进一步分析 NG 模块度的定义，记社区 S_i 中节点的度值之和为 D_i，定义

$$A(S_i,S_j)=\sum_{l\in S_i,k\in S_j}a_{lk} \tag{2}$$

这样有

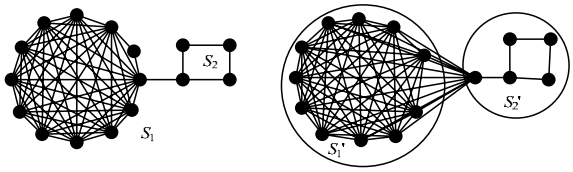
$$Tre=\frac{\sum_{i=1}^{n}A(S_i,S_i)}{D},\ a_i=\frac{D_i}{D} \tag{3}$$

这里 D 是指网络中所有节点的度值之和。NG 模块度可以表示为

$$Q=\frac{\sum_{i=1}^{n}A(S_i,S_i)}{D}-\sum_{i=1}^{n}\left(\frac{D_i}{D}\right)^2=1-\frac{\sum_{i\neq j}A(S_i,S_j)}{D}-\sum_{i=1}^{n}p_i^2 \tag{4}$$

其中，$p_i=\frac{D_i}{D}$，并且有 $\sum_{i=1}^{n}p_i=1$；项 $\frac{\sum_{i\neq j}A(S_i,S_j)}{D}$ 表示社区间的边所占的比例。可以看出，当 p_i 近似相等时，即社区是均质的时候，项 $\sum_{i=1}^{n}p_i^2$ 逼近于它的最小值。要获得最大的 NG 模块度，必须尽量减少社区间边的数目。因此，在每个社区的节点度值之和近似相等的情况下，NG 模块度可以得到合理的结果。

然而，当 p_i 彼此差异较大时，NG 模块度则达不到预期的效果。例如，倘若有一个 p_i 趋近于 1，不妨设为 p_1，无论社区间的边的数目是多少，NG 模块度都将趋近于 0。因此，在这种情况下，NG 模块度是得不到满意解的。下面通过一个简单的例子来加以说明。考虑由 2 个社区 S_1 和 S_2 构成的网络，其中，S_1 是由 100 个节点构成的完全子图，S_2 是由 4 个节点构成的环，并且 S_1 和 S_2 仅有一边相连接，如图 2(a)所示。很明显网络被自然地分割成 S_1 和 S_2 2 个社区，但是对于这种分割，模块度值才仅仅是 0.001 6。这里存在另一种不太合理的分割方法，如图 2(b)所示，它的 NG 模块度是 0.001 8(模块度值大于第 1 种情况)。



(a)分解 1　　(b)分解 2

图 2　NG 模块度的局限性示例

通过理论分析和举例说明，可以看出，NG 模块度不利于度量社区大小差异较大的情形。然而，绝大多数实际网络的社区大小差异较大，因而使用 NG 模块度可能达不到预期的效果。

3 LC 模块度

在 Newman 和 Girvan 看来，网络的社区结构就是把网络划分成各个小的社区，使得社区内部的连接非常紧密，而社区间的连接则比较稀疏。假定 $S_1,S_2,\cdots,S_n$ 是网络 N 的一个分解。网络分解是否合理的问题就等价于以下 2 个问题：(1)社区内部的连接是否紧密；(2)社区内部的连接数是否大于社区间的连接数。本文定义 2 个概念来描述以上 2 个问题。

第 1 个概念是社区 S_i 的连接密度 $L(S_i)$，定义如下：

$$L(S_i)=\begin{cases}\dfrac{2E(S_i)}{n_i(n_i-1)} & n_i>1\\ 1 & n_i=1\end{cases} \tag{5}$$

其中，n_i 表示社区 S_i 的节点数；$E(S_i)$ 表示社区 S_i 内部的边数。很明显连接密度 $L(S_i)$ 描述了社区 S_i 内部连接的紧密程度。

第 2 个概念是社区 S_i 的内聚系数 $Coh(S_i)$，定义如下：

$$Coh(S_i)=\begin{cases}\dfrac{E(S_i)}{E(S_i)+\sum_{j\neq i}A(S_i,S_j)} & E(S_i)>0\\ 0 & E(S_i)=0\end{cases} \tag{6}$$

其中，$i\neq j$，并且 $A(S_i,S_j)$ 表示连接社区 S_i 和 S_j 之间的边数。当 $Coh(S_i)>1/2$ 时，$E(S_i)>\sum_{j\neq i}A(S_i,S_j)$，就表示社区 S_i 内部的连接边数大于 S_i 和其他社区之间的连接数。因此，$Coh(S_i)$ 描述了社区 S_i 的独立性。

根据社区的连接密度和内聚系数的概念，$L(S_i)Coh(S_i)$ 既能描述社区 S_i 的内部连接紧密程度，又能描述社区 S_i 与网络中其他社区的独立性。于是，新的模块度定义为

$$M(S_1,S_2,\cdots,S_n)=\begin{cases}\dfrac{\sum_{i=1}^{n}L(S_i)Coh(S_i)}{n} & n>1\\ 0 & n=1\end{cases} \tag{7}$$

其中，$S_1,S_2,\cdots,S_n$ 是网络的一个分解。

为了区别于 NG 模块度，本文称这种新的模块度为 LC 模块度(采用式子 $L(S_i)Coh(S_i)$ 的首字母表示)。从式(5)~式(7)可以看出，LC 模块度与社区的连接密度和内聚系数相关，与社区的内部节点度值之和无关。本文通过图 2 来举例说明 LC 模块度的优点。不难发现，图 2(a)和图 2(b)的 LC 模块度分别为 0.766 6 和 0.502 0。图 2(b)的 LC 模块度略低于图 2(a)的模块度，也即第 1 种分解更为合理。

第 2 个例子是出自社会网络分析的一个经典问题——空手道俱乐部网络。20 世纪 70 年代初期，Wayne Zachary 用了 2 年的时间来观察美国一所大学中的空手道俱乐部成员间的相互社会关系。基于这些成员在俱乐部内部及外部的社会关系，他构造了他们之间的关系网。由于该俱乐部主管和校长关于是否抬高俱乐部收费问题而发生分歧，致使该网络最终被分裂成了 2 个分别以主管和校长为核心的小俱乐部，如图 3 所示，节点 1 和节点 33 分别代表了俱乐部主管和校长，而方形和圆形的节点分别代表了分裂后的小俱乐部中的各个成员。

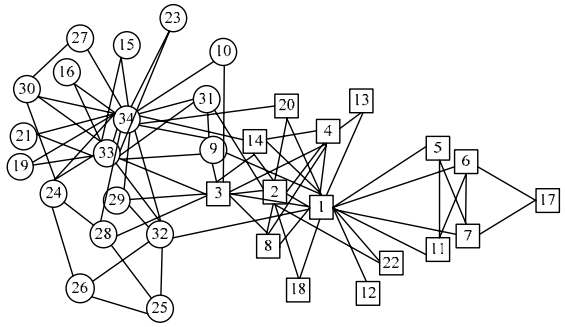


图 3　空手道俱乐部网络

本文参照文献[3]的方法，即结合谱分析和凝聚算法分析网络的社区结构：

(1)求该网络拉普拉斯矩阵的所有特征值(由小到大排序)及其对应的特征向量。

(2)对于任一正整数 J，选取拉普拉斯矩阵的前 J 个非零特征向量，然后将网络中的节点投影到由这几个特征向量所组成的多维空间上。

(3)根据投影空间中投影点之间的距离远近，分离得到不同的区域。

(4)上一步所得到的不同投影区域所对应的网络节点集即是所发现的社区。

在步骤(3)中，2 个投影节点之间的距离可以是欧几里得距离，也可以是 2 个投影节点向量(将原点连接到投影节点即构成投影节点向量)之间的夹角。一般认为基于向量夹角的距离的分离效果较为理想[2]。

得到网络节点在特征向量空间中的投影后，对于小型网络，可用观察法获得社区分解。但对于大型网络，则必须借助于系统步骤利用计算机实现，本文采用层次聚类法对投影空间中的节点进行聚类，根据节点之间的距离(由小到大)将节点聚合：先连接距离最近的 2 个节点，下一步将上一步连接起来的 2 个节点看成一个节点，而其他节点到该新节点的距离可以定义为到组成新节点的 2 个节点的距离中较大的距离，然后再重复上一步，直到最终将所有的节点聚合成一个节点。上述过程如果进行到某一步停止，则得到网络的一个分解。当本文考虑了所有可能的正整数 J 时，就得到了网络的许多分解，这样就产生了一个问题：哪一个分解是最合理的。这就是模块度所起的真正作用。

对于 NG 模块度，当空手道俱乐部网络中的节点投影到由前 4 个非零特征向量所组成的四维空间时，得到最大模块度值(见图 4)。该最大值所对应的分解由 5 个社区组成。它们是：{5, 6, 7, 11, 17}, {9, 10, 15, 16, 19, 21, 23, 27, 30, 31, 33, 34}, {1, 2, 3, 4, 8, 13, 14, 18, 20, 22}, {12}和{24, 25, 26, 28, 29, 32}。而对于 LC 模块度，当该网络中的节点投影到由前 2 个非零特征向量所组成的二维空间时，得到最大模块度值(见图 4)。LC 最大的模块度值所对应的分解由 3 个社区组成。它们是：{5, 6, 7, 11, 17}, {9, 10, 15, 16, 19, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34}和{1, 2, 3, 4, 8, 12, 13, 14, 18, 20, 22}。很明显 LC 模块度值所对应的分解更接近实际情况。

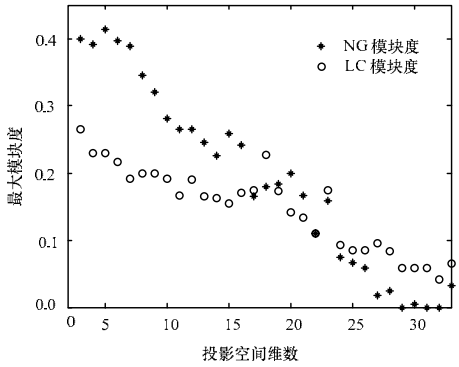


图 4　空手道俱乐部网络的 NG 模块度和 LC 模块度

本文引用 Newman 和 Girvan 给出的计算机生成的仿真网络[1]。该网络有 128 个节点，分成 4 个社区，每个社区的节点数均为 32 个。对每个任意的节点对，分别放置一条边，这 2 个节点位于 2 个不同的社区内的概率为 p_{out}，而位于同一个社区内的概率是 $1-p_{out}$。构建网络使得每个节点的度的期望值都为 16。本文设 p_{out} 以 0.05 为步长，由 0.05 变到 0.40。NG 模块度的最大值和 LC 模块度的最大值所对应的分解恰好完全一样，如图 5 所示。这个结果显示，复杂网络中社区大小分布趋向均匀时，LC 模块度和 NG 模块度一样有效。

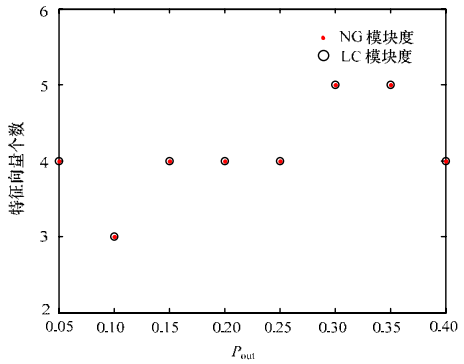


图 5　LC 模块度的性能测试

4　结束语

本文通过理论分析和具体例子说明了 Newman 和 Girvan 提出的模块度不利于度量社区大小差异较大的情形，其原因在于 NG 模块度依赖于社区的大小。据此，本文做了如下工作：

(1)提出连接密度的概念，连接密度刻画了社区内部节点连接的紧密程度，此概念不依赖于社区的大小；

(2)提出内聚系数概念，刻画了社区内部节点与社区外部节点之间的独立程度，此概念也不依赖于社区的大小；

(3)将连接密度与内聚系数相结合，提出新的模块度，旨在既考虑到社区的独立性，又考虑到社区内部的紧密性；

(4)借助几个经典的实际网络与人造网络对新的模块度进行了测试，发现新模块度不仅适用于社区大小相似的情形，而且也适用于社区大小差异较大的情形。

下一步的研究方向是如何借助连接密度与内聚系数构造出更自然的模块度。

(下转第 232 页)