复杂网络中的社团结构发现方法

邓智龙 淦文燕
(解放军理工大学指挥自动化学院 南京 210007)

摘 要 社团结构是真实复杂网络异质性与模块化特性的反映。深入研究网络的社团结构有助于揭示错综复杂的真实网络是怎样由许多相对独立而又互相关联的社区形成的,使人们更好地理解系统不同层次的结构和功能,具有广泛的实用价值。总结了目前常用的社区发现方法,包括经典的GN算法、模块度优化算法、基于网络动力学的方法以及统计推断方法;用社区划分基准测试网络Zachary对上述算法进行了实验,对这几类算法的时间复杂度和优缺点进行了比较分析。最后,对复杂网络的社区结构发现算法的研究进行了展望。

关键词 复杂网络,社团结构,社区发现,聚类
中图法分类号 N94 **文献标识码** A

Community Structure Detection in Complex Networks

DENG Zhi-long GAN Wen-yan
(*Institute of Command Automation, PLA University of Science and Techology, Nanjing 210007, China*)

Abstract Many networks of interest in the sciences, including social networks, computer networks, are found to divide naturally into communities or modules. Community structure can reflect the heterogeneity and modularity of the real-world networks. Finding the communities within a network is a powerful tool for understanding the structure and the functioning of the network. We reviewed some most popular methods for detecting community, including GN algorithm, modularity-based methods, dynamic algorithms, and the methods based on statistical inference. We used the standard testing network Zachary to test the above-mentioned methods, and analysed the time complexity and conclude the advantages and disadvantages of this methods. Finally, prospected of study on community detection methods.

Keywords Complex networks, Community structure, Community detection, Clustering

现代网络科学的发展为我们理解复杂系统带来了巨大的进步,随着许多大型真实世界网络数据集的获得,人们发现许多真实网络都具有某些共同的特性。1998年,Watts和Strogatz在Nature上发表论文,阐述了实际复杂网络的"小世界"效应[1],即网络节点间具有较小的平均最短路径长度,对数依赖于网络的规模。1999年,Barabási和Albert在Science上发表论文,指出许多真实网络的度分布遵循幂律分布,称为无标度网络[2]。此外,复杂网络研究还表明,真实网络不仅具有小世界和无标度等特性,还呈现明显的社区结构(community structure)[3]。

所谓社团(community,又称为cluster,module,cohesive subgroup),是一个节点内聚的子图,子图内部节点之间存在较多的连接,而子图之间的连接相对较少。社区结构是网络模块化和异质性的反映,表示真实网络是由许多不同类型节点组合形成的。如社会网中具有某种特殊功能的组织就是由聚在一起的个体(人)构成的,同样地,人体的各个器官也可以说是具有某种功能的社团。我们称网络具有社团结构,是指网络是由若干个社团或聚类构成的。这些社团的最主要的拓扑特征是社团内部节点的连接相对紧密,而社团间连接则相对稀疏[3]。如图1所示,网络包含3个社团,在这些社团内部,节点之间连接很紧密,而社团之间的连接则稀疏得多。例如,WWW可以看作是由大量网站社团组成的,其中同一个社团内部的网站为同一类型或观点相似的网站,这是因为同类网站的互链数比非同类网站的要高得多[4]。在生物网或社会网中,也可以根据不同的性质把节点划分为不同的社团。研究这些网络中的社团结构,可以帮助人们理解其结构和功能。目前,社团结构分析在计算机科学、生物学、物理学和社会学中都有极其广泛的应用[3,4]。

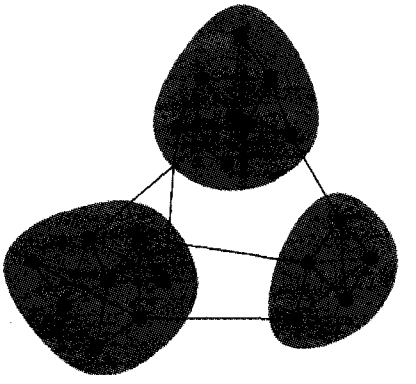


图1 网络中的节点分成的3个社团(取自文献[5])

本文第1节论述社团结构的定义及度量;第2,3节详细

本文受国家973计划项目(2007CB310800),国家自然科学基金项目(61035004)资助。
邓智龙(1986—),男,硕士生,主要研究方向为复杂网络、人工智能,E-mail:jxjadzl@gmail.com;**淦文燕**(1971—),女,副教授,主要研究方向为数据挖掘、复杂网络、人工智能。

介绍了当前的最流行的几类社区发算法，并对其进行了实验，分析了这些算法的特性；最后总结和展望了社区划分算法的研究。

1 社团结构的含义及其度量

关于网络社团结构，目前还没有一个被广泛认可的严格定义。较为常见的是基于相对连接密度的描述：网络中的节点可以分成组，组内连接稠密而组间稀疏。但是描述中提到的稀疏与稠密都没有明确的判断标准。目前常用的社团结构定义包括：基于网络局部拓扑的社团定义、基于模块度的社团定义和基于节点相似度的社团定义等。

基于网络局部拓扑的社团定义着眼在网络的子图结构上，若一个子图满足某种定量条件，则被认为是一个社区，如n-派系[6]、强社团[7]和弱社团[7]等。

基于模块度的社团结构定义认为：具有社团结构的网络拓扑与空模型之间存在明显的差异。所谓空模型[8]，是Newman和Girvan提出的在保持原始网络度序列的约束下，节点之间的边是随机连接的随机图。模块度是指原始网络中社区结构内部的边数比例与空模型中社团内部节点之间的期望边数所占的比例的差值。模块度的定义非常简洁、直观，一经提出便得到了广泛的应用，它不仅可以作为评价算法优劣的一个标准，还衍生出了基于最优化模块度的一系列社区发现算法，这类算法将在2.1节中介绍。

基于节点相似度的社团结构定义把相似的节点归属于同一个社区，核心在于节点的相似度计算。常用的方法是把网络节点映射为欧氏空间中的一个点，用点间距离来度量节点间的相似性。周海军[9]提出利用节点之间随机游走的往返时间来度量两个节点之间的距离和相似性，取得了不错的效果。

2 常用的复杂网络社团结构发现算法

近10年，由于网络新科学的兴起，各个领域的学者纷纷利用网络科学研究各自领域的一些网络，因此产生了很多具有领域背景的社团结构发现算法，如社会学方法、计算机科学方法、物理学方法等。这些算法性能不一，有的只能处理较小的网络，而有的则能对百万节点的网络进行社区划分；有的需要输入社区数目，而有的则会直接给出最优划分。本文只介绍4类最常用的社区划分算法：GN算法、模块度优化算法、基于网络动力学的算法和基于统计推断的算法。

2.1 GN算法与模块度优化算法

GN算法是Girvan和Newman在2002年提出的一个社区发现的分裂算法[3]。根据社区的描述，社区内部节点连接稠密，社区间的连接相对稀疏，社区间的少数连接将成为社区间通信时通信流量必经之路。考虑网络中某种形式的通信并寻找到具有最高通信流量的边，去除该边将获得网络最自然的分割。由此，Girvan和Newman等人引入边介数来度量网络的通信流量，提出基于边介数的社区发现算法，简称GN算法。

给定网络$G=(V,E)$中的任一边$e\in E$，其介数可定义为所有节点对之间的最短路径中经过该边的路径数，即：

$$B(e)=\sum_{v_i\neq v_j\in V} n_{ij}(e) \tag{1}$$

式中，$n_{ij}(e)$表示节点$v_i,v_j\in V$间最短路径中包括边e的数量。显然，介数越高的边对网络连通性越重要。如果删除网络中介数很高的边，很可能会影响网络的连通性，使得任意两点间最短路径长度增加。基于上述思想，GN算法迭代计算网络中每条边的介数并删除介数最大的边，直至网络中所有边被去除，每个节点自成一个社区。

GN算法无须指定社区个数，可以将网络分解成任意数量的社区，但无法确定最优的社区结构。事实上，即使明显不存在社区结构的随机网络，GN算法仍然会产生强制的层次社区结构。此外，GN算法的时间复杂度比较高，迭代计算所有边的介数的时间开销为$O(mn)$，总的时间复杂度为$O(m^2n)$。

针对GN算法的强制划分问题，Newman等引入模块度(modularity)[8]来评价社区分解的合理性。其思想是一个好的社区划分，其内部节点连接数应远大于它对应的空模型的节点连接数。

具体来说，假定网络$G(V,E)$可被分解为k个社区$G_1,\cdots G_k$，定义一个$k\times k$的对称矩阵$e=(e_{ij})_{k\times k}$，其中e_{ij}表示网络中连接社区G_i和G_j的边所占的比例。对角元之和$Tre=\sum_{i=1}^{k}e_{ii}$表示连接社区内部节点的边所占的比例，行(列)元素之和$a_i=\sum_{j=1}^{k}e_{ij}$表示连接社区G_i中节点的边所占的比例，显然a_i只依赖于社区G_i中节点的度值，与网络中是否存在社区结构无关。如果将网络看作是一个在节点度值给定的前提下生成的随机网络，任选一条边连接社区G_i和G_j的概率为a_ia_j，则连接社区内部节点的概率为$\sum_{i=1}^{k}a_i^2=e^2$，其中，$\|\ \|$表示矩阵的所有元素之和。由此，可以得到网络社区划分的模块度定义：

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

一般情况下，好的社区划分内部节点连接概率Tre应远大于空模型中内部节点的连接概率$\|e^2\|$，即模块度越大，社区划分的质量越好。Newman等的研究表明，具有社区结构的真实网络的模块度通常取值在0.3～0.7之间[8]。

模块度的定义由于独立于特定的社区发现算法，因此可以用作衡量社区划分合理性的基准度量。从这种意义上说，社区发现问题可以等价于模块度优化问题。考虑到网络可能的划分与网络规模呈指数关系，人们提出了多种启发式模块度优化算法，如Newman快速算法[10]、极值优化方法[11]、谱优化方法[5]等来对大规模网络进行社区划分。

2.2 基于网络动力学的算法

2.2.1 Potts模型

Potts模型是统计力学中最流行的模型之一[12]，它描述一个粒子系统可以有q个不同的状态。这些粒子之间的相互作用是铁磁性(ferromagnetic)的。因此，系统倾向于把粒子有序排列。在零温度状态时，所有的粒子都处在同一个状态。但如果额外提供一个反磁性的作用力，那么系统的基态可能并不是所有粒子有序排列，而是同时存在不同的粒子自旋态(spin values)。如果把粒子对应为具有社区结构的网络中的节点，粒子的相互作用仅仅在于其邻居之间，则可以认为，具有相同自旋态的粒子属于同一个社区，从而得到网络的社区划分。基于这个想法，Reichardt和Bornholdt[13]提出了一种社区检测方法，它把图映射到一个零温度状态的有最近邻居相互作用的q-Potts模型。模型的哈密顿函数(Hamiltonian)，即它的能量为：

$$H=-J\sum_{i,j}A_{ij}\delta(\sigma_i,\sigma_j)+\gamma\sum_{s=1}^{q}\frac{n_s(n_s-1)}{2} \tag{3}$$

式中，A_{ij} 为网络邻接矩阵中的元素；σ_i 为节点 i 的自旋态，也可以看作是社区的编号；n_s 为状态为 s 的粒子数目；J，γ 为耦合参数；H 由两个对立项构成：第一项是经典的 Potts 模型的能量，倾向于把所有粒子排成有序状态，即把所有节点放在一个社区中，以达到能量的最大值，取负号则是最小值。第二项则相反，它的最小值在每个粒子都是不同状态时，即每个粒子都自成一个社区时取得。我们的目标是寻找合适的划分，使得能量 H 最小。γ/J 表示式(3)中前后两项的相对重要性：通过调节 γ/J，可以探求系统中不同层次的模块性。如果 γ/J 设定为图 g 中边的平均稠密度 $\delta(g)$，那么粒子在子图中排列有序，就对应于子图内部边密度超过 $\delta(g)$，同时外部边密度小于 $\delta(g)$，这样会使得 H 更小。确定目标函数后，采用模拟退火算法来进行搜索。具体过程如下：

(1)给定系统一个初始温度，给网络中每个节点随机地赋予 q 个自旋态中的一个。

(2)随机选择一个节点，改变它的自旋态。

(3)如果新产生的系统能量 H 比未修改之前的系统能量更小，即 $\Delta H=H_{后}-H_{前}<0$，则接受对这个节点自旋态的修改；如果 $\Delta H=H_{后}-H_{前}>0$，就在(0,1)之间随机选择一个数 ε，若 $\varepsilon<e^{-\beta\cdot\Delta H}$，其中 $\beta=\frac{1}{T}$，则也接受新的自旋态，否则保持不变。

(4)回到第(2)步，遍历网络中的所有节点。

(5)降低系统温度，重复上面的步骤，直到温度降到某个值为止。此时，根据节点的自旋态把节点分成若干个社团。

可以证明，当初始状态个数 q 较大时，算法的结果并不依赖于 q 值，即得到的社团个数与 q 无关。如果我们把粒子间的相互作用范围扩大到无限，即网络中任何两个粒子间都有相互作用，此时模型的哈密顿能量为：

$$H=\sum_{i<j}\delta(\sigma_i,\sigma_j)(\gamma-A_{ij}) \tag{4}$$

这种方法也可以很容易地推广到加权网络上，只需要把邻接矩阵 A 替换成网络权矩阵 W。Song 等在 2006 年提出了一种基于随机磁场伊辛模型(FRFIM)的聚类技术[14]。对于一个权矩阵为 W 的网络，它的哈密顿能量为：

$$H=-\frac{1}{2}\sum_{i,j}W_{ij}\sigma_i\sigma_j-\sum_i B_i\sigma_i \tag{5}$$

式中，$\sigma_i=\pm1$，B_i 为节点 i 的随机磁场。这个模型的性能依赖于磁场的选择，Song 等把除了两个节点外的所有节点的磁场设为 0，非零的两个节点表示为 s，t。它们具有无限的作用距离和相反的符号，这等价于把粒子 s，t 设为有相反的自旋态。算法的思想很简单，如果 s，t 分别是两个社区的中心节点，那么它们会把它们的自旋态作用于社区内的其他节点。所以最小能量的状态是网络分为两个子图，一个子图内部节点都具有正的自旋态，另一个都具有负的自旋态。这两个子图则对应于我们要得到的社区。这个问题等价于一个最大流问题。s，t 分别为源点和汇点，具体可参见文献[14]。需要注意的是 s，t 的选择问题。在缺少结构信息的情况下，根据算法，我们可能针对所有可能的 s，t 节点对进行算法迭代，当 s，t 恰在同一个社区内时，得到的划分结果很可能是错误的。对于稀疏网络，算法的时间复杂度大约为 $O(n^{2+\theta})$($\theta\approx1.2$)，只能处理非常小的网络。但是如果我们知道网络中的重要节点，算法的复杂度会降为 $O(n^{\theta})$($\theta\approx1.2$)，此时可以处理上百万节点规模的网络。

2.2.2 随机游走

随机游走[15]可以用来发现社区结构。如果一个图有较强的社区结构，那么随机游走者将由于社区内部边的密度较高，走到社区内部边的概率更大。这里我们介绍最流行的基于随机游走的聚类算法。

首先，周海军[9]等用随机游走来定义每对节点之间的距离：i，j 之间的距离 d_{ij} 定义为一个随机游走者从节点 i 出发，到节点 j 结束经过的平均步数。定义了一个节点 i 的全局吸引子为节点 i 的最近节点(从节点 i 出发，随机游走最近的节点)，同时定义节点 i 的局部吸引子为它的最近邻居。这里可以根据其是局部吸引子还是全局吸引子定义两类社团：节点 i 应该和它的吸引子在同一个社区中，那些把 i 作为吸引子的节点也应该与 i 在同一个社区中。社区必须为极小子图，因此它不能包含更小的根据这个标准划分出的社区。把该算法应用在真实网络中，如 Zachary[16]和美国大学足球联赛网[3]上，结果显示，这种方法可以发现有意义的划分。

计算节点距离矩阵要求解 n 个线性代数方程式，时间为 $O(n^3)$。但是，一个精确的距离矩阵计算是没有必要的，由于计算节点的吸引子可以通过仅仅考虑节点周围的一个局部范围，因此，这种方法也可以应用在大型图上。

2.2.3 同步

同步[17]是指系统中的单元都处于一个相同或相似的状态，是由于系统中单元之间的相互作用而产生的一种涌现现象，在自然和社会现象中普遍存在。同步也可以被用来发现社区，如果开始某个节点有一个振荡，且相位随机，由于邻居间存在相互作用，振荡会首先在社区内实现同步，但是要达到整个网络的同步则需要较长时间。如果我们观察网络随时间演化，那么可以发现在同一个社区中的节点状态稳定相似并且这种状态持续时间较长。这一现象是由 Arenas[18]等发现的。他们使用一个耦合二维向量的 Kuramoto 振荡器。发出一个合适频率的振荡。在 Kuramoto 模型中，节点 i 的相位 θ_i 根据下面的公式动态变化：

$$\frac{d\theta_i}{dt}=\omega_i+\sum_j K\sin(\theta_j-\theta_i) \tag{6}$$

式中，ω_i 为节点 i 的固有频率，K 为振荡器之间的耦合强度，累加表示所有其他节点的振荡对节点 i 的影响。如果节点间的相互作用超过了一个根据固有频率分布范围决定的某个阈值，那么这个过程产生了一个同步。如果把动力学过程放在一个图上，每个振荡器仅与它的最近邻居耦合。为了揭示局部同步的效果，Arenas 等引入了局部有序参数

$$\rho_{ij}(t)=\langle\cos[\theta_i(t)-\theta_j(t)]\rangle \tag{7}$$

来衡量振荡器 i 和 j 之间的平均相关性。在不同的初始条件下计算这个平均值。通过在一段给定时间 t 内可视化相关性矩阵 $\rho(t)$，可以区分出同步节点构成的社团。实际中，采用的是动态连接矩阵 $D_t(T)$ 的方法，$D_t(T)$ 是一个二元矩阵，通过阈值化 $\rho(t)$ 中的元素得到。从 $D_t(T)$ 的谱中，在 t 时刻，可以获得不连通构件的个数。我们把构件的数目作为时间 t 的函数。在某些特定时间尺度上，可能出现一段较长的稳定期，即网络构件的个数随时间不变。此时就表示该图存在稳定的社区结构(见图 2)。实验表明，此时网络对应的划分具有较高

的模块度。稳定期出现在不同的时间尺度上，则意味着网络存在层次结构。经过一段长时间 t 后，所有的振荡器都同步了，则表示整个系统表现为一个单个的构件。

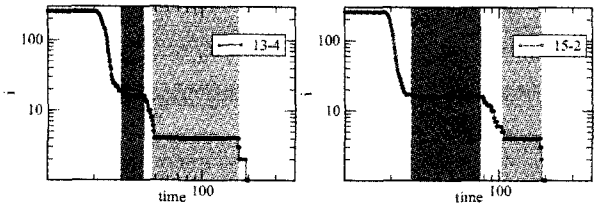


图 2　不同边密度的两层社区结构的网络的同步构件数 i 随时间 time 的变化(取自文献[18])

同步算法的时间复杂度为 $O(n^2)$，应用在 Zachary 网和 GN 的基准测试网络上，该算法能给出一个优秀的结果。但是当网络中社区的大小相差较大时，同步的社区发现算法可能并不十分可靠，但是在这方面的测试还没有研究。

2.3　基于统计推断的方法

统计推断[19]是从一个观察集或者证据和一个假设模型出发，对数据集利用数理统计方法发现数据的某些特性。这里我们要研究的数据集是一个图，假设模型必须满足实际网络的拓扑结构，即节点之间的连接关系。这一节回顾一些找到最佳拟合原图的、具有社区结构的假设模型的技术。

贝叶斯推断是用证据来推断某个给定假设模型为真的概率。它由两个部分组成，一个是证据，即有关于图的信息 D，另一个是参数集为 $\{\theta\}$ 的统计模型。贝叶斯推断首先写出似然度 $P(D|\theta)$。因此目标函数为选择一个参数集 $\{\theta\}$，来最大化似然度 $P(D|\theta)$，也即是最大化后验概率 $P(\theta|D)$。因为根据贝叶斯公式有：

$$P(\{\theta\}|D)=\frac{1}{Z}P(D|\{\theta\})P(\{\theta\}) \tag{8}$$

其中，$P(\{\theta\})$ 为模型参数的先验分布，

$$Z=\int P(D|\{\theta\})P(\{\theta\})\mathrm{d}\theta \tag{9}$$

这里的主要困难是参数集的先验概率 $P(\{\theta\})$ 不容易得到，而且式(8)的计算也比较困难。图的社区划分问题可以看作是一个特殊的统计推断问题，这里的证据是图的结构信息，一般就是图的邻接矩阵。另外，假设模型则是我们对网络的一种假设划分。通常情况下，参数集 $\{\theta\}$ 表示为形式 $(\{q\},\{\pi\},k)$，其中 q 表示节点的社团指派，π 表示模型的参数，k 为聚类的数目。这里我们介绍最常用的随机分块模型。

随机分块模型[19]是一种通用的随机网络模型，可用来发现网络社区，同时也可生成作为基准的人工网络。在最简单的随机分块模型中，所有的 n 个节点被指派到 K 个社团中。节点对之间独立地放置一条无向边(不考虑多边的情况)，放置边的概率和节点无关，仅仅与节点隶属的社团相关。具体来说，如果节点 i 隶属的社团表示为 g_i，那么可以定义一个 $K\times K$ 矩阵 ψ,ψ_{g_i,g_j} 表示节点 i,j 之间有一条边的概率。应用在社区发现中，用一个随机分块网络来对给定网络进行拟合，这种方法在社会网文献中称为后验分块建模。随机分块模型比传统方法更具一般性，它可以发现很多形式的结构，而不仅仅是有密集连接的简单社团。

为了使计算更加简单，下面的网络中将允许存在多边和自循环边，尽管在真实网络中并没有这种边，但对于稀疏网络，并不会有显著影响其结果。

令我们要进行社团划分的原始图表示为 G，它是节点数为 n 的无向图，邻接矩阵为 A。现在我们假设节点之间的边独立地服从泊松分布。注意 A_{ij} 为节点 i 的自循环边的两倍。定义 ω_{rs} 为邻接矩阵 A_{ij} 的期望值，其中 i,j 分别隶属于社团 r,s 中。现在可以根据这个参数写出拟合出图 G 的概率：

$$P(G|\omega,g)=\prod_{i<j}\frac{(\omega_{g_ig_j})^{A_{ij}}}{A_{ij}!}\exp(-\omega_{g_ig_j})\times\prod_i\frac{(\frac{1}{2}\omega_{g_ig_i})^{\frac{A_{ii}}{2}}}{(\frac{A_{ii}}{2})!}$$

$$\exp(-\frac{1}{2}\omega_{g_ig_i}) \tag{10}$$

因为 $A_{ij}=A_{ji}$，$\omega_{rs}=\omega_{sr}$，所以式(10)可以简化为：

$$P(G|\omega,g)=\frac{1}{\prod_{i<j}A_{ij}!\ \prod_i 2^{\frac{A_{ii}}{2}}(\frac{A_{ii}}{2})!}\times\prod_{rs}\omega_{rs}^{\frac{m_{rs}}{2}}$$

$$\exp(-\frac{1}{2}n_rn_s\omega_{rs}) \tag{11}$$

式中，n_r 为社团 r 中的节点数目。m_{rs} 为社团 r,s 之间的边数，当 $r=s$ 时，为两倍的社团内部边数。我们的目标是在不知道模型参数 ω_{rs} 和节点的社团指派的情况下，最大化这个拟合概率。在大多数情况下，我们只需最大化这个概率的对数。忽略常数和不依赖于参数和社团指派的项，式(11)的对数为：

$$\log P(G|\omega,g)=\sum_{rs}(m_{rs}\log\omega_{rs}-n_rn_s\omega_{rs}) \tag{12}$$

分两步最大化式(12)，首先考虑模型参数 ω_{rs}，然后再考虑社团指派 g_i。模型参数的最大似然估计 $\hat{\omega}_{rs}$ 可以通过简单的变形得到：

$$\hat{\omega}_{rs}=\frac{m_{rs}}{n_rn_s} \tag{13}$$

把式(13)代入式(12)中，得到：

$$\log P(G|\omega,g)=\sum_{rs}m_{rs}\log(\frac{m_{rs}}{n_rn_s})-2m \tag{14}$$

式中，$m=\frac{1}{2}\sum_{rs}m_{rs}$，为网络的总边数。去掉这个常量，得到：

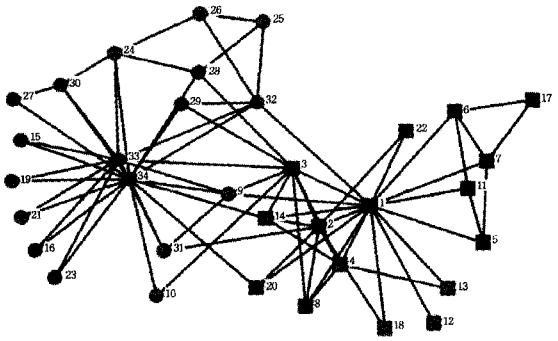
$$\log P(G|\omega,g)=\sum_{rs}m_{rs}\log(\frac{m_{rs}}{n_rn_s}) \tag{15}$$

现在我们的目标是最大化这个量，得到的社团指派即为网络的最佳拟合模型。上面的标准随机分块模型中，节点 i,j 之间的连接概率 A_{ij} 只与 i,j 所处的社团相关，而与 i,j 节点它们自身的度无关。但是我们知道，度大的节点它的连接概率也应该大，因此，Karrer，Newman 对其进行了改进，把节点的度信息加入到模型中，提出了一个度修正随机分块模型，具体可见文献[20]。

3　实验结果与比较

这里，选取社区划分的基准网络空手道俱乐部社会关系网络 Zachary[16]来测试上述算法的有效性。Zachary 网络是由 Wayne Zachary 观察一所大学的跆拳道俱乐部长达 3 年(1970～1972 年)之久得出的，其数据通过情报和俱乐部对大学活动的记录获得。节点代表俱乐部的成员，边代表成员间的友谊。34 号节点是主管员 John A.，1 号节点是空手道教练 Mr. Hi。该网络包括 34 个节点和 78 条边，如图 3 所示。实验结果如表 1 所列，可以看出，谱优化算法和随机游走模型对网络的划分正确率为 100%，划分效果最好；GN 算法、极值优化算法以及随机分块模型对网络的划分也只错了一个节点，其中 GN 算法划分错误的节点是 3 号节点，而 3 号节点仅在拓扑上，可以被划分到任一社区中。同步模型把网络划分

成3个社团，其中社团1被分成了两个社团，与真实划分相差较远。



方形代表教练一方，节点数为16，圆形表示主管一方，节点数为18

图3 空手道俱乐部成员关系网络

表1 算法对Zachary网络的划分结果

算法名称	社区1正确划分节点数 n_1	社区2正确划分节点数 n_2	划分正确率 $(n_1+n_2)/34$
GN算法	15	18	97%
极值优化算法	16	17	97%
谱优化算法	16	18	100%
随机游走模型	16	18	100%
同步模型	11	18	85%
随机分块模型	16	17	97%

另外，真实研究的复杂网络规模通常较大，如社会网与Web网等。社区发现算法的性能也是实际应用中需要考虑的问题，我们分析了上述几类算法的时间复杂度，如表2所列。可以发现，当提前知晓网络中重要节点时，Potts模型具有最小的时间复杂度；Newman快速算法、随机分块模型和同步模型具有相同的时间复杂度。在配合模块度指标下，Newman快速算法可以对网络进行较好的划分，且不需要提前输入社区数目；随机分块模型具有优良的时间复杂度，对网络的划分效果也较好，缺点是需要提前知道社团个数；同步模型算法相对复杂，但具有其他算法不具有的优点，可以根据同步时间发现网络不同层次的社区结构。划分效果最佳的是谱优化算法和随机游走模型算法，虽然时间复杂度相对高一点，但对于10万个节点以下的中等规模网络仍是最佳的选择。

表2 算法的时间复杂度和优缺点

算法分类	算法名称	时间复杂度	优缺点
传统算法	GN算法	$O(m^2n)$	时间复杂度高；算法结束需要人为的控制；划分效果较好
基于模块度最优化算法	Newman快速算法	$O(n^2)$	时间复杂度低，可以分析大型网络；结果可能不是全局最优
	极值优化	$O(n^2\log n)$	速度较快；可以得到极优的模块度值；依赖初始划分
	谱优化	$O(n^2\log n)$	速度较快；划分效果优良
基于网络动力学算法	Potts模型	$O(n^\theta)(\theta\approx1.2)$	在知道重要节点的前提下，速度极快；算法较复杂
	随机游走模型	$O(n^2\log n)$	算法较快；划分效果优良
	同步模型	$O(n^2)$	时间复杂度低；可以发现网络层次结构；算法相对较复杂
基于统计推断算法	随机分块模型	$O(n^2)$	时间复杂度低；算法简单；需要提前输入社区数目

结束语 社团结构是真实复杂网络异质性与模块化特性的反映。深入研究网络的社区结构有助于揭示错综复杂的真实网络是怎样由许多相对独立而又互相关联的社区形成的，可使人们更好地理解系统不同层次的结构和功能特性，具有广泛的实用价值。针对复杂网络中的社区检测问题，本文总结了目前常用的社区发现方法，包括经典的GN算法、模块度优化算法、基于网络动力学的方法以及统计推断方法；并用社团划分的基准测试网络Zachary对上述算法进行了测试，对这几类算法的时间复杂度和优缺点进行了总结。

近10年，人们对网络社团划分问题进行了大量的研究，提出了几十种社区划分方法，作者所在的课题组也提出了一种基于拓扑势的网络社区发现方法[21]，在实验网络中取得了较好的效果。尽管有这些努力，网络的社区划分还是没有形成一个整体的框架，社区发现算法应该如何才能发现网络真实的社区结构仍是一个悬而未决的问题。在这方面，Newman等提出的空模型及模块度为人们评价社区划分算法的划分质量做了有益探索，但是空模型只保留原始网络的度信息，其他网络结构信息，如介数、聚集系数等则无法应用在社区划分的评价之中。

传统的社区发现算法大都实验在简单的无向无权网络上，而真实的网络可能是节点有权重、边是有向的。在这方面，虽然也有一定研究，但大都基于无向无权网络的算法的简单扩展，算法的效能还有较大提升空间。另外一个重要的方面，即真实网络的社区往往是重叠的，而且具有层次性，例如，我们可以同时属于若干个社团，社团之间可能存在着隶属层次关系。考虑真实复杂网络社区结构的这些特点是我们进行下一步研究的重要方向。

参考文献

[1] Watts D S. Strogatz. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393: 440-442

[2] Barabási, Albert. Emergence of scaling in random networks[J]. Science, 1999, 286: 509-512

[3] Girvan M, Newman M E J. Community structure in social and biological networks[J]. PNAS, 2002, 99: 7821-7826

[4] Gleiser P, Danon L. Community structure in jazz[J]. Advances in Complex Systems, 2003, 6: 565-573

[5] Newman M E J. Modularity and community structure in networks [J]. PNAS, 2006, 103(23): 8577-8582

[6] Palls G, Deruyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society [J]. Nature, 2005, 435(7043): 814-818

[7] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. PNAS, 2004, 101: 2658-2663

[8] Newman M E J, Girvan M. Finding and evaluating community structure in networks[J]. Phys. Rev. E, 2004, 69(2): 026113

[9] Zhou H. Distance, dissimilarity index, and network community structure[J]. Phys. Rev. E, 2003, 67(6): 061901

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

[11] Duch J, Arenas A. Community detection in complex networks using extreme optimization[J]. Phys. Rev. E, 2005, 72: 027104

[12] http://en. wikipedia. org/wiki/Potts_model. [EB/OL]. 2011-11-10

[13] Reichardt J, Bomholdt S. Detecting fuzzy community structures in complex networks with a Potts model[J]. Phys Rev Lett, 2004, 93(21): 218701

[14] Son S-W, Jeong H, Noh J D. Random field Ising model and community structure in complex networks[J]. Eur. Phys. J, 2006, B50(3):431
[15] Hughes B D. Random Walks and Random Environments: Random Walks[M]. Oxford, UK: Clarendon Press, 1995
[16] Zachary W W, Anthropol J. Res. 1977, 33:452
[17] Kuramoto Y, Oscillations C. Waves and Turbulence [M]. Berlin, Germany: Springer-Verlag, 1984
[18] Arenas A, Diaz-Guilera C, Perez-Vicente J. Synchronization Reveals Topological Scales in Complex Networks[J]. Phys. Rev. Lett, 2006, 96(11):114102
[19] Fortunato S. Community detection in graphs [J/OL]. arXiv: 0906.0612v2[physics. soc-ph], 2010
[20] Karrer B, Newman M E J. Stochastic blockmodels and community structure in networks[J/OL]. arXiv: 1008.3926v1[physics. soc-ph], 2010
[21] 淦文燕, 赫南, 李德毅, 等. 一种基于拓扑势的网络社区发现方法[J]. 软件学报, 2009, 20:8

(上接第 95 页)

间、数据文件传输开始和结束的时间以及数据文件在临时目录 Dropbox 中被删除的时间等。通过这些日志信息的记录，可以实时地观察到某个时刻数据传输的阶段，同时也可以根据错误信息较快地进行问题排查和故障诊断。图 6 是一个数据文件传输过程中日志信息的截图。

3.8 监视模块

为了更加形象化地监视大批量数据传输系统传输过程中的效果，监视模块在日志模块采集日志的基础上，采用 Web 页面图形化的方法实时显示每个时刻的传输结果，包括文件传输的个数、文件的大小(byte 单位)等，具体如图 7 所示。

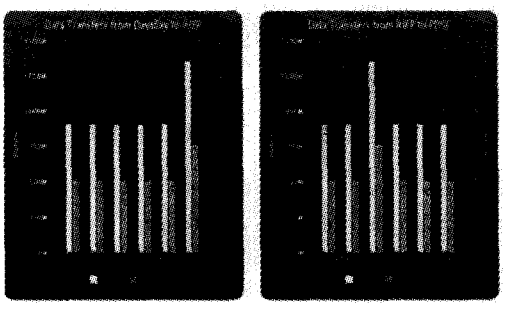


图 7 监视模块中数据文件传输过程效果图

此外，监视模块会对每天传输到各个数据中心和计算中心的文件进行统计，并在当天晚上 12 点发邮件通知管理员，以便管理员进行数据统计和排查问题。

3.9 配置管理模块

配置管理模块利用通过 JBoss 的 MBean 提供的页面管理服务，图形化地管理系统的各个子模块。在 Web 页面上，管理员可以设置 DAQ 数据的磁盘阵列与系统缓冲区的映射关系、远程数据中心和计算中心的机器名、Fetcher 程序轮询 Dropbox 的时间间隔、发送模块和接收模块用于通信的邮件名、数据存放路径、传输超时时限、数据文件的自动重传次数、传输的数据文件类型的注册、传输失败文件存放目录等。

4 系统应用

本文介绍的大批量数据传输系统已经在大亚湾中微子实验中进行部署，并完成了 4 次测试数据的传输，已经完成了 14.1TB 的数据传输，取得了良好的效果。部署图如图 8 所示，分别在大亚湾现场、高能所和高能所的合作单位部署了大批量数据传输系统，采用数据中继的方式，现场的数据传输系统从现场数据磁盘缓冲区中获取原始数据后，将其发送到高能所的数据传输系统，高能所的数据传输系统再将其中继到其合作单位，合作单位的数据传输系统接收数据，并进行线下分析。

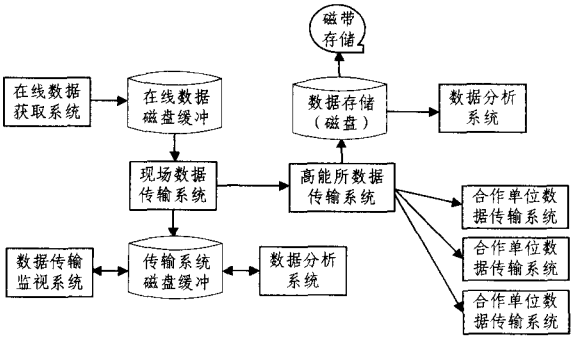


图 8 大批量数据传输系统在大亚湾中微子实验中的部署图

结束语 本文描述的大批量数据传输系统已经应用于大亚湾与高能所的物理实验数据传输过程中，取到了良好的应用效果。但是 DAQ 获取数据速率较大时，使用数据中继方式传输数据，数据接收的时延会比较长，从而影响远程的数据中心分析人员对数据进行分析的实时性。经过分析，产生该问题的原因有两个方面，一是系统部署的问题，具体指大批量数据传输系统的数据处理和数据存储使用的磁盘是同一块磁盘，由于磁盘自身的 IO 问题导致互相影响，目前已经将数据处理和数据存储使用的磁盘分离，以对时延的缩短产生良好的效果。二是系统源代码本身的冗余问题，这是今后改进的方向。

参 考 文 献

[1] Allcock W, Bresnahan J, Kettimuthu R, et al. The Globus Striped GridFTP Framework and Server[C]// Proceedings of the 2005 ACM/IEEE conference on Supercomputing (SC'05). Washington, DC, USA: IEEE Computer Society, 2005:54-54
[2] Allcock W, Bresnahan J, Bresnahan A, et al. GridFTP: Protocol Extension to FTP for the Grid[C]// Grid Forum Internet-Draft, 2001. Washington, DC, USA: IEEE Computer Society, 2001:21-24
[3] Momjian B. PostgreSQL: introduction and concepts. Bruce Momjian, PostgreSQL: introduction and concepts [M]. Boston, MA, Addison-Wesley Longman Publishing Co., Inc., 2001:47-58
[4] Stark S. JBoss Administration and Development[M]. The JBoss Group, USA, 2003:145-159
[5] Kounev S, Weis B, Buchmann A. PerformanceTuning and Optimization of J2EE Applications on the JBoss Platform[J]. Journal of Computer Resource Management, 2004, 2(113):129-135
[6] 聂思敏, 张吉龙, 谭有恒. 羊八井宇宙线观测数据实时传输及处理系统[J]. 核电子学与探测技术, 2007, 2(1):14-17
[7] Patton S. Spade and Data Movement[EB/OL]. http://dayabay.ihep.ac.cn/DocDB/0055/005598/001/DataMovement.pdf, 2010