基于信息扩散的多尺度重叠社团快速探测算法

李慧嘉
(中央财经大学管理科学与工程学院 北京 100080)

摘 要 现有的社团分析方法由于需要网络的全局信息,并且只能在单一的尺度上划分社团,因此不利于分析大规模的科技社会网络。提出了一种新颖的多尺度社团结构快速探测算法,其只利用网络的局域信息就可以模拟复杂网络中的多尺度的社团结构。该方法通过优化表示网络统计显著性的拓扑熵,来寻找有最佳统计意义的社团结构。为了得到具体的社团归属,算法只需利用局部信息的扩散来更新归属向量便能够收敛到局部极小值,因此具有较低的计算复杂性。它不需要指定具体的社团数量,便能够找到每个节点与具体社团的归属关系,从而能够自然地支持模糊社团的划分。理论分析和实验验证共同表明,该算法可以快速而准确地发现社会网络和生物网络中的各种功能社团。

关键词 社团结构,复杂网络,重叠性,信息扩散,多尺度
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.09.024

Fast Algorithm for Detecting Multi-scale Overlapping Community Structure Based on Information Spreading

LI Hui-jia
(School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100080, China)

Abstract Most existing community detection methods require the complete graph information, thus is impractical for large-scale technological and social networks. This paper proposed a novel algorithm for the fast detection of multi-scale overlapping community structure. It does not embrace the universal approach but instead tries to focus on local ties and model multi-scale interactions in these networks. It identifies leaders and modules around these leaders using local information. It naturally supports overlapping information by associating each node with a membership vector that describes its involvement of each community. Our method for the first time optimizes the topological entropy of a network and uncovers communities through a novel dynamic system converging to a local minimum by simply updating the membership vector with very low computational complexity. Both theoretical analysis and experiment show that the algorithm can detect communities in social and biological networks fast and accurately.

Keywords Community structure, Complex network, Overlapping, Information spreading, Multi-scale

1 引言

由于现实世界中的诸多系统以网络形式存在,并具有很高的复杂性,如科技网络的WWW网络和Internet网络、社会系统中的科学家合作网和Email关系网、生物网络中的蛋白质交互网和新陈代谢网等,因此复杂网络的研究具有重要的理论价值和实际意义。复杂网络[1-7]的结构与网络的功能有着紧密的关系(如鲁棒性、传递性等),因此发现并确定网络的正确拓扑结构及性质(如小世界性质和无标度性质)是一个非常重要的课题。在复杂网络的各种性质中,针对社会网络和生物网络的社团结构探测在最近几年已经成为一个非常受人关注的问题[14,15]。属于一个紧密相连的社团中的节点,更有可能拥有共同的属性或动态过程。以WWW网络为例,成组的网页链接更有可能具有共同的相关主题。这些集合的网页可能对应于某些种类的社团。因此,搜索引擎可能会为了增加精度而专注于这一具体部分的搜索结果,从而又减少了搜索耗费。

事实上,一些具体的目标函数已经被用来衡量社团划分[2-7]。其中,模块度Q是其中应用最为广泛的一个[2-4],而且已经被很多工作验证和应用。然而,其中包括模块度优化方法在内的大多数方法需要了解整个网络的结构,在确定全局信息的基础上进行社团的划分。这种信息需求上的约束对于大型复杂的网络是不切实际的,因为要了解整个网络完全是一个挑战。此外,统计方法只能检测到最显著的社团结构,而经常忽略它们可能存在的多尺度拓扑结构[5-7]。这些方法不具有提供粗粒化的网络认知,从而不能勾勒其组织或不确定的节点集合可能具有的共同的隐藏功能或性能。

针对这些局限性,本文提出了一种新颖的多尺度社团结构快速探测算法,其主要应用在社会网络和生物网络中。不同于传统算法,该算法只利用网络的局域信息来模拟复杂网络中的多尺度的社团结构模式。我们的方法通过优化表示网络统计显著性的拓扑熵,来寻找有最佳统计意义的社团结构。我们注意到网络的拓扑熵函数是非凸的,因此期望通过传统的标准优化算法找到全局最小值是不切实际的。因此我们开

到稿日期:2013-11-13 返修日期:2014-01-17 本文受国家自然科学基金项目(91324203,11131009),中财121人才工程青年博士发展基金(QBJ1410)资助。

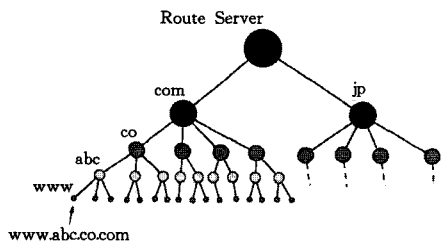
李慧嘉(1985-),男,博士,讲师,主要研究方向为社会网络、数据挖掘、运筹学,E-mail:Hjli@amss.ac.cn。

发了一个新的动态系统，该系统只需利用局部信息的扩散来更新归属向量便能够收敛到局部极小值，从而具有较低的计算的复杂性。它不需要指定具体的社团数量，便能够找到每个节点与具体社团的归属关系，因此能够自然地支持模糊社团的划分。理论分析和实验验证共同表明，该算法可以快速而准确地发现社会网络和生物网络中的各种功能社团。

2 网络层次结构和领袖节点

给定一个网络 $G=(V,E)$，包含 n 个节点，我们可以将这些节点划分为 a 个社团。在每一个社团中我们假设可以选择一个“领袖”节点。领袖节点应该有两个属性：它应该与社团内其他节点紧密相连，在必要的时候它能够与其他领袖节点相互通讯。如果在每个组执行一个分布式算法，使得领导节点之间能在一个更高层次上面进行沟通，节点就可以享受更快的收敛速度。

人们可以自然地将社会网络与层次网络结构关联起来。在一个这样的层次结构中，领袖节点比其他一些节点更加重要，因此位于一个更高的层次水平上。以 WWW 中的 DNS 网络为例，当查找 IP 地址时路由服务器是一个天生的领袖节点并且位于最高层次上（见图 1）。最有影响力的节点，路由服务器，位于层次结构树中的最高点。该服务器包括“www”，“abc”，“co”和“com”，“com”位于较低的水平。为了获得 IP 地址，用户需要从最高水平的路由服务器到最低层次的 www 服务器进行查询。因为层次结构是节点管理传播的结果，我们相信识别这样一个层次网络结构将自然产生一个社团结构。领袖节点影响最大的区域可以自然地定义为该领袖节点的社团。因此，社团可以通过发现所有的领袖节点和其统辖的节点来确定。这种方式获得的社团可以很自然地解释清楚。同时，另一个直观的性质是同一社团的节点的最短路径应该最短。给定一个网络，节点只知道周围局部的相关信息[8,9]，其中包括它们的邻居节点信息。任何一个节点若希望提高自己的表现，就需要知道更多关于网络的信息。这个信息可以通过改进节点的邻居顺序选择提高其性能，但这将花费大量的计算复杂性。最完整的全局信息为邻接矩阵。由于每个节点只有有限的存储空间、能量和计算能力，因此很难直接用邻接矩阵进行数据处理。我们的目标是设计一个方案，为每个节点提供一个小的向量使其包括全局网络的压缩信息，即节点相对于其他节点的位置。理想情况下这可以通过分布式的方式实现[12,13]。



节点的大小描述了层次结构中的不同层次，
更大的节点定位在较高的水平上

图 1 具有 IP 为“www. abc. co. com”的 DNS 网络的层次结构

此外，一个用来揭示社会的模块网络的高效的方法可采用多尺度的方式。这个多尺度的社团结构能够提供具有粗粒化优势的系统表示，从而能够描述组织结构并确定一些节点可能有隐藏的共同属性。大多数方法找到的节点只满足硬划分的条件，即每个节点分配到一个且只有一个社团中，而重叠分区是不兼容。所以，为了找到高质量的功能社团结构，我们应该寻找软划分方式[9,10]的算法。

3 社团划分算法

对于一个具有 n 个节点的网络 $G=(V,E)$，我们开发了一种新颖的分布式的算法，即可以只使用局部信息将节点分为“领袖”或者“一般”节点。此外，算法以多尺度的方式给每个一般节点赋予一个归属向量，表明领袖节点的多尺度影响。这在一定程度上可以解释网络内部的隐藏信息。迭代包括 3 个步骤，描述如下。

3.1 计算节点领导力

首先，我们计算网络中每个节点 i 的领导力 $f(i)$。领导力 $f(i)$代表节点 i 的意见在整个网络中的重要性。节点领导力 $f(i)$的方程定义为：

$$f(i)=\sum_{j=1,d_{ij}\leqslant\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor}^{n} e^{-\frac{d_{ij}}{\delta}} \tag{1}$$

其中，d_{ij} 表示顶点 i 和 j 之间的的最短距离。$\delta\in(0,+\infty)$代表影响因素，用来控制节点间的相互作用范围。根据指数函数 $e^{-\frac{d_{ij}}{\delta}}$ 的特性，每个节点到其他节点的影响范围近似为$\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor$。

当 d_{ij} 大于$\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor$时，指数方程的值迅速降低到 0，所以我们可以用 δ 控制一个节点的影响范围，而计算 $f(i)$只需要在范围 $d_{ij}\leqslant\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor$即可。对一个网络边的密集区域，其中的点往往具有较高的领导力。最大领导力的节点意味着它和其他节点有最多数量的关联，可以被看作是领袖节点的候选。因此，我们可以利用节点的领导力来表示网络中的节点的重要性。

3.2 识别领袖节点

识别社团的领袖节点对于分析复杂网络的性能是非常重要的。很多方法可以定义“关键节点”，如最大度节点或中心性最大的节点[11]。在此使用节点的领导力来搜索领袖节点。根据社团结构的定义，社团内部边的连接密度大于社团之间的边。每个社团都代表一个边关联相对较高的区域，社团内边连接相对紧密，社团的领袖节点则具有最高的领导力。此外，不同的社团被局域领导力最低的节点分开，领导力最低的节点也就是边界节点。

我们注意到，在有些极端情况下，两个或两个以上的领袖节点是邻接关系，那么将这些领袖节点分组在一起，成为同一个社团的领袖节点。例如，在一个全连接网络，所有节点都是同一个社团的领袖节点，而在环形网络中，每个节点是一个独立社团的领袖节点。具体地说，如果两个最高领导力节点的最短距离小于$\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor$，就把它们分在同一个社团。寻找领袖节点只需要一个简单的广度优先搜索，如果找到了就选择一个随机的节点重复该过程直到收敛。该步骤的计算复杂度为 $O(m)$，其中 m 是网络中边的数目。

3.3 利用随机游走动态确定归属向量

在这一步，我们的目标是制定一个机制，为每个节点分配一个小的向量，该向量包括压缩后的全局信息，使该节点能够知道与其他节点之间的相对位置。基于网络中的随机游走动态系统[16,17]，我们定义了一个节点的归属向量。假定一个网络中有 a 个领袖节点 $l_1,l_2,\cdots,l_a$ 和 $n-a$ 个常规节点。我们将任意顺序分配给这些领袖节点，然后为每个常规节点赋予一个归属向量，确定每个常规节点和社团的归属关系。节点 i

的归属向量表示为 $x_i=(x_i^1,x_i^2,\cdots,x_i^a)\in R^a$，其中 $x_i^k(t)$ 表示在 t 时刻节点 i 相对于社团 k 的归属关系。

算法流程的操作如下：首先将领袖节点的归属向量赋为单位向量，这 a 个单位向量并不改变。对于常规节点 i，x_i^k 初始在[0,1]($k=1,2,\cdots,a$)之间均匀随机赋值。然后，我们将 x_i 的每行进行标准化，使得对于所有的领袖节点 k，x_i^k 的和为 1。在每次迭代 t 时，归属向量按照以下规则进行迭代：

$$x_i^k(t+1)=\frac{1}{\sum_j a_{ij}+1}\left[x_i^k(t)+\sum_j a_{ij}x_j^k(t)\right] \tag{2}$$

其中，$A=\{a_{ij}\}$ 是网络的邻接矩阵，$a_{ij}=1$ 表示节点 i 和 j 相互连接，否则 $a_{ij}=0$。我们注意到在所有的时间 t，$\sum_k x_i^k(t)=1$。方程(2)等价于 $X(t+1)=PX(t)=(I+D)^{-1}(A+D)X(t)$，其中 $P=(I+D)^{-1}(A+D)$ 是随机游走矩阵。其实，领袖节点 l_k($k=1,2,\cdots,a$)对常规节点 i 的影响力 x_i^k 等价于随机游走者从 i 开始到达 l_k 而不是其他领袖节点的概率[19]。如果网络是连通图，则迭代 $\lim_{t\to\infty}x_i(t)$ 将收敛到一组唯一的向量，这些向量可以自然地表示成节点属于一个特定社团的概率。所以，虽然领袖节点只知道局部信息，但我们可以使用随机游走动态地找到归属信息，使其包含压缩后的全局信息。

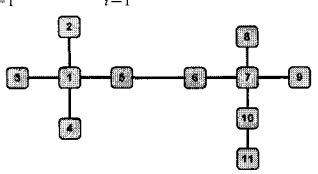
4 算法相关部分的一些描述

在本节中，我们给出算法的几个重要特性的描述，包括计算影响因素，识别多尺度的社团结构，利用局域信息确定领袖节点，确定重叠节点和估计该算法的复杂性。

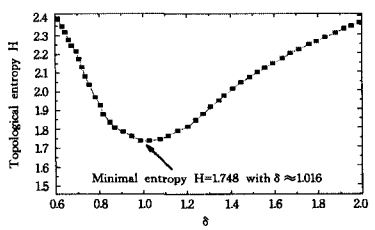
4.1 确定不同规模的影响因子

根据节点领导力的定义，算法仅由一个参数即影响因素控制。我们可以自然地使用它来控制社团的规模。在这里，我们引入拓扑熵 H 的概念，它通过选择合适的熵值来表示一个网络的统计显著性：对网络 $G=(V,E)$，$V=v_1,v_2,\cdots,v_n$，V 的领导力为 $f(1),f(2),\cdots,f(n)$，那么拓扑熵定义为：

$$H=-\sum_{i=1}^{n}\frac{f(i)}{\sum_{i=1}^{n}f(i)}\log\left[\frac{f(i)}{\sum_{i=1}^{n}f(i)}\right] \tag{3}$$



(a) 一个 11 个节点的小网络



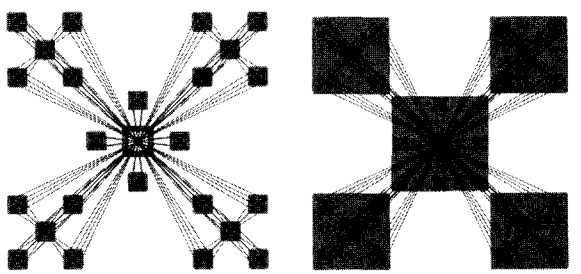
(b) 拓扑熵 H 随着影响因子 δ 的变化趋势

图 2

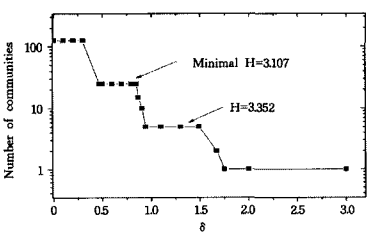
较小的 H 值意味着一个稳定合适的社团划分。下面是一个简单的例子，我们用一个包含 11 个节点的网络，如图 2(a)所示，来计算不同的对应的拓扑熵。如图 2(b)所示，当 δ 从 0 增加时，相应的拓扑熵开始减小，并在 $\delta=1.01$ 时达最小值1.74。当 δ 离开最优值后，熵开始增加并最终达到最大值。

因此，寻找最佳的，就相当于最小化一个单参数的非线性函数 $H(\delta)$，我们可以用许多算法来求解，例如随机搜索算法或者模拟退火算法。然而，小的 H 对应的社团数目即使不是最小，也仍然是有意义的。特别地，根据领袖节点的性质，节点的影响范围大约是 $\left\lfloor\frac{3\delta}{\sqrt{2}}\right\rfloor$。当 $0<\delta<\sqrt{2}/3$ 时，两个节点之间是没有关联的。由于不存在关联，每个节点都单独属于一个社团，此时社团的数目为 n。同样，当 $\sqrt{2}/3<\delta<2\sqrt{2}/3$ 时，节点只与它的邻居进行交互。随着 δ 值的增长，节点可以影响范围越来越广的节点，因此，领袖节点和相应的社团的数目会逐步减小。最后，当 $\delta\geqslant\sqrt{D}/3$ 时(其中 D 是网络的直径)，每一个节点对可以相互影响网络中的任意一个节点。

为了说明算法可以随着 δ 的变化发现多尺度的社团结构，我们在一个经典的分层无标度网络即 RB125 网络[18]中，进行了测试。RB125 是由 Ravasz 和 Barabasi 提出的。对应于图 3(a)和(b)，该网络最小的拓扑熵为 $H=3.107$，其次是 $H=3.352$，这显示了两种不同的尺度网络结构，值得我们研究和讨论。图 3(c)展示了 H 随着不同社团数目的变化情况。我们观察到持续最长的网络社团结构分别为 25 和 5 个社团，这两种情况显示了两个不同的层次结构。其中，25 个社团的分区和 5 个社团的分区在原图中高亮显示。



(a) 具有 25 和团的划分，其中拓扑熵为 $H=3.107$　(b) 具有 5 个社团的划分，其中拓扑熵为 $H=3.352$



(c) 社团数目随着 δ 的变化情况

图 3　RB125 对应的层次无标度网络

4.2 利用局部信息确定社团归属

在算法中，我们可以只通过局部信息来识别社团的领袖节点。通过探测社团的领袖节点，可以获得最有影响力的网络子结构，这是非常有用的拓扑信息。如果删除领导者，人们可以预期的网络将遭受严重的破坏，然后会分裂成几个较小的不相连的子部分。领袖节点所处的层次或领袖节点领导的社团、在此区域领袖节点的意见具有最大的领导力，比如对病毒传播的影响。因此，该算法可以自动确定领袖节点的数目，即网络中社团的数目。算法的一个有趣的特征是虽然它会自动检测到最佳的领袖节点数目，但我们仍可以指定并输入一个特定的社团数目，用以划分具体的社团结构。

4.3 确定重叠节点

需要指出的是，绝大多数社团检测方法都假设网络中的社团是不相交的，即每个节点只属于一个非重叠的社团。通常称这些方法为“硬划分”算法。然而，许多实际网络中的社

团往往具有一定程度的重叠部分[9,10]。我们的算法的一个重要特性是计算每个节点对所有社团的归属度向量，而不是局限于单一的社团。我们可以很容易地识别同时属于多个社团的重叠节点，因此我们的方法是一个“软划分”算法。此外，可以有效地发现社团内的领袖节点和它的隶属节点，这对于分析网络的功能具有非常重要的作用。

4.4 算法计算复杂性

该算法的总体复杂性取决于算法中3个部分复杂性的最高的部分。下面依顺序分析它们的复杂性。第一步是计算节点的领导力 $f(i)$。我们需要计算节点间最短路径长度的指数函数 $d_{ij} \leqslant \left\lfloor \frac{3\delta}{\sqrt{2}} \right\rfloor$，此步骤的复杂性至少为 $O(m)$，m 为网络边的数目。实际上，当网络为稠密图时，这一步的计算复杂度最坏为 $O(n^2)$。接下来一步为确定网络社团的领袖节点，这是通过搜索所有具有局域最高领导力的节点来实现的。这一步可以通过一个简单的广度优先搜索来完成，其复杂性为 $O(m)$。最后一步非常类似于一个一致线性过程，其复杂度为 $O(n)$，与随机游走过程相似。因此，最高的计算复杂性出现在第一步，即计算节点的领导力。它的复杂性取决于网络连接的稠密程度，稠密图需要更多的计算复杂性。算法的总体复杂度最好为线性的 $O(m)$，最坏为 $O(n^2)$。

5 实验

本节分别在人工网络(LFR网络)和一些现实的社会网络：空手道俱乐部网络、科学合作网络、大型语义网络和一个大规模蛋白质交互网络上检验我们的算法。结果表明，算法可以有效而且准确地发现复杂网络中的多尺度社团结构。

5.1 人工基准网络

我们通过在人工基准网络上与其他5个著名的算法进行比较来验证算法的有效性。这些算法包括：Newman快速算法[2]、Danon方法[20]、Louvain方法[21]、Infomap方法[22]和clique percolation方法[10]。我们使用由Lancichinetti等人提出的LFR人工基准网络。此基准网络具有无标度的网络度分布和社团规模分布，因此该网络是一个更贴近现实网络、更加精确的测试对象。许多参数被用来控制LFR网络的生成：节点数目 N、节点的平均度 $\langle k \rangle$、最大节点度 $\max_k$、混合比例 μ(每个节点与其他社团的节点共享 μ 比例的边)、最小社团规模 $\min_c$ 和最大社团规模 $\max_c$。μ 的变化范围为 $\lfloor 0,1 \rfloor$，用来确定网络的模糊性。较大的 μ 代表更加模糊的社团结构。在测试中，我们使用默认的参数配置，其中 $N=1000$，$\langle k \rangle = 15$，$\max_k = 50$，$\min_c = 20$，$\max_c = 50$。

为了评估一个社团检测算法的精确性，我们使用归一化后的互信息(NMI)进行测量。测试的重点是能正确地发现网络内在的社团规模。实验结果显示在图4上，其中y轴表示NMI的值，曲线上的每个点的值是50次独立实验的平均值。从结果可以看出，当 $\mu \leqslant 0.3$ 时，所有的算法都有很好的表现，NMI都大于0.85。与其他5个算法相比，我们的算法有相当不错的表现，其精度仅在 $0.35 \leqslant \mu \leqslant 0.5$ 情况下比clique percolation方法略差。然而，Clique percolation算法的时间复杂性大于 $O(n^3)$，几乎相当于广度优先搜索(BFS)算法。由于实际网络可能与人工合成网络在拓扑性质上有一些不同，下面考虑几个广泛使用的真实网络的例子，以进一步评估本算法的表现。

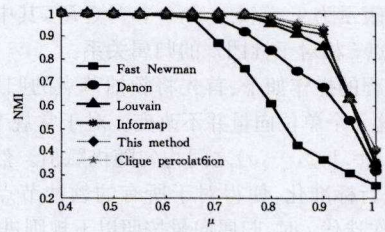
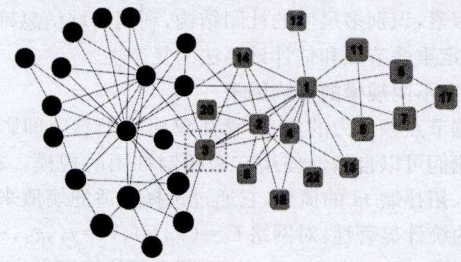


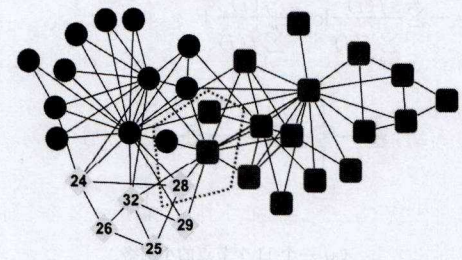
图4 6个算法得到的NMI的对比

5.2 空手道俱乐部网络

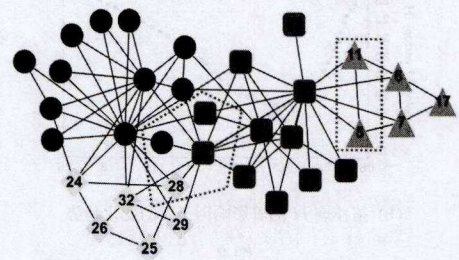
在20世纪70年代初的两年时间里，Wayne Zachary观察并记录了一所美国大学空手道俱乐部成员之间的社会关联情况。他根据俱乐部成员之间的亲密和疏远关系构建了网络[23]。一次偶然的机会，俱乐部的管理员和首席教师之间因为是否提高俱乐部收费产生了纠纷，结果俱乐部最终一分为二，形成两个较小的社团，而社团成员围绕在管理员(节点1)和首席教师(节点33)周围。我们最小化拓扑熵 $H(\delta)$，得到了最优值 $\delta = 1.85$ 和相应的 $H = 3.914$。如图5(a)所示，在最优情况下算法发现的划分不仅与原有的分区完全匹配，而且还准确确定了节点1和33为社团的领袖节点，分别代表了管理者和首席教师。根据归属矩阵，节点3被检测作为一个重叠的节点，因为它在两个社团之间的关联几乎相等。其实，节点3也位于社团之间的边界上，所以能够很清楚地解释。



(a) 当 δ 为最优值1.85时的社团划分



(b) 当 δ 从最优值降到1.41时的社团划分



(c) 当 δ 进一步降低到0.933时的社团划分

图(a)(b)(c)中社团分别用不同的形状表示，重叠节点用虚线圈住

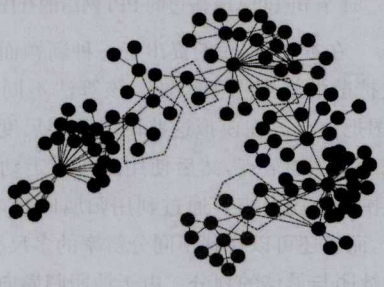
图5 空手道俱乐部网络的划分

相比于最优的情况，当 δ 减少到1.41时，熵 $H = 4.139$ 也是非常小的。社团检测的情况如图5(b)所示，这种情况揭示了另一个规模的空手道俱乐部成员之间的关系。节点28成为另一个领导力最高的节点，而且4个最不稳定的节点，包

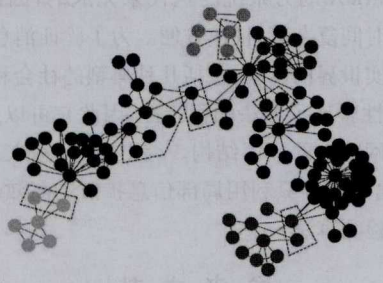
括节点 3,10,20,28 用虚线曲线标记。所以在这种情况下,它们是重叠的节点。此时空手道俱乐部网络中检测到的社团数目为 3。此外,当 δ 减少 0.93 时,H 变成 4.436,我们得到了一个 4 个社团的划分,如图 5(c)所示。这个社团划分与 Newman 在文献[2,3]中检测的一致。我们发现 6 个重叠的节点,其构成了社团之间模糊的界限。因此,使用不同的 δ 能够反映在实际网络中社团的多尺度属性。

5.3 科学家合作网络

Newman 收集的科学家合作网络[3]已在很多文献中进行了研究和检测。该网络包含 118 个节点(科学家或文章作者),以及代表他们在 archive 上合作发表文章的边。图中网的边缘不仅代表文章,我们感兴趣的主要是他们的研究方向是否相同。在最小的熵 $H=5.447$ 时,$\delta=1.493$,我们的算法检测到 8 个社团。图 6(a)显示在最优的情况下检测到的社团结构,此时的情况与文献[3,4]是完全一样的。这很好地验证了我们算法的准确性。然而,我们相信本方法同样可以做一个合理的在视觉上"粗粒化"的划分。所以当 δ 从最优值增大到 1.749 时,相应的熵为 $H=6.483$。节点的影响力范围随着 δ 扩增,社团的数目也在相应地减少。从图 6(b)中,我们注意到一些"不产生影响的"社团,如灰色的社团,分别被更有实力的黑色社团所合并。最后,我们可以得到 6 个社团,用肉眼容易分辨出。这些分辨出的多尺度社团对理解网络功能和结构是非常宝贵的,能帮助我们分析处理大型的网络数据。此外,根据归属向量我们发现了一些重叠的节点,在图 6 中用虚线的曲线圈住。这些节点位于两个或多个社团的边界位置,代表有多个研究方向或跨学科研究背景的科学家。也许这样的节点在复杂网络中起到了桥接两个或两个以上社团的作用。能够找到这样的重叠节点是我们算法的一个显着的特征,有利于分析许多社会和生物网络的自然特征。



(a) 当 δ 等于最优值 1.493 时的科学家合作网络划分



(b) 当 δ 增大到 1.749 时的科学家合作网络划分

在两个子图(a)和(b)中,重叠节点分别被虚线圈住

图 6 科学家合作网的划分

5.4 大规模语义网络

语义网络包含 7207 短语和 31784 条语义之间的关联,边的权重是根据句子中短语共同出现的次数来计算的[10]。为了可视化,我们的算法输出了一个转换后的邻接矩阵,即同一个社团的节点组合在一起(如图 7(a)所示),来表示网络的层次结构。幂率形式的社团规模累积分布在图 7(a)中显示。在最优 $\delta=2.931$ 的情况下,共有 569 个社团被发现,此时最小熵 $H=5.952$。最大的社团规模为 139,最小的规模为 2,平均的社团规模为 12.57。近似地,我们可以认为这是一个幂率行为,也就是大多数社团的规模都比较小,只有少数是大规模社团。其中,我们选择了 4 个比较有趣的社团列举如下:

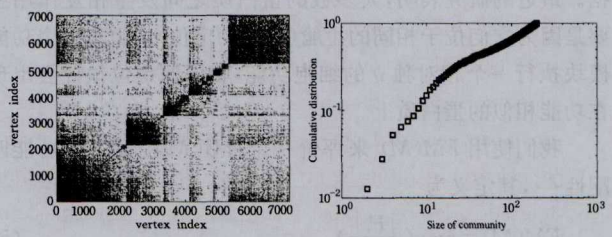
社团 1={Scientist, Inventor, Genius, Gifted, Brilliant, Intelligent, Smart, Science, Intelligence, Musician};

社团 2={Violin, Instrument, Cello, Band, Tuba, Clarinet, Orchestra, Trumpet, Trombone, Oboe, Woodwind, Symphony, Flute, Bass, Viola, Fiddle};

社团 3={Ovation, Sitting, Low, Descent, Up, Step, Ascend, Elevator, Ascent, Staircase, Stairwell, Climb, Steps, Ladder, Stairs, Wake, Stairway, Rise, Escalator, Stair, Down, Standing, Resting, Using};

社团 4={Nails, Hammer, Carpenter, Screw, Screwdriver, Tool, Pliers, Wrench, Sickle, Mechanic, Phillips}。

这 4 个社团中列出的模块都是合理的,而且同一个社团的所有元素都具有相同的语义。在这些元素中,{Musician, Intelligence}被发现是社团 1 和 2 之间的重叠节点,并且{Using, Tool, Mechanic} 是社团 3 和 4 之间的重叠节点。我们可以很容易地认识到,这些重叠的词组都具有模糊语义,对研究短语共同出现具有很高价值。



(a)语义网络转换后的邻接矩阵　　(b)社团规模的幂率累计分布图

图 7

由于大型的语义网络内在的社团结构通常是未知的,因此值得本利用指标来定量地评价我们算法的表现。在这里,我们采用 Newman 和 Girvan 提出的著名的模块度 Q[2] 来衡量我们的算法。Q 的定义如下:

$$Q=\sum_{i=1}^{c}[\frac{l_i^{in}}{L}-(\frac{d_i}{2L})^2] \tag{4}$$

其中,c 是社团的数量,L 是网络中边的总数,l_i^{in} 和 $d_i=2l_i^{in}+l_i^{out}$ 分别是社团 i 内的边的数目和顶点度的总和。图 8 展示了模块度 Q 的拓扑熵 H 随着 δ 变化的情况。我们可以看出,变化的趋势是较低的 H 值对应着较大值 Q 值。结果表明,网络合理的社团结构对应于一个确定的 δ 值。所以,我们的算法可以有效地发现合理的真实世界的社团结构。

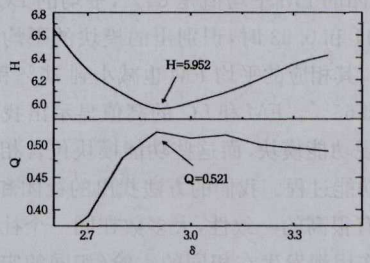


图 8 不同的 δ 的情况下模块度 Q 和拓扑熵 H 的对比

5.5 大规模的酵母蛋白质交互网络

我们将算法应用到一个大型的酵母蛋白质交互网络中进行社团发现，从而找到其功能模块以显示其性能。大型非冗余(没有自连边和重边)的酵母蛋白质相互作用数据是从文献[24]中获得的，它构建了一个高品质的酵母蛋白质交互作用网络，其中包含 3025 个蛋白质(节点)和 6888 条相互作用(边)。这些模块的生物意义和显著性可以通过 MIPS 上已知的功能注释和蛋白质复合物的生物学意义(慕尼黑蛋白质序列信息中心)数据库[25]来进行检验。在这里，我们重点检验所提出的方法在探测大规模生物网络功能模块中的表现。

这里的基本假设是，所识别的社团中含有的蛋白质涉及相同的功能、进化过程或生物单元的功能模块。为了验证这个想法并注释找出的社团，我们将 MIPS 中的酵母功能产品目录(FunCat)的数据库功能注释进行对比。P-值用来给一个特定的蛋白质赋予统计意义，经常被用来作为一个标准以分配给每个社团一个主要的功能。FunCat[26]是一种为各种生物功能的蛋白质的功能进行注释的数据库。主要分支表现为层次、树状结构，共有 1307 个功能类别和 6 个级别的增加特异性。利用专家从文献中收集的一个显著的功能注释表，并结合 MIPS 数据进行比较和分析。此外，蛋白质的定位数据是有价值的信息资源，有助于阐明真核细胞蛋白质的功能。大型定位数据库提供了现有品种的蛋白质组的基本定位数据。最近的研究表明，大多数的蛋白质之间发生相互作用主要是因为它们位于相同的主舱(同样的定位)。由于一个功能模块执行一个相对独立的细胞功能，相同的定位预计会出现在功能相似的蛋白质上。

我们使用 $FM(M)$ 来评价一个社团 M 中元素的功能匹配性[26]，其定义为

$$FM(M)=\max\left(\frac{|F_i|}{|V_M|}\right) \tag{5}$$

其中，$|F_i|$ 表示模块 M 中具有功能 F_i 的蛋白质的数量，$|V_M|$ 表示模块 M 中的蛋白质的数量。此外，我们采用的蛋白质定位数据[24,25]包括 23 个不同的亚细胞位置信息，将其用于验证我们的算法。我们使用 $LC(M)$ 来评估模块 M 中的蛋白定位的覆盖面：

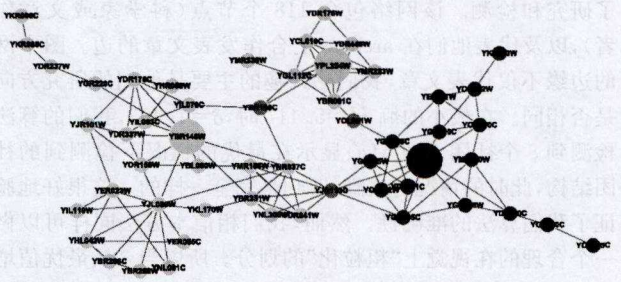
$$Lc(M)=\max\left(\frac{|L_i|}{|V_M|}\right) \tag{6}$$

其中，$|L_i|$ 表示模块 M 中具有定位 L_i 的蛋白质的数量，$|V_M|$ 表示模块 M 中的蛋白质的数量。

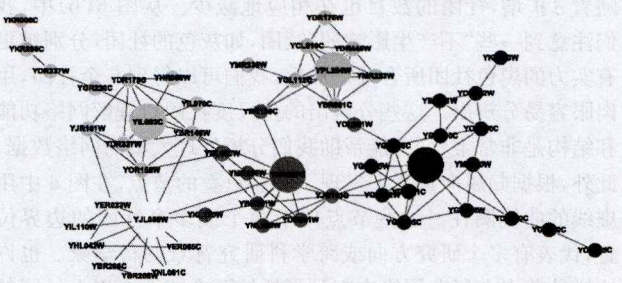
在表 1 中，N 代表检测到的社团数目，Max 表示最大的社团规模，Min 代表最小的社团规模。从表 1 所列的结果可以看到，N 随着 δ 从最优的 1.85 到 0.93 逐渐递减。如表 1 所列，当 δ 等于最优值 1.85 时，检测出的社团平均大小是 22.7，所有社团的 FM 平均值是 84%，平均的 LC 值为 72%。当 δ 等于 1.41 和 0.93 时，识别出的模块的平均大小减少至 15.5 和 13.5，其相应的平均 FM 也减小到 76%和 72%，平均 LC 为 71%和 65%。FM 和 LC 的高值显示出我们的方法可以有效地发现功能模块，而这些功能模块包含相同的生物单位和相同的功能过程。我们的方法找出的社团和 MIPS 中的功能分类具有很高的一致性，大多数在同一个社团的蛋白质之间的相互作用都发生在相同的主舱(相同的定位)。最后，专门展示算法检测出的由 54 个节点和 119 条边组成的一个子集，如图 9 所示。两个尺度的划分可以容易地用肉眼来分辨和判断。这些多尺度的分区将是非常宝贵的，有助于我们了解大尺度结构的生物系统的各种功能。

表 1　3 个不同标度下社团划分结果的对比

δ	H	N	Max	Min	平均社团规模	平均 FM	平均 LC
1.85	3.914	133	126	4	22.7	84%	72%
1.41	4.138	196	118	3	15.5	76%	71%
0.93	4.536	224	107	3	13.5	72%	65%



(a) 当 δ 等于最优值 1.85 时的 PPI 网络的社团划分



(b) 当 δ 降到 1.4 时的网络社团划分

用大的节点表示领袖节点，重叠节点用黑色圆圈圈住

图 9　54 个节点和 119 条边的 PPI 网络的社团结构

结束语　在本文中，我们提出了一种新颖而高效的基于多尺度信息扩散的社团探测算法。该算法不同于现有的做法，而是试图把重点放在模拟这些网络的多尺度局域关联上面。它首先确定领袖节点，然后使用随机游走动态过程确定领袖节点周围的社团节点。通过利用归属向量，不仅能够发现重叠社团，而且还可以找到不同分辨率的多尺度社团，包括“粗粒化”的社团与最佳的划分。由于社团归属向量的计算是通过依靠 Markov 动力系统的迭代来完成的，因此它具有接近于线性的时间复杂度，快速方便。为了验证消息，我们将算法应用于真实世界的网络，包括几种典型的社会和生物结构。结果的合理性验证了方法的可行性，因此它可以用来检测各种现实复杂网络中的社团结构。综上所述，本文从一个全新的角度，提出了一种只利用局部信息扩散的高效快速的多层次社团结构检测算法。

参 考 文 献

[1] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science, 1999, 286(5439): 509-512

[2] Newman M E J. Fast algorithm for detecting community structure in networks[J]. Phys. Rev. E, 2004, 69: 066133

[3] Girvan M, Newman M E J. Community structure in social and biological networks[J]. Proc. Natl. Acad. Sci, 2002, 99(12): 7821-7826

[4] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys. Rev. E, 2004, 69: 026113

[5] Li H J, Zhang X S. Analysis of stability of community structure across multiple hierarchical levels[J]. Europhys. Lett., 2013, 103: 58002

[6] Li H J, Wang Y, Wu L Y, et al. Community structure detection based on potts model and spectral characterization[J]. Europhys. Lett., 2012, 97: 48005

[7] Li H J, Wang Y, Wu L Y, et al. Potts model based on a Markov process computation solves the community structure problem effectively[J]. Phys. Rev. E, 2012, 86: 016109

[8] Muff S, Rao F, Caflisch A. Local modularity measure for network clusterizations[J]. Phys. Rev. E, 2005, 72(5): 056107

[9] Gregory S. Finding Overlapping Communities Using Disjoint Community Detection Algorithms[M] // Complex Networks. Springer Berlin Heidelberg, 2009: 47-61

[10] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435: 814-818

[11] 沈华伟,程学旗,陈海强,等. 基于信息瓶颈的社区发现[J]. 计算机学报, 2008, 31(4): 677-686

[12] Raghavan U, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks[J]. Phys. Rev. E, 2007, 76(3): 036106

[13] Pujol J M, Bejar J, Delgado J. Clustering algorithm for determining community structure in large networks[J]. Phys. Rev. E, 2006, 74(1): 016107

[14] 王观玉. 基于聚类的复杂网络社团发现算法[J]. 计算机工程, 2011, 37(10): 58-60

[15] 杨博,刘大有,金弟,等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1): 54-66

[16] Noh J, Rieger H. Random walks on complex networks [J]. Phys. Rev. Lett., 2004, 92(11): 118701

[17] Lai D, Lu H, Nardini C. Enhanced modularity-based community detection by random walk network preprocessing [J]. Phys. Rev. E, 2010, 81(6): 066118

[18] Ravasz E, Barabasi A L. Hierarchical organization in complex networks[J]. Phys. Rev. E, 2003, 67: 026112

[19] Baras J S, Hovareshti P. Efficient and robust communication topologies for distributed decision making in networked systems [C] // Proceedings of 47th IEEE Conference on Decision and Control. 2008: 2973-2978

[20] Danon L, Duch J, Guilera D, et al. Comparing community structure identification[J]. J. Stat. Mech., 2005, 29: 09008

[21] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. J. Stat. Mech., 2008, 10: 10008

[22] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure [J]. Proc. Natl. Acad. Sci., 2008, 105(4): 1118-1123

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

[24] Huh W, et al. Global analysis of protein localization in budding yeast[J]. Nature, 2003, 425: 686-691

[25] Jansen R, Gerstein M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction[J]. Current Opinion in Microbiology, 2004, 7: 535-545

[26] Ruepp A, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genome[J]. Nucleic Acids Res, 2004, 32: 5539-5545

(上接第 114 页)

[14] 张国基,李璇,刘清,等. 基于广义信息域离散轨迹变换的随机数生成器[J]. 物理学报, 2012, 61(6): 0605021-0605029

[15] 董俊,朱文,蒲秀英,等. 物理真随机数发生器的设计[J]. 电光与控制, 2013, 20(2): 93-96

[16] 周庆,胡月,廖晓峰. 基于鼠标轨迹和混沌系统的真随机数产生器研究[J]. 物理报, 2008, 57(9): 5413-5418

[17] 李璇. 三元组密钥流发生器的机理及应用研究[D]. 广州: 华南理工大学, 2012

[18] Wang Fu-lai. A universal algorithm to generate preudo-random numbers based on uniform mapping as homemophism[J]. Chin. Phys. B., 2010, 19(9): 0905051-0905056

[19] Li Xuan, Zhang Guo-ji, Liao Yu-liang. Chaos-based true random number generator using image[J]. InternationalConference on Computer Science and Service System, 2011, 6(3): 2145-2147

[20] 张国基,徐浩,黎凤鸣. 基于广义信息域的高级加密系统及方法: 中华人民共和国, CN 101394268 B[P]. 2011. 05. 18

[21] 张国基,刘清,黎凤鸣,等. 基于广义信息域的动态加解密方法: 中华人民共和国, CN 101383703 B[P]. 2011. 04. 27

[22] 冯富强,陈举鹏. 在低信噪比条件下 DS-DS 信号的检测和参数估计[J]. 通信学报, 2002, 23(9): 63-68

[23] 贾志成,蒋文娟,李建娜. 基于蒙特卡罗的扩频通信仿真与分析[J]. 通信技术, 2007, 28(8A): 206-209

[24] Palmore J. Computer arithmetic, Chaos and Fractals [J]. Physical D., 1990, 42(1): 99-110

[25] 王亚东. 混沌序列在扩频通信中的研究与应用[D]. 西安: 西安电子科技大学, 2007

[26] 曾璐,谢晓尧. 基于 MATLAB 扩频通信系统误码率的研究[J]. 通信技术, 2011, 44(11): 25-29

[27] 吴成茂,李杜鹃,王保平. 混沌扩频通信及其误码率[J]. 西安邮电大学学报, 2013, 18(3): 10-13

[28] 黄乘顺,李星亮. 基于混沌的扩频通信系统及性能分析[J]. 通信技术, 2008, 41(12): 37-39

[29] 张国基,刘清,黎凤鸣,等. 基于广义信息域的伪随机码发生器及其发生方法: 中华人民共和国, CN 101364868 B[P]. 2012. 02. 01