

多尺度的社团结构稳定性分析

李慧嘉¹⁾ 李慧颖²⁾ 李爱华¹⁾

¹⁾(中央财经大学管理科学与工程学院 北京 100081)

²⁾(清华大学自动化系 北京 100084)

摘 要 社团结构分析是一项非常重要且具有挑战性的工作,已经引起来自不同领域学者的广泛关注.在该文中,作者创新性地结合 Potts 模型和 Markov 动态过程,提出了衡量多尺度社团结构稳定性的完整理论框架.对于给定的网络,该文通过揭示网络社团结构及其自旋动态的局域一致行为之间的关系,可以不使用特定的算法而直接获得社团结构相关的重要隐藏信息,比如社团结构的稳定性和在多个时间尺度的社团结构的最佳数量.它还克服了传统方法的不足,如模块度 Q 的分辨率局限性问题.进一步基于理论分析,该文给出一个无参数的社团结构探测算法.该算法通过计算每个节点的归属向量,可以识别网络的模糊社团结构,从而在多个层次上描述了每个节点参与重叠社团的程度.同时该文也证明了算法的可扩展性和在实际大型网络上的有效性.

关键词 社团结构; Potts 模型; Markov 过程; 稳定性; 多尺度; 社会计算; 社交网络; 复杂网络

中图法分类号 TP393

DOI号 10.3724/SP.J.1016.2015.00301

Analysis of Multi-Scale Stability in Community Structure

LI Hui-Jia¹⁾ LI Hui-Ying²⁾ LI Ai-Hua¹⁾

¹⁾(School of Management Science and Engineering, Central University of Finance and Economics, Beijing 100081)

²⁾(Department of Automation, Tsinghua University, Beijing 100084)

Abstract The analysis of stability of community structure is an important problem for scientists from many fields. Differently from the previous studies, we established a new framework to study the dynamics of Potts model for community structure detection by using the Markov process, which has a clear mathematical explanation. Critical topological information regarding multivariate spin configuration could also be inferred from the spectral significance of the Markov process. We test our framework on some example networks and find that it does not have resolution limitation problems at all. Finally a parameter-free algorithm is developed to determine fuzzy communities based on the belonging vectors across multiple timescales. Results have shown the model we proposed is able to uncover the hierarchical structure in different scales effectively and efficiently.

Keywords community structure; Potts model; Markov process; stability; multi-scale; social computing; social networks; complex networks

1 引 言

社团结构分析是复杂网络研究中非常重要的部

分,已经吸引了来自不同学科的科学家的广泛关注.从直观上来讲,社团是网络中的一些子集,在子集内部边的密度要比子集之间的边的密度大.在起始阶段,社团研究分析局限于规则网络.最近,由于一

些真实的网络特征^[1-7]被发现,例如无标度网络特性,很多的研究开始集中于实际世界的网络.在这个层面上,社团结构可以揭示实际网络的拓扑结构和功能,能够在很多研究领域^[8-10]得到广泛的应用.

社团结构分析中一个非常重要的指标是模块度函数,它第一次由 Newman^[1]提出并用来评估网络的社团划分.最优化模块度能够得到合理有效的社团划分并且已经被广泛应用到各种实际网络中^[11-13].除了对模块度概念的研究,很多科学工作者开始致力于研究并发现网络内部动态过程的性质.具体而言,研究人员已经开始探索社团结构和网络中的动态过程之间的相关性.例如,Arenas 等人^[2-3]指出同步过程能够显示复杂网络的拓扑特性.此外,Markov 动态过程也被广泛地应用在网络上,用于揭示社团结构^[14-17].在文献[18]中,Markov 过程被引入并定义了网络节点之间的粗粒化距离,并根据粗粒化距离提出了网络社团划分的优化算法.另外,作者在文献[19]中提出了对网络划分进行排序和量化的稳定性指标,该指标定义为社团化结构网络上 Markov 过程的自协方差,能够在形式上对应多种著名的划分指标,如模块度 Q .

除了同步和 Markov 过程,Potts 动态模型也是一种常见的热动力学模型,已经被用于挖掘网络中的社团结构.利用 Potts 模型发现社团结构,也被叫做超顺磁聚类法,自从被 Blatt 等人^[20]首次引入后已经得到了深入的研究.在该模型中,我们赋予网络中每个点一个 Potts 自旋值用以表示点的社团归属,而且在临近的自旋值之间存在着关联.然后社团结构可以通过类似的自旋配置来揭示,自旋配置内部比外部有更多的相互作用.从方法层面上来说,如邻居定义的依赖性,相互作用的类型,可能的中间状态和数据集的规模大小,都已经得到深入的研究^[21-25].Reichardt 和 Bornholdt^[26]引入了一种新的 Hamiltonian 函数,能够在全局的约束下确定合适的社团结构.该方法使得我们能够将社团结构映射到一个零温度的 q 状态 Potts 最近邻相互作用模型.

我们注意到,Markov 过程的动力学可以自然地反映社团结构的内在特性与自旋配置的局部一致性行为.然而,Potts 模型的动态和 Markov 过程之间的关联,一直没有得到深入的研究.Potts 模型和 Markov 过程都属于动力学的范畴,能够高效地反映组织中个体的动态演化性质与特征.但是它们在反映的具体形式和理论内容方面有较大差别.其中,Potts 模型有着非常形象的演化动态描述,能够利用

自旋值(spin value)随温度的变化,有效地描述社团归属随时间变化的动态过程.但是 Potts 模型没有具体的动态演化数学模型,目前为止大部分工作只是利用最优化 Hamiltonian 函数去划分出社团的归属.而 Markov 过程有着非常成熟的理论基础,能够轻松地反映个体状态的动态演化,但是 Markov 过程并不能利用随机游走者(random walker)的位置直接反映社团的归属演化.因此如果能够把 Potts 模型形象具体的动态描述和 Markov 过程成熟的数学基础结合起来,便能够发挥两种动力学模型的优势,更加有效地研究社团结构的内在特征.在本文中,我们创新性地结合 Potts 模型和 Markov 动态过程,研究自旋配置的局域一致现象,揭示网络社团结构与自旋动态亚稳定性之间的关系,并进一步提出了新颖的多尺度社团结构稳定性.对于任何给定的网络,我们不需要使用特定的算法就可以直接获得社团结构相关的重要信息,比如社团结构的稳定性度量和多个时间尺度的社团结构的最佳数量.它还克服了传统方法的不足,如模块度 Q 的分辨率局限性问题.基于理论分析,接着我们给出一个无参数的社团结构探测算法.该算法通过将每个节点与归属向量相结合,可以对网络的模糊社团结构进行识别,从而描述了每个节点参与重叠社团的程度.同时我们也证明了算法的可扩展性和在实际大型网络上的有效性.

2 Potts 模型和自旋关联矩阵

Potts 模型通过将网络模拟成一个非同质的铁磁系统,将每个数据点看成网络中具有标号的节点,网络中的每个节点都关联一个标号.在这里,标号被认为是网络的自旋值,其中自旋值是 Potts 模型中的一个重要概念,能够表示数据点状态随温度的变化情况.我们定义系统的结构为节点及其相互之间的关联,这种关联被温度所调控.当系统的温度较低时,所有节点的标号是相同的(所有的自旋值是均衡的),这对应于网络中只存在一个社团.随着温度的上升,单一的社团开始分裂而且节点间相对比较弱的自旋关联逐渐消失.

对于一个具有 N 个节点的无权无自环网络,自旋配置 $\{S\}$ 赋予每个节点 i 一个自旋标号 s_i ,这个自旋标号取整数 $s_i = 1, \dots, q$. 假设系统 S 的自旋值有 q 个不同的状态,那么该系统的 Potts 模型 Hamiltonian 函数 $H(S)$ 有以下形式

$$H(S) = \sum_{\langle ij \rangle} J_{ij} (1 - \delta_{s_i s_j}), \quad i, j = 1, \dots, N \quad (1)$$

其中的求和包括所有相邻的节点, 表示为 $\langle ij \rangle$, J_{ij} 表示节点 i 和节点 j 之间的交互力度. 当 $s_i = s_j$ 时, $\delta_{s_i s_j}$ 值为 1, 否则为 0. 这里 J_{ij} 被设为

$$J_{ij} = J_{ji} = \frac{1}{\langle k \rangle} \exp \left[-\frac{(d_{ij})^2}{2} \right], \quad i, j = 1, \dots, N \quad (2)$$

其中 $\langle k \rangle$ 表示节点的平均邻居节点数量, d_{ij} 是节点 i 和节点 j 之间的欧氏距离. 关联值 J_{ij} 是一个 d_{ij} 的单调递减函数, 而且如果我们最小化 $H(S)$, 自旋值 s_i 和 s_j 会按照相应 d_{ij} 的大小而渐次趋向于相同的数值.

为了指示和量化两个自旋值之间的关联性和紧密程度, 我们定义自旋关联 C_{ij} 为 $\delta_{s_i s_j}$ 的平均值

$$C_{ij} = \langle \delta_{s_i s_j} \rangle \quad (3)$$

它表示自旋变量 s_i 和 s_j 具有相同值的概率. C_{ij} 的值位于区间 $[0, 1]$ 之中, 代表着自旋变量 i 和 j 从没有交互到完全一致的连续变化情况. 当 J_{ij} 确定之后, 在同质系统中存在着两种不同的相位. 当温度比较高时, 系统处于顺磁相位而且自旋值处于混乱状态.

此时, 对于所有的节点 i 和 j , $C_{ij} \approx \frac{1}{q}$, q 是不同的自旋值的数目. 当温度比较低时, 系统转入铁磁相位, 所有的自旋子会趋近相同的自旋方向. 此时对于所有的节点对 (i, j) , $C_{ij} \approx 1$.

当系统不是同质的而是具有社团结构时, 自旋变量不仅仅只有顺磁相位或者铁磁相位. 我们假设温度会随着时间 t 的增长而降低, 它们的标度是一一对应的, 那么系统随着温度的降低到达全局稳定状态之前, 自旋变量会经历层次结构的局域一致状态(亚稳定状态), 如图 1 所示. 在每一个局域一致状态, 相同社团内节点的自旋值是相同的, 这时整个系统被划分为一些不同的自旋值局域一致区域(社团). 相应的, 我们可以计算出每个局域一致状态的开始和结束时刻, 用来分析这个状态的稳定程度. 当处于局域一致状态时, 节点的自旋值会处于稳定状态不会改变. 社团结构可以和局域一致状态关联起来, 对于一个具有强社团结构的网络, 每个社团非常明显, 社团内的节点很容易达到局域一致状态, 因此该状态的进入时间会很早. 同时, 由于社团之间关联很少, 这就意味着节点很难离开该局域一致状态, 因此该状态的结束时刻会很晚. 综上所述, 对于一个强社团结构的网络, 社团的局域一致状态的开始和结束时刻会相差很大.

一旦 J_{ij} 确定下来, 我们可以用 Swendsen-Wang

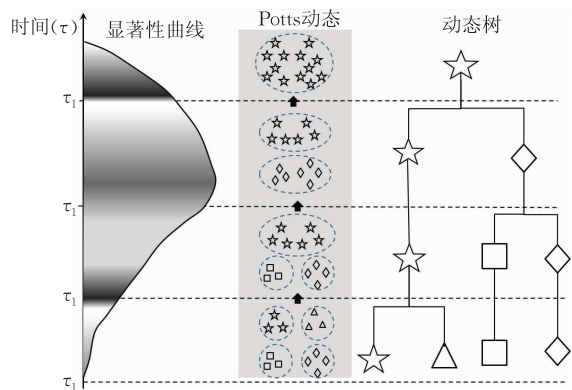


图 1 随着时间的增加, 拥有 4 个社团的自旋结构会经过一些局部一致状态, 最终达到全局稳定的状态. 不同自旋值区域用不同的形状描述. 在时间 τ_4 时 ($\tau_4 > \tau_3 > \tau_2 > \tau_1$, τ_i 表示系统中有 i 个不同自旋状态的开始时间), 我们将 4 个局部一致的自旋状态分布对应 4 个社团. 在时间 τ_3 时, 五角形和三角形混合在一起. 在 τ_2 时, 方形和菱形混合在一起, 这种混合显示在他们的结构层次上. 最后, 在 τ_1 时刻, 整个系统只有一个自旋状态, 所有的节点具有相同的自旋分布

(SW) 蒙特卡洛算法来计算 C_{ij} , 该算法的优势是它的自动关联时间比常规方法小很多. 对于一个具有 N 个节点的网络, SW 算法简述如下.

(1) 随机生成初始的自旋变量配置 $S_1 = (s_1, s_2, \dots, s_N)$, 其中 s_i 代表节点 i 的自旋值, 初始值从 $1 \sim q$ 之中随机选取, 不同的自旋值数目为 $q = N/2$.

(2) 基于系统配置 S_1 来产生配置 S_2 :

① 访问所有的节点对 $\langle i, j \rangle$. 节点对 $\langle i, j \rangle$ 具有关联值 $J_{ij} > 0$, 其中 J_{ij} 的计算只依赖于网络邻接矩阵, 节点 i 和节点 j 按照下面的概率冻结起来:

$$P_{ij}^f = 1 - \exp \left(-\frac{J_{ij}}{T} \sigma_{s_i s_j} \right) \quad (4)$$

其中如果 $s_i = s_j$ 那么 $\delta_{s_i s_j} = 1$, 否则为 0, T 代表温度. 然后计算所有的自旋变量对, 并对所有的变量对加入一个冷冻连接, 这里冷冻连接的作用是将关联节点之间状态保持固定而不会随着时间发生变化.

② 我们定义 SW 聚类为该聚类中包含的所有自旋变量之间都有冷冻连接关联它们. 因为节点之间的关联发生冻结当且仅当它们具有相同的自旋值, 所以我们只需要识别相同自旋值的节点就可以划分社团.

③ 对于每个 SW 社团, 我们从 $1, 2, \dots, q$ 中随机抽取一个数, 然后将这个数赋予该社团的所有点. 当所有的社团都被赋予数值之后, S_2 就生成了.

(3) 重复步(2). 最后我们根据式(3)计算 C_{ij} .

初始我们设可能的自旋值数目为 $q = N/2$ ($N/2$

是社团可能的最大数目,因为如果社团数目大于 $N/2$,有些社团将仅包含一个节点而不存在团内边). 对于一个特定的点,我们从 $1 \sim q$ 中随机挑选一个数并将它设为自旋值.

3 Markov 随机模型

Markov 过程是一个非常强大的工具,已经被科学家广泛地应用到网络分析中^[27-29]. 为了建立社团结构和 Potts 模型局域一致行为之间的关联,我们引入 Markov 随机模型用来计算网络的谱显著性. 设 $A=(V, E)$ 代表一个网络,其中 V 是节点的集合, E 是网络中边的集合. 考虑一个定义在网络 A 上的 Markov 随机过程,一个随机游走者每一步从一个节点沿着边随机移动到邻接节点. 假定 $X_t = X_i, t \geq 0$ 代表游走者的所在位置, $P\{X_t = i_t, 1 \leq i \leq N\}$ 代表经过 t 时刻之后,游走者到达节点 i 的概率. 对于 $i_t \in V$, 有 $P(X_t = i_t | X_0 = i_0, X_1 = i_1, \dots, X_{t-1} = i_{t-1}) = P(X_t = i_t | X_{t-1} = i_{t-1})$. 这意味着随机游走者的下一个状态只由它当前的状态决定. 因此,这个随机过程是一个定义在状态空间 V 上的离散 Markov 链. 而且,因为 $P(X_t = j | X_{t-1} = i) = p_{ij}$, 所以 X_t 是同质的,其中 p_{ij} 代表随机游走者从节点 i 移动到节点 j 的转移概率.

为了将 Markov 过程与 Potts 模型的动态机制进行关联,我们定义 p_{ij} 为

$$p_{ij} = \frac{C_{ij}}{\sum_{j=1}^N C_{ij}} \quad (5)$$

其中 C_{ij} 代表式(3)中的自旋变量关联. 通过这种定义,随机过程理论和有限状态 Markov 过程就可以应用到社团结构分析当中. 假设矩阵 $\mathbf{P}$ 为状态概率转移矩阵,我们可以定义

$$\mathbf{P} = \mathbf{D}^{-1} \mathbf{C} \quad (6)$$

其中矩阵 $\mathbf{D}$ 是矩阵 $\mathbf{C} = (C_{ij})$ 的对角度矩阵. 定义 $p_{ij}^{(t)}$ 为经过 t 时刻后,随机游走者节点 i 移动到节点 j 的转移概率,我们有

$$p_{ij}^{(t)} = (\mathbf{P}^t)_{ij} \quad (7)$$

对于这个遍历性的 Markov 过程, $\mathbf{P}^t$ 中的元素对应着状态之间经过 t 时刻后的转移概率. 为了计算转移矩阵 $\mathbf{P}^t$, 我们利用矩阵 $\mathbf{P}$ 的特征分解. 设定 $\mathbf{P}$ 的特征值为 $\lambda_k, k=1, \dots, N$, 它的右特征向量和左特征向量分别为 $\mathbf{u}_k$ 和 $\mathbf{f}_k$, 它们满足正交关系

$$\mathbf{u}_k \mathbf{f}_l^T = \delta_{kl} \quad (8)$$

矩阵 $\mathbf{P}$ 的谱的表示形式为

$$\mathbf{P} = \sum_k \lambda_k \mathbf{u}_k \mathbf{f}_k^T \quad (9)$$

由此可得

$$\mathbf{P}^t = \sum_k \lambda_k^t \mathbf{u}_k \mathbf{f}_k^T \quad (10)$$

我们假设矩阵 $\mathbf{P}$ 的特征值排列为 $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_N|$. 根据谱聚类理论^[30-31], $\mathbf{P}^t$ 可以表示为 N 项的加和

$$\mathbf{P}^t = \sum_{k=1}^N \lambda_k^t \frac{\mathbf{u}_k \mathbf{u}_k^T \mathbf{D}}{\mathbf{u}_k^T \mathbf{D} \mathbf{u}_k} \quad (11)$$

从式(11)可以看出,每一个矩阵 $\mathbf{P}^t$ 的具体形式都只依赖于 $\mathbf{P}$ 的特征系统. 因为 $\mathbf{P}$ 是对称关联矩阵 $\mathbf{C}$ 标准化之后的结果,所以我们可以得到 $\mathbf{u}_n^T \mathbf{D} \mathbf{u}_m = I_{nm}$. 由于矩阵 $\mathbf{P}$ 的最大特征值 $\lambda_1 = 1$, 所以当时间 $t \rightarrow \infty$ 时, 有 $\mathbf{P}^{(t)} = \mathbf{P}^\infty = \frac{\mathbf{u}_1 \mathbf{u}_1^T \mathbf{D}}{\mathbf{u}_1^T \mathbf{D} \mathbf{u}_1}$. 每一个 Markov 随机过程的初始状态都会收敛到静态分布 $\mathbf{P}^{(0)}$, 它对应着整个系统的自旋变量最终会随着温度的下降收敛到完全相同的数值. 当时刻 $t \rightarrow \infty$ 时,除了最大的特征值 $\lambda_0^t = 1$, 其他所有的特征值 λ_k^t 会趋向于 0. 另一个极端情况下,当时刻 $t=0$ 时, $\mathbf{P}^t$ 会变成静态分布矩阵. 该矩阵所有的列都是各不相同的,此时整个系统分裂为尽可能多的不同自旋值.

虽然矩阵 $\mathbf{P}^t$ 能很有效的关联 Potts 模型的动态过程,但是它需要预处理到对称的情况. 我们简单地将 $\mathbf{P}^t$ 转换为对称矩阵 $\mathbf{G}^{(t)} = (\mathbf{P}^t + (\mathbf{P}^t)^T)/2$ 的形式,它可以定义为 t 时刻的自旋关联矩阵. 矩阵 $\mathbf{G}^{(t)}$ 和矩阵 $\mathbf{P}^t$ 的特征系统之间的关系有下面的引理.

引理 1. 矩阵 $\mathbf{G}^{(t)}$ 和矩阵 $\mathbf{P}^t$ 的特征值和相应的特征向量完全相同.

引理 1 的证明是显然的. 根据引理 1, 因为矩阵 $\mathbf{G}^{(t)}$ 拥有和矩阵 $\mathbf{P}^t$ 完全相同的特征系统,我们可以用 $\mathbf{G}^{(t)}$ 来发掘 Potts 状态的动态特征. 同样,很多算法也可以用 $\mathbf{G}^{(t)}$ 去寻找最合理的划分,比如 K-means 算法^[30]和 GN(Giroan-Newman)算法^[32].

4 Potts 模型中社团多时间尺度稳定性分析

在本节,我们根据上面分析的 Potts 模型和 Markov 随机过程的关联揭示多时刻的社团结构稳定性,很多有用的信息,比如最优的社团个数和任意时刻的网络结构稳定性等都能够有效地表示出来.

假设我们可以将网络 A 划分成 K 个不相交的

社团或者群 $V_k \subset V, k \in 1, 2, \dots, K$, 这些社团 $V_1, V_2, \dots, V_K$ 一起构成了点集合 V 的一个划分. 每个社团中节点的数目可以表示为 $N_k = |V_k|$. 从数值上, 我们把社团结构的动态过程模拟为自旋系统的变化过程, 并将时间因素考虑在内. 因此, 我们定义一个特定社团 V_k 的显著性为内部关联程度的比例

$$S_k^{(t)} = \sum_{i,j \in V_k} \frac{[\mathbf{G}^{(t)}]_{i,j}}{N_k} \quad (12)$$

$S_k^{(t)}$ 可以看作时间标度 t 的一个方程, 我们可以利用它来研究社团结构随时间的变化趋势. 假定社团的个数为 K , 我们可以通过最大化下面这个目标函数来找到合适的社团结构

$$J_K^{(t)} = \sum_{k=1}^K S_k^{(t)} = \sum_{k=1}^K \sum_{i,j \in V_k} \frac{[\mathbf{G}^{(t)}]_{i,j}}{N_k} \quad (13)$$

该目标函数可以解释为每个社团 V_k 的显著性 S_k 的加和. 式 $J_K^{(t)}$ 可以关联到一些常见的划分函数, 比如它就是对切边准则 (Ratio cut criterion) 的一种扩展. 因为目标函数 $J_K^{(t)}$ 定义为团间边数目的总和除以总的边数, 这里我们将切边准则中的邻接矩阵 $\mathbf{A}$ 替换为自旋关联矩阵 $\mathbf{G}^{(t)}$. 而且我们也可以发现 $J_K^{(t)}$ 实际也是模块度 Q 第 1 部分的扩展.

下一步的论述将通过目标函数的矩阵形式来展开. 我们将节点 i 的社团归属用向量 $\mathbf{x}_i = (x_i^k)$ 表示, 它代表节点 i 归属于某一特定社团的概率, 元素 x_i^k 表示点 i 的归属向量的第 k 个元素. 硬划分和非相交的集合 V_k 要求向量 $\mathbf{x}_i$ 和 $\mathbf{x}_j$ 是正交的. 目标函数 $J_K^{(t)}$ 可以根据向量 $\mathbf{x}_k$ 重写为

$$J_K^{(t)} = \sum_{k=1}^K \frac{\mathbf{x}_k^T \mathbf{G}^{(t)} \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{x}_k} \quad (14)$$

最大化该目标函数有下面的约束条件: $\mathbf{x}_k \in \{0, 1\}$ 和如果 $i \neq j$ 有 $\mathbf{x}_i^T \mathbf{x}_j = 0$. 通过将向量 $\mathbf{x}_k$ 集中到矩阵 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K)$, 式 (14) 能够重写成矩阵的形式

$$J_K^{(t)} = \text{tr} \{ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{G}^{(t)} \mathbf{X} \} \\ = \text{tr} \{ (\mathbf{X}^T \mathbf{X})^{-1/2} \mathbf{X}^T \mathbf{G}^{(t)} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1/2} \} \quad (15)$$

其中矩阵 $\mathbf{X}^T \mathbf{X}$ 是正交的. 接下来利用 $\mathbf{Y} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1/2}$ 我们可以将优化问题简化为 $J_K^{(t)} = \text{tr} \{ \mathbf{Y}^T \mathbf{G}^{(t)} \mathbf{Y} \}$. 条件 $\mathbf{Y}^T \mathbf{Y} = \mathbf{I}_K$ 是自动满足的, 这是因为

$$\mathbf{Y}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1/2} (\mathbf{X}^T \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1/2} = \mathbf{I}_K \quad (16)$$

向量 $\mathbf{y}_K$ 因此具有单位长度而且满足正交性. 这个优化问题可以根据矩阵 $\mathbf{Y}$ 重写为

$$\max_{\mathbf{Y}^T \mathbf{Y} = \mathbf{I}} \text{tr} \{ \mathbf{Y}^T \mathbf{G}^{(t)} \mathbf{Y} \} \quad (17)$$

定理 1. Rayleigh-Ritz 定理. 假设 $\mathbf{L}$ 是一个对称的 $N \times N$ 矩阵, 它拥有特征值 λ_k 和对应的特征

向量 $\mathbf{u}_k, k=1, \dots, N$. 那么问题

$$\max \sum_{k=1}^K \mathbf{y}_k^T \mathbf{L} \mathbf{y}_k, \text{ s. t. } \mathbf{y}_l^T \mathbf{y}_k = \mathbf{I} \quad (18)$$

的最优解 $\text{opt}\{\mathbf{y}\}$ 为 $\mathbf{u}_1, \dots, \mathbf{u}_K$, 进而 $\max \sum_{k=1}^K \mathbf{y}_k^T \mathbf{L} \mathbf{y}_k$ 等于 $\sum_{k=1}^K \lambda_k$.

Rayleigh-Ritz 定理^[31] 告诉我们式 (17) 中的最大化问题可以通过将向量 $\mathbf{Y}$ 的列元素对应到矩阵 $\mathbf{G}^{(t)}$ 的前 K 个最大的特征向量来求解. 我们将矩阵 $\mathbf{P}$ 的特征向量进行排序 $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_N|$, 而 λ_k 对应的特征向量表示为 $\mathbf{u}_k$. 因此式 (17) 所代表的优化问题的结果为特征向量矩阵 $\mathbf{Y} = \mathbf{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_K\}$. 社团的显著性就对应于相应特征值的 t 次方

$$S_k^{(t)} = \frac{\mathbf{u}_k^T \mathbf{G}^{(t)} \mathbf{u}_k}{\mathbf{u}_k^T \mathbf{u}_k} = \lambda_k^t \frac{\mathbf{u}_k^T \mathbf{u}_k}{\mathbf{u}_k^T \mathbf{u}_k} = \lambda_k^t \quad (19)$$

经过多个时间尺度后自旋配置将收敛到完全一致的状态, 随时间增长而消失的小特征值阐述了不同自旋态的消失和相对较弱的社团结构的同化, 而这些社团特征对应于不同的特征向量. 出于社团划分的目的, 我们对中间时间尺度的局部一致状态更为感兴趣. 如果我们想识别出 z 个社团, 我们期望着在一个特定时刻 t , 矩阵 $\mathbf{P}^t$ 的特征值 λ_k^t 只在区间 $k=1, \dots, z$ 内显著大于 0, 这可以通过确定时刻 t , 使得 $|\lambda_k^t|^t \approx 0$.

从另一个角度, 因为对特征值进行排序为 $\lambda_1 = 1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_N|$, 那么 t 时刻的社团强度, λ_k^t 可以看作 k 自旋状态的鲁棒性程度. 由此, 特征差值 $\lambda_k^t - \lambda_{k+1}^t$ 能够表示 t 时刻从 $(k+1)$ 自旋状态转移到 k 自旋状态的“难度”. 给定关联矩阵 $\mathbf{G}$, 我们可以通过计算最大的特征差值 $\lambda_k^t - \lambda_{k+1}^t$, 得出特定时刻 t 最适合的自旋变量的可能个数.

t 时刻最优的社团个数 Λ 可以通过最大的特征值差来推断出, 用来衡量最稳定的状态. $\Lambda(t)$ 可以正式定义为

$$\Lambda(t) = \arg[\max_k (\lambda_k^t - \lambda_{k+1}^t)] \quad (20)$$

从全局的角度来看, 如果社团个数 Λ 在最长的时间内没有改变, 我们可以把它当做最优的社团个数, 表示为 Ψ .

社团个数 Λ 可能会保持一段时间不会改变, 然而隐藏在我们模型中的自旋配置的动态变化仍不清楚. 为了揭示自旋动态的变化细节, 我们需要确定自旋配置所代表的社团结构在什么时间尺度是最稳定

的.在一定程度上,最稳定的状态可以代表整个网络的最优社团结构.因此,我们定义每一个时间刻度上社团结构的稳定性 $\theta(t)$,作为最稳定的自旋状态的稳定性:

$$\theta(t) = \lambda'_{\Delta(t)} - \lambda'_{\Delta(t)+1}$$

(21)

我们希望从 $\theta(t)$ 的变化趋势,可以发现最稳定的社团结构所达到的时间刻度.在这个时候,自旋配置的分布代表着最合理的社团结构.此外,从整体的角度来看,我们可以使用 $\Delta(t)=q$ 社团对应的最大稳定性, $\Gamma(q)=\max\{\theta(t) \mid \Delta(t)=q\}$,来表示一个网络的鲁棒性,定义为拥有 q 个社团的网络结构的稳定性.因为 $\Gamma(q)$ 能直接描述和评价网络的结构特征而不需要具体的网络划分,用它来评价网络的模块化性质是很方便的.

为了验证该模型可以揭示不同规模的网络层次结构^[18-19],图 2 和图 3 给出了多层次社团结构的两个例子.图 2(a)展示了 RB125 网络,这是一个多层次的无标度网络,它首先由 Ravasz 和 Barabasi 给出^[33].从图中可以看出,对应于 5 个和 25 个模块的区域划分在分辨率方面是最有代表性的.其次,图 3(a)展示了由 Arenas 等人^[34]提出的 H13-4 网络,这是 1 个度均匀的网络,有两个预定义层次网络标度.该网络的第 1 个层次水平由 4 个拥有 64 个节点的模块组成,第 2 层由 16 个拥有 16 个节点的模块组成.两个层次水平上的划分都在原始网络上高亮显示.

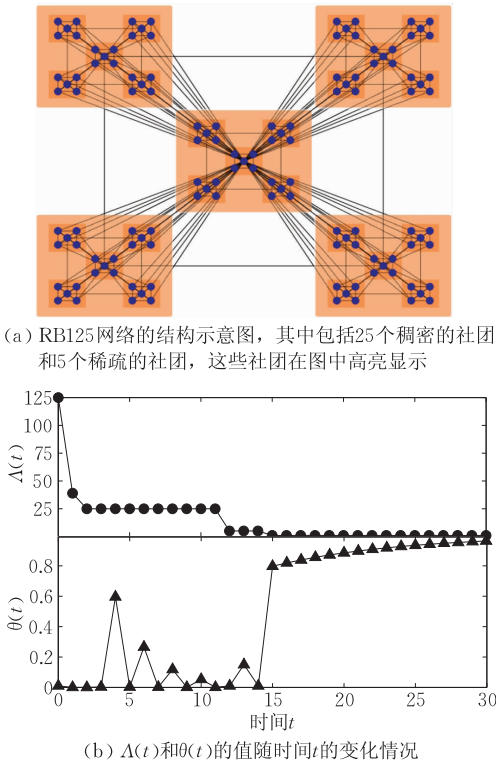


图 2 RB125 网络

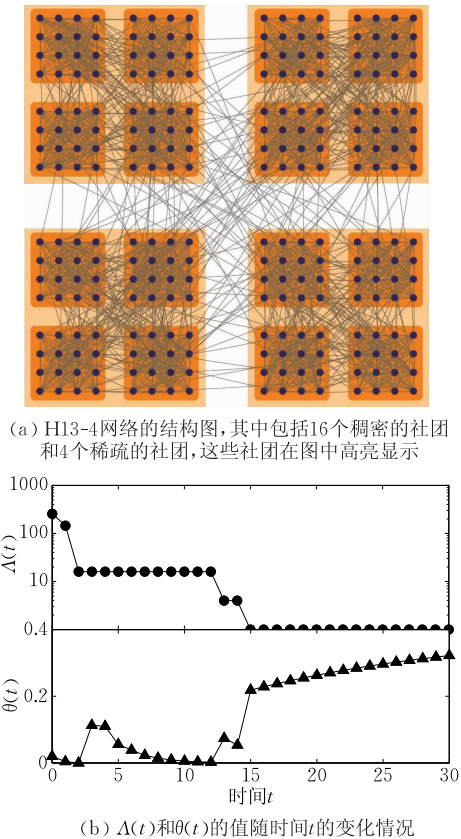


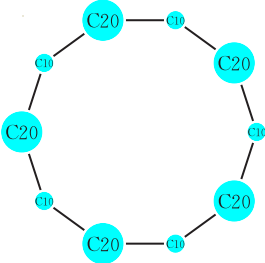
图 3 H13-4 网络

在这两个例子中,持续时间最长的 $\Delta(t)$ 揭示了隐藏在网络中的实际的层次结构.这些网络层次的显著性,可以通过相应的持续时间长度来衡量.时间持续越久,该层次的显著性越强.从图 2(b)和图 3(b),我们可以观察到 25 和 16 是在 RB125 和 H13-4 网络中是最优的社团个数,它们的持续时间分别也是最长的.然而,5 个模块的划分和 4 个模块的划分也是合理的分区,它展示了模糊等级的网络层次.这些结果与这两个网络的生成机制和层次模式完全一致.

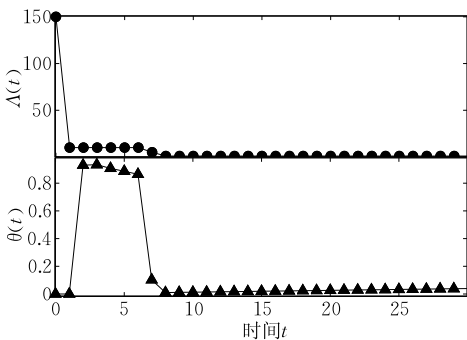
此外,我们还展示出,这两个例子的稳定性的变化趋势可以揭示自旋配置的动态变化.从图 2(b)和图 3(b)可以看出,特定 Δ 相应的稳定性 $\theta(t)$ 不是抛物线形状.因此,我们不能轻易地找到全局最优解.然而,一些局域稳定性的最大值也能够代表比较好的社团结构.我们可以找到这些稳定性的局域最大的时间尺度 t ,并将 $\mathbf{G}^{(t)}$ 应用到特定的社团划分方法中.此外,稳定性在所有特定的 Δ 结束时将达到最低值,而当转移过渡到新的状态时,稳定性开始增加.我们可以使用 $\theta(t)$ 来评估复杂网络的模块化性质,较大的 $\theta(t)$ 意味着更显著的网络社团结构.因此,人们可以找到最大值 $\theta(t)$ 对应的社团数目,用它来指示社团结构的鲁棒性.图 3(b)展示了 H13-4 网络,当 $t=4$ 时数目为 16 的社团结构稳定性 $\Gamma(16)=0.48$,

大于当 $t=7$ 时 $\Gamma(4)=0.31$. 这表明, 该网络包含 16 个社团要比包含 4 个社团更稳定. 同样, 对于在图 2 (b) 中展示的 RB125 网络, 当 $t=25$ 时 25 个社团对应的 $\Gamma(25)=0.48$ 要大于当 $t=7$ 时对应的 $\Gamma(5)=0.31$. 社团结构的鲁棒性 $\Gamma(q)$ 倾向于小而明显的社团单元, 这与 Arenas 等人^[2] 提出的结果是相同的, 在许多实际网络中也非常合理.

最后, 我们强调稳定性指标和 Newman 提出的模块度 Q 之间的差异. 模块度 Q 是一个非常著名的指标, 用来评价一个网络的一个特定的划分方案. 它定义为“社团内边的比例, 减去如果边随意分布而不考虑社团结构时预期的相同的量值, 即随机化的社团内部边比例”. 相同网络的不同的划分方案将得到不同的 Q 值, 而大的 Q 值意味着更好的划分. 我们提出的 Δ 和 Γ 能直接描述和评估基于谱性质的网络结构, 而不是一个具体的网络划分. 因此, 一个特定网络只有一个自我确定的 Δ 和 Γ 值, 不管有多少划分方案. 拥有较大的 Δ 和 Γ 值的网络具有更强的社团结构. 此外, Fortunato 等人指出模块度 Q 存在分辨率限制问题, 即小的社团存在一个内在尺度, 小于这个尺度就不能通过最大化模块度 Q 检测出. 然而, 如图 4 所示, 当一个团环包含不同尺度的团 (即异质社团的规模), 内在的社团结构可以用 Δ 准确的揭示出来. 因此, 利用 Δ 和 Γ , 我们可以定量地比较模块化结构不同的复杂网络.



(a) 一个团环网络的示意图. 每个团用圆形来表示, 团的规模用 C20 (包含 20 个节点) 或者 C10 (包含 10 个节点) 来表示



(b) $\Delta(t)$ 和 $\theta(t)$ 的值随时间 t 的变化情况

图 4 团环网络

5 一种新型的社团划分算法

为了能够执行社团划分, 我们提出一个基于关联矩阵 $\mathbf{G}^{(t)}$ 的特征值分解的算法. 该算法使我们能够识别多尺度的网络社团结构. 基于上面几节的分析, 我们将相关的多尺度社团结构^[35] 与特征值和特征向量的动态变化结合起来.

实数对称矩阵 $\mathbf{G}^{(t)}$ 的特征值 λ_k 和特征向量 $\mathbf{u}_k$ 可以通过求解下列特征方程来得到

$$\mathbf{G}^{(t)} \cdot \mathbf{u}_k = \lambda_k \cdot \mathbf{u}_k, \quad k = 1, \dots, N \quad (22)$$

当时间 t 非常小的时候, 该方程具有 N 种不同的结果. 假设特征向量经过了归一化 ($\sum_i \mathbf{u}_k(i) = 1$). 每个显著性 $S_k(t) = \lambda_k^t$ 都与一个特定的社团相互关联 (该向量中的元素有相同的自旋值). 对于特定的社团 k , 它的内部结构与特征向量 $\mathbf{u}_k$ 相互对应.

特征向量归一化后 ($\sum_i \mathbf{u}_k(i) = 1$), 它的元素通过 $\mathbf{u}_k^2(i)$ 可以衡量每个点 i 相对于社团 k 的归属关系. 结合社团结构的稳定性和向量 $\mathbf{u}_k^2(i)$, 点 i 在 t 时刻对社团 k 的“绝对”归属关系可以通过下列的归属向量来表示

$$\mathbf{R}_k^{(t)}(i) = \lambda_k^t \mathbf{u}_k^2(i) \quad (23)$$

如果 $\mathbf{R}_k^{(t)}(i)$ 最大, 则节点 i 属于社团 k .

根据式 (23), 我们可以观察归属矩阵随着时间推移的演化规律. 当时间尺度 $t \rightarrow \infty$, 除了最大特征值 $\lambda_0 = 1$ 之外所有的特征值 λ_k 都趋近于 0. 此时, 根据归属矩阵的定义, 所有的节点属于同一个社团. 在另一个极端, 当 $t=0$ 时, 归属矩阵 $\mathbf{R}$ 却成为特征向量矩阵 $\mathbf{U}^2$. 所有的列向量是不同的, 而且社团的数目等于矩阵的维数. 在这里, 我们有兴趣在中间的时间尺度中寻找最大的稳定性指标 ($\theta(t)$), 此时自旋配置代表最稳定的社团结构.

因此, 我们首先在时间间隔中利用 $\Delta(t)$ 确定最优的社团个数. 然后, 我们挑选的时间刻度 t , 使得该时刻稳定性 $\theta(t)$ 达到最大. 在许多实际网络中, 社团的组成是一种硬划分, 每个节点只属于一个社团. 通常这种情况限制太大, 因为位于社团边界的点会分享不同社团的一些共性, 发挥网络中的扩散和过渡作用. 在我们的工作中, 归属矩阵 $\mathbf{R}$ 将划分扩展到概率的形式, 从而将社团划分扩展到模糊划分的范畴, 其中每个节点可能同时属于不同的社团, 这更加符合现实世界的特征.

最后, 我们计算最稳定的时间 t 时的归属矩阵

R. 该框架整个过程归纳到算法 1. 在该算法的过程中, 基于 SW 算法计算自旋关联矩阵 C 的时间复杂性低于 $O(N^2)$. 我们很容易估算出计算耗费主要在稀疏矩阵 G 的特征系统计算上面, 它约为 $O(N^2)$. 算法中的其他步骤是一些简单的矩阵计算. 最后, 我们得到的算法 1 的整体计算复杂度是 $O(N^2)$. 我们的算法是一个与参数无关的方法, 在实际网络中很容易实现并应用.

算法 1. 社团划分算法.

输入: 网络的邻接矩阵 A

输出: 归属矩阵 R

- 1. 计算自旋关联矩阵 C .
- 2. 基于 C , 计算 Markov 转移概率矩阵 P 和 G .
- 3. 计算矩阵 G 的特征值和对应的特征向量.
- 4. 计算最优的社团个数 Ψ 和最大稳定性 $\Gamma(\Psi)$ 对应的时刻 t .
- 5. 根据式 (23), 计算归属矩阵 R .
- 6. 返回: 输出归属矩阵 R .

6 实 验

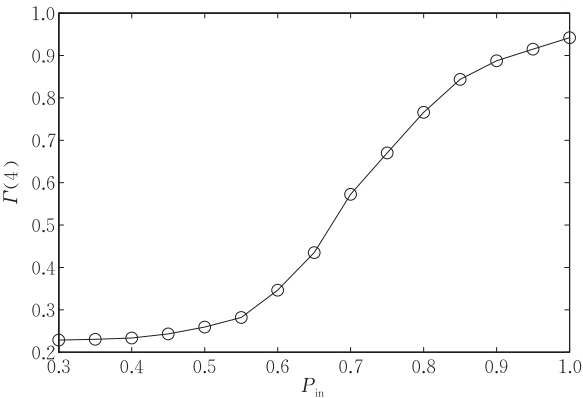
在本节中, 我们将测试所提出算法的性能. 我们设计并实现了 3 个实验, 主要有两个目的: (1) 评估该算法的划分精度; (2) 应用到实际的大型网络.

6.1 人工网络

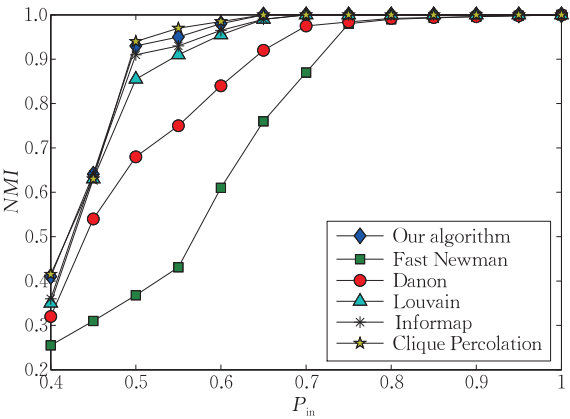
为了验证算法的有效性, 我们在人工网络上与其他 5 个著名的算法进行比较. 这些算法包括: Fast Newman 算法^[32]、Danon 算法^[36]、Louvain 算法^[37]、Infomap 算法^[18] 和 Clique Percolation 算法^[38]. 我们广泛使用的 Ad-Hoc 网络模型^[1], 它又叫 GN 人工网络, 是由 Givan 和 Newman 首先构建的一个随机合成网络. 整个网络共 128 个节点, 分属于 4 个预定义的社团, 每个有 32 个节点. 节点的平均度为 16, 社团内的边的比例表示为 P_{in} . 随着 P_{in} 下降, Ad-Hoc 网络的社团结构变得越来越模糊, 相应地, 它们的 $\Gamma(4)$ 会从 0 上升到 1, 如在图 5(a) 所示.

我们用 NMI(Normalized Mutual Information) 指标^[35]来评估每个算法产生的社团划分的优劣, 来验证网络社团的内在尺度能否被正确地发现. 实验结果展示在图 5(b) 中, 其中 y 轴代表 NMI 的值, 曲线上各点都是 50 个实验结构的平均. 我们可以看到, 所有的算法在 P_{in} 大于 0.7 都有优越的性能, NMI 值都大于 0.85. 通过和其他 5 个算法进行比较, 我们算法的表现是最好的, 其精度仅稍比 Clique Percolation 算法在 $0.5 \leq P_{in} \leq 0.65$ 时差一点. 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(Breadth First Search) 算法. 相比之下, 我们算法的时间复杂性是非常低的 ($O(N^2)$), 可以容易地在各种网络中实现.



(a) $\Gamma(4)$ 随 P_{in} 的变化情况



(b) 6 个算法的精确性对比

图 5 人工网络实验结果

6.2 美国大学足球队网络

由于具有非常清晰自然社团结构, 美国大学足球队网络作为一个基准的例子已经得到了广泛的应用. 该网络的数据是由 Girvan 和 Newman^[32] 收集的. 这个网络代表了 2000 赛季美国第一等级美式足球比赛的赛程, 网络中的节点代表 115 支球队, 而边代表一年中 613 场比赛关系. 整个网络可以很自然地分为 12 个不同的组. 作为比赛, 相对于不同的组成员之间, 比赛一般都更频繁地发生在同一组的成员间.

首先, 我们计算 Δ 和相应的稳定性 $\theta(t)$, 结果如图 6 所示. 结果表明, 美国大学足球队网络的社团最优数目为 $\Delta=12$, 与真实情况吻合. 稳定性 $\theta(t)$ 在 $t=4$ 达到 $\Gamma(12)=0.31$. 然后, 我们把算法应用到网络, 将网络划分为 12 个社团, 结果如图 7 所示. 算法划分的正确率超过 93%, 这意味着划分出的社团结构与真实的网络具有的结构非常吻合. 事实上, 基于优化模块化 Q 的方法通常只会找到 11 个社团, 而

且由于网络的模糊性,划分的正确率也比较低. 我们的算法能够揭示网络的实际特性,对许多现实网络是非常有价值的. 此外,我们的算法识别了 5 个有趣的重叠节点,它们分别是图 7 中的 Navy、Notre Dame、Gentral Florida、Utah State 和 Connecticut. 这些节点都位于社团之间的边界上,可以被看作是一些相对独立的俱乐部,用肉眼可以很轻易地识别出来.

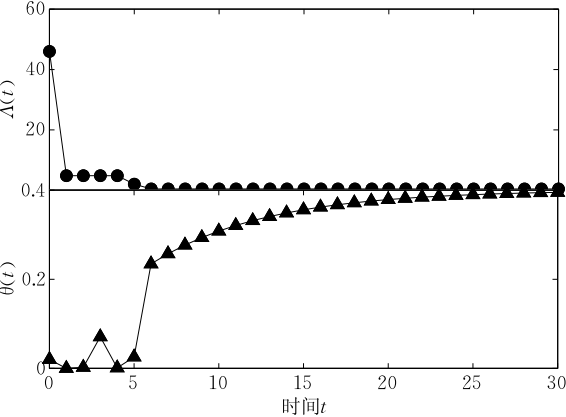


图 6 美国大学足球队网络中 $\Delta(t)$ 和 $\theta(t)$ 的计算结果

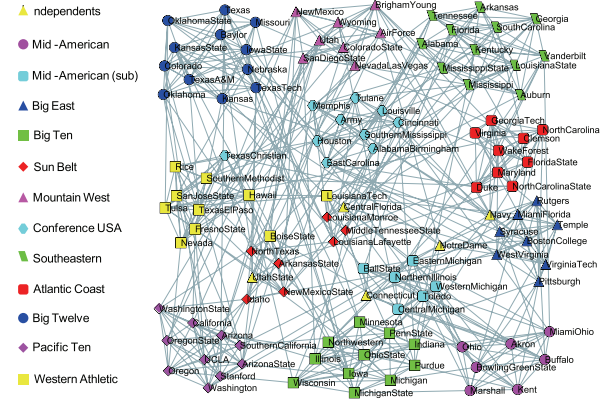


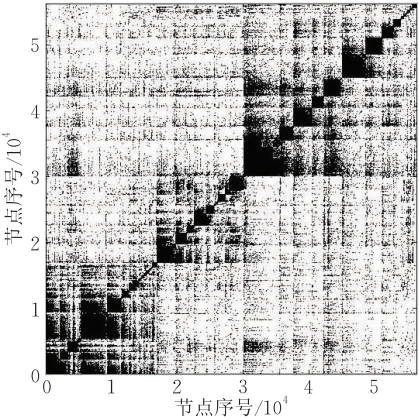
图 7 美国大学足球队网络的划分结果(同一个社团内的节点被赋予相同的形状)

6.3 科学家合作网络

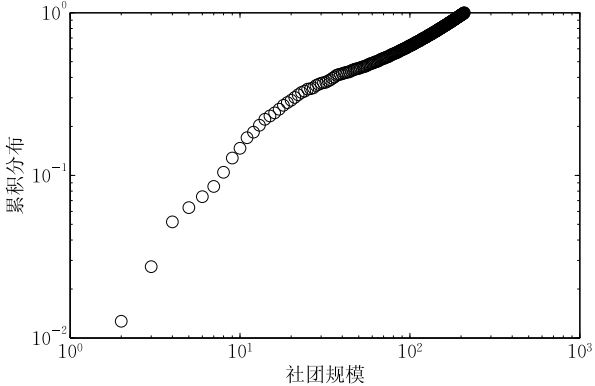
最后我们在大规模的科学家合作网上测试算法的性能,该网络是由 Girvan 和 Newman^[39] 收集的,它根据合著发表在 arxiv.org 的论文描述了 56 276 位物理学家之间的合作关系. 从整体上来看,这个网络包含 315 810 个加权边. 为了可视化,算法输出了一个转换后的具有层次结构的邻接矩阵(即同一个社团的节点组合在一起). 从图 8(a)展示 的转换后的邻接矩阵可以看出,科学家网络具有非常强的社团特征. 在这些社团中,3 个最大的分别代表 3 个主要研究领域:凝聚态物质、高能物理(包括理论、现象学和核物理学)和天体物理学.

我们将社团规模的分布以幂率图的形式在图 8

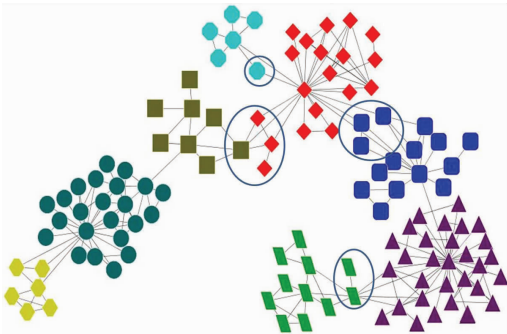
(b)给出,可以看出该分布是一种典型的无标度分布,这和很多实际网络是一致的. 从总体来说,我们通过优化社团稳定性一共探测出 737 个社团,最大的社团规模为 195,最小的为 2,而社团的平均规模是 76. 在这些社团中,1433 个社团对之间具有模糊的重叠节点. 5%的最大的社团包含大约 25.4%的节点,而其他社团规模都比较小.



(a) 科学家合作网的转换后的邻接矩阵



(b) 社团规模的分布图



(c) 包括8个社团的子图,每个社团用不同的形状来表示(10个重叠节点被四边形封闭,用来显示社团的边界)

图 8 科学家合作网络实验结果

3 个最大社团符合实际的物理研究领域,最大的是固体物理学,第 2 大的是超核物理,第 3 大的是理论天体物理学. 此外,我们从中截取了 1 个包括

8 个社团的子网络,在图 8(c)中显示. 我们利用归属矩阵找出该子网络包括 10 个重叠的节点,用圆形圈了起来. 该划分结果与文献[40-41]完全相同. 我们通过在大型网络上运行算法,其高效的性能表明,我们的方法非常有效而且对进一步研究具有非常大的价值.

由于模块度函数 Q 是衡量社团划分优劣的最著名的标准,因此我们将 3 种著名的基于模块度优化的算法——GN 算法^[32]、Fast Newman 算法^[1]和 Danon 算法^[36]应用于科学家合作网,并分别计算划分的模块度值. 为了展示性能,我们将 3 种算法的 Q 值和我们的算法进行比较,具体数值展示在表 1 中. 结果显示,我们算法的准确性是最高的,由此我们可以看出,我们提出的算法具有模块度方法所没有的优越性能.

表 1 4 种算法模块度 Q 值的比较

算法	GN 算法	Fast Newman 算法	Danon 算法	本文的算法
Q 值	0.38	0.43	0.49	0.55

7 结 论

在本文中,我们提出了一个新颖的基于多尺度稳定性的社团结构分析框架,该框架能够发现网络的社团结构和 Potts 模型局部一致状态谱特性之间的关联. 很多重要的社团结构信息都可以通过 Markov 过程和谱显著性计算得出,如社团结构的稳定性和社团的最佳数量. 在理论分析的基础上,我们还开发了一个能够检测模糊社团结构的算法,通过在模拟人工网络 and 实际大型网络上的运行,我们框架的有效性得到了理论证明和实验验证.

本文在方法论方面的重要创新点是将 Potts 模型和 Markov 随机过程结合起来解决社团问题,这是因为两种模型的共同的作用都是描述个体的动态演化. 我们利用 Potts 模型中的 $spin$ 值去描述节点的社团归属,然后将这种归属的动态演化用 Markov 模型的数学理论去推导和计算,具体是用 Markov 转移概率去模拟自旋值的变化趋势. 我们也可以反过来用特定形式 Hamiltonian 函数去衡量随机游走者的活动度量,这也是一个非常新颖的课题.

致 谢 中国科学院数学与系统科学研究院的章祥荪研究员、王勇副研究员对本文提出了很多有益的建议,中央财经大学管理科学与工程学院的贾传

亮副教授对本文的修改提供了帮助,在此一并表示感谢. 最后,特别感谢评审老师,他们细致耐心地审查了稿件!

参 考 文 献

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

[2] Arenas A, Diaz-Guilera A, Perez-Vicente C. J. Synchronization reveals topological scales in complex networks. *Physical Review Letters*, 2006, 96(11): 114102

[3] Pikovsky A, Rosenblum M, Kurths J. *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, 2003

[4] Meila M, Shi J. A random walks view of spectral segmentation // *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics*. San Francisco, USA, 2001: 166-169

[5] Ronhovde P, Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E*, 2009, 80(1):016109

[6] Son S W, Jeong H, Noh J D. Random field Ising model and community structure in complex networks. *The European Physical Journal B: Condensed Matter and Complex Systems*, 2006, 50(3): 431-437

[7] Hughes B D. *Random Walks and Random Environments*. Oxford: Clarendon Press, 1996

[8] Prem K G, David M B. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 2013, 110(36): 14534-14539

[9] Sun Peng-Gang, Gao Lin, Yang Yang. Maximizing modularity intensity for community partition and evolution. *Information Sciences*, 2013, 236(1): 83-92

[10] Li Hui-Jia, Xu Bing-Ying, Zheng Liang, Yan Jia. Integrating attributes of nodes solves the community structure partition effectively. *Modern Physics Letters B*, 2014, 28(5): 1450037

[11] Yan Qiang, Wu Lian-Ren, Zheng Lan. Social network based microblog user behavior analysis. *Physica A*, 2013, 392(7): 1712-1723

[12] Weng Li-Lian, Menczer F, Ahn Y Y. Virality prediction and community structure in social networks. *Scientific Reports*, 2013, 3(1): 2522

[13] Traag V A, Krings G, Van Dooren P. Significant scales in community structure. *Scientific Reports*, 2013, 3(1): 2930

[14] Li Hui-Jia, Zhang Xiang-Sun. Analysis of stability of community structure across multiple hierarchical levels. *Europhysics Letters*, 2013, 103(5): 58002

[15] Zhou Hai-Jun. Distance, dissimilarity index, and network community structure. *Physical Review E*, 2003, 67 (6): 061901

[16] Pons P, Latapy M. Computing communities in large networks using random walks. *Lecture Notes in Computer Science*, 2005, 3733: 284-293

- [17] Hu Yan-Qing, Li Meng-Hui, Zhang Peng, et al. Community detection by signaling on complex networks. *Physical Review E*, 2008, 78(1): 016115
- [18] Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, 105(4): 1118-1123
- [19] Delvenne J C, Yaliraki S N, Barahona M. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America*, 2010, 107(29): 12755-12760
- [20] Blatt M, Wiseman S, Domany E. Superparamagnetic clustering of data. *Physical Review Letters*, 1996, 76(8): 3251-3255
- [21] Wang S, Swendsen R S. Cluster Monte Carlo algorithms. *Physica (Amsterdam)*, 1990, 167(3): 565
- [22] Wiseman S, Blatt M, Domany E. Super-paramagnetic clustering of data. *Physical Review E*, 1998, 57(4): 3767-3783
- [23] Agrawal H, Domany E. Potts ferromagnets on coexpressed gene networks: Identifying maximally stable partitions. *Physical Review Letters*, 2003, 90(5): 158102
- [24] Ott T, Kern A, Steeb W, Stoop R. Sequential clustering: Tracking down the most natural clusters. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, 11(2005): 110145
- [25] Berthier L, et al. Direct experimental evidence of a growing length scale accompanying the glass transition. *Science*, 2005, 310(5755): 1797-1800
- [26] Reichardt J, Bornholdt S. Detecting fuzzy communities in complex networks with a Potts model. *Physical Review Letters*, 2004, 93(21): 218701
- [27] E Wei-Nan, Li Tie-Jun, Vanden-Eijnden E. Optimal partition and effective dynamics of complex networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, 105(23): 7907-7912
- [28] Zhou Hai-Jun. Network landscape from a Brownian particle's perspective. *Physical Review E*, 2003, 67(4): 041908
- [29] Von Luxburg U. A tutorial on spectral clustering. *Statistics and Computing*, 2007, 17(4): 395-416
- [30] Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(8): 8888
- [31] Yang Bo, Liu Ji-Ming, Feng Jian-Feng. On the spectral characterization and scalable mining of network communities. *IEEE Transactions on Knowledge and Data Engineering*, 2012, 24(2): 326-337
- [32] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, 2004, 69(2): 026113
- [33] Ravasz E, Barabasi A L. Hierarchical organization in complex networks. *Physical Review E*, 2003, 67(2): 026112
- [34] Arenas A, Fernandez A, Gomez S. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 2008, 10(5): 053039
- [35] Li Hui-Jia, Zhang Jun-Hua, Liu Zhi-Ping, et al. Identifying overlapping communities in social networks using multi-scale local information expansion. *European Physical Journal B*, 2012, 85(6): 109
- [36] Danon L, Duch J, Guileria D, Arenas A. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005, 09(2005): P09008
- [37] Blondel V D, Guillaume J L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 10(2008): P10008
- [38] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks innatureandsociety. *Nature*, 2005, 435(7043): 814-818
- [39] Newman M E J, Leicht E. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, 104(23): 9564
- [40] Yang Bo, Liu Ji-Ming, Liu Da-You. Characterizing and extracting multiplex patterns in complex networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 2012, 42(2): 469-481
- [41] Li Zhen-Ping, Zhang Shi-Hua, Wang Rui-Sheng, et al. Quantitative function for community detection. *Physical Review E*, 2008, 77(1): 036109



LI Hui-Jia, born in 1985, Ph. D., assistant professor. His research interests include data mining, complex networks, and information retrieval.

LI Hui-Ying, born in 1983, Ph. D., assistant professor. Her research interests include bioinformatics, complex networks, and data mining.

LI Ai-Hua, born in 1978, Ph. D., associate professor. Her research interests include data mining, decision making, and management science.

Background

Community structure detection is a main focus of complex network studies. It has attracted a great deal of attentions

from various scientific fields and formed an intermediate level between the microscopic and macroscopic description of

networks. Furthermore, it is able to provide the knowledge about the relation between the function and structure of complex networks. Thus, the study of community structure plays a crucial role in the research of complex networks.

Recent researches on community detection devote most efforts to reflecting the intrinsic properties of community structure's dynamics. In this paper, we show that the local uniform behavior of Potts model across multiple timescales in the representation of the Markov variables could naturally reveal the network's hierarchical community structure. Critical topological information regarding multivariate spin configuration could also be inferred from the spectral significance of the Markov process. Furthermore, a novel algorithm is developed to determine fuzzy communities based on the optimal number of communities and the stability across multiple time-scales. The effectiveness and efficiency of the framework are theoretically analyzed as well as experimentally validated.

This research is supported by the National Natural Science Foundation Youth Projects under Grant Nos. 71401194 and 71401188; the National Nature Science Foundation Major Projects under Grant Nos. 91324203 and 11131009; "121" Youth Development Fund of CUFE Grant No. QBJ1410. The first two funds aim at studying community structure characteristics in social network specifically. The authors have made some in-depth researches on the measure of community structure, the method to uncover the hierarchical and overlapping community structure in large networks, and its application on finding the intrinsic properties of community structure. The last three funds aim to the study on complex networks from many scientific fields. The main focus is on three aspects: the statistical properties of complex networks, the model of the evolution of complex networks, the dynamics of complex networks. The study of community structure properties is a fundamental task of these works.