基于动态系统的网络社团线性探测算法

李慧嘉^{1*}, 严冠¹, 刘志东¹, 李桂君¹, 章祥荪²

1. 中央财经大学管理科学与工程学院, 北京 100081;

2. 中国科学院数学与系统科学研究院, 北京 100090

E-mail: hjli@amss.ac.cn, yg_cufe@126.com, liu_phd@163.com, ligj@cufe.edu.cn, zxs@amt.ac.cn

收稿日期: 2015-09-06; 接受日期: 2016-02-17; 网络出版日期: 2016-08-08; * 通信作者

国家自然科学基金 (批准号: 11131009, 71401194, 71271223, 71473285 和 91324203) 资助项目

摘要 社团探测技术对于理解和分析现实世界网络具有非常重要的意义. 本文提出了一种新的动态社团探测算法, 利用迭代技术高效而准确地揭示网络中的社团结构. 首先引入一种新型的基于离散时间的动态系统, 描述社团归属的从随机状态到最优划分的演化轨迹, 进一步利用严格的数学分析找出了社团归属收敛到最优的条件. 另外, 本文还创新性地提出了划分指标函数的一般化形式, 通过选择不同的参数, 可以引申到几乎所有著名的指标函数. 本文算法非常高效, 计算复杂度分析显示算法需要的时间与稀疏网络节点的数量呈线性关系. 除此之外, 为了确定社团的最优数目, 本文利用 Markov 状态转移矩阵及其特征系统给出了具体而严格的求解证明. 最后, 本文将算法应用到人工网络 and 实际网络中, 结果显示算法不仅具有极高的准确性, 还能够揭示很多有用的隐藏信息, 如层次结构和社团交互模式等.

关键词 社团探测 动态系统 层次结构 线性时间 隐藏特征

MSC (2010) 主题分类 05C82, 35R02, 05C72

1 引言

许多大规模的自然和工程系统可以用复杂网络来刻画和表示, 其中节点用来描述个体, 边用来表示个体间的联系 (参见文献 [1–4]), 如 Internet 网络、WWW 网络、电力网络、交通运输网络、新陈代谢网络、蛋白质交互网络、人际关系网和科学家合作网等 (参见文献 [5–12]). 社团结构 (community structure) 是复杂网络的一个重要特征, 它将网络分成具有密集内在联系的子群. 同一子群中的节点通常拥有共同的性质和紧密的关系, 准确地探测社团结构能够有效地促进网络系统的控制、演化和优化, 例如, 进行精确的营销和推荐^[5, 6]、中西医药物靶点的发现^[4, 7]、预测和发掘网络社交媒体的趋势和发展等 (参见文献 [13–16]).

虽然有不少经典的探测方法, 但是我们对很多基本的社团相关问题理解仍然不够深刻. 例如, 现实世界中社团结构的社会或生物意义是什么? 为什么网络中社团会自然地呈现出层次结构^[17, 18]? 怎样阐释社团的一些特性 (如显著性和稳健性) 和网络拓扑结构性质之间的关系? 这些问题的答案至今

英文引用格式: Li H J, Yan G, Liu Z D, et al. A linear community detection algorithm based on dynamical system in networks (in Chinese). Sci Sin Math, 2017, 47: 241–256, doi: 10.1360/N012015-00284

仍然十分模糊. 虽然经典优化方法和启发式算法被广泛应用到社团探测中, 但它们的基本思想是比较社团内外的直观属性和拓扑特性. 为了达到一定的准确性, 这些方法需要很高的计算复杂度, 而且结果会出现一些不可避免的缺陷, 如分辨率限制 (resolution limit) ^[19] 和错误识别 (misidentification problem) ^[10] 等. 因此, 设计一个准确高效的社团探测算法非常必要且具有挑战性.

为了回答上述问题并探索网络社团的内在特征, 本文首先引入一种新型的基于离散时间的动态系统 (如图 1(a) 所示), 通过迭代能够自动地最优化指标函数, 将社团归属从随机状态收敛到最优的划分.

图 1 (a) 根据特定的迭代机制, 我们利用一种新颖的动态系统 $X(t+1) = f(X(t))$ 来寻找最优的社团归属 $X(t)$ (用不同的颜色深度代替), 指标函数 H 用来衡量 t 时刻的社团归属 $X(t)$ 的优劣. 这里利用动态系统 $X(t+1) = f(X(t))$, 我们可以最大化方程 H 并得到 $X(t)$ 的固定的最优结果. (b) 一个拥有三个层次的社团结构的网络模式示意图, 其中一般化的观测网络 (observed network) 位于底部, 最高层次描述了一个核心 - 边缘网络 (核心模式), 细化最高层次可以得到低一级的层次

我们利用严格的数学分析找出了社团归属收敛到最优的条件, 用来给出理想的划分结果. 具体来讲, 本文的创新点主要有两个方面:

(1) 根据严格的数学推导, 我们证明了通过迭代动态系统, 可以最优化特定指标函数并得到相应社团划分, 该划分是动态系统轨迹的固定的稳定值. 在一定条件下, 动态系统收敛得到的社团, 就是最优的划分结果.

(2) 我们提出一种新颖的一般化目标函数, 可以将著名的指标函数整合成统一的形式, 这样我们可以通过选择不同的参数, 来找到合适的指标函数.

利用动态系统, 我们接着提出了一个高效的动态社团探测算法. 该算法具有非常快的速度, 其时间复杂度在稀疏网络上为 $O(n)$, 其中 n 是网络中点的个数. 特别地, 为了确定社团的最优数目, 本文利用 Markov 状态转移矩阵及其特征系统给出了具体而严格的推导. 最后, 我们将算法应用到人工网络 and 实际网络中, 结果显示, 算法不仅具有极高的准确性, 还能够揭示很多有用的信息, 如层次结构^[12]和社团交互信息等 (如图 1(b) 所示).

2 相关研究

近年来涌现出多种探测网络社团结构探测算法, 其中优化模块度 (modularity) 函数^[8,9]是最广为人知的方法. 模块度函数的增加代表了社区划分的优化, 因此, 社区探测问题可以转化成为模块度的最大化问题. 最近 Potts 模型在社团结构分析中的应用也非常广泛 (参见文献 [11, 12]). Potts 模型将社团的标号用自旋状态来表示, 并且利用系统自旋能量值来评估社团划分的质量. 假设 A 是网络 G 的邻接矩阵, σ_i 是节点 i 的社团标号, 考虑 δ 函数 (当 $\sigma_i = \sigma_j$ 时有 $\delta(\sigma_i, \sigma_j) = 1$, 否则 $\delta(\sigma_i, \sigma_j) = 0$). Reichardt 和 Bornholdt^[18] (RB) 提出了一个泛函 Milton 函数来刻画能量函数:

$$H_{RB}(\{\sigma\}) = -\frac{1}{2} \sum_{i \neq j} (A_{ij} - \gamma_{RB} p_{ij}) \delta(\sigma_i, \sigma_j), \quad (2.1)$$

其中 σ 是社团标号集合, p_{ij} 为两点之间有边的概率, $\gamma_{RB}, p_{ij} \in \mathbb{R}$. 该模型能够通过调整分辨率系数 γ_{RB} 来找到不同规模的社团, 这是模块度方法所不具有的.

Hofman 和 Wiggins^[20] 提出了一个一般化的概率模型并定义了标准化的能量函数:

$$H_{HW}(\{\sigma\}) = -\frac{1}{2} \sum_{i \neq j} (W_L A_{ij} - W_G) \delta(\sigma_i, \sigma_j) + \frac{1}{2} \sum_{\mu=1}^K h_\mu \sum_{i=1}^n \delta(\sigma_i, \mu), \quad (2.2)$$

其中

$$W_G = \log \frac{1 - p_{out}}{1 - p_{in}}, \quad W_L = \log \frac{p_{in}}{p_{out}} + W_G,$$

p_{in} 是社团团内边的比例, p_{out} 是社团团间边的比例, 该能量函数通过 Bayes 方法进行最优化处理.

除此之外, 标签传播方法是一种接近线性时间的快速社团探测方法^[21]. 简而言之, 算法以分配给每个节点不同的社团标志为起点, 然后每个节点通过替换和标记最常用的相邻节点来更新自己的标签. 标签传播与寻找局部能量最小值的零温度动力学 Potts 模型等同优化下列函数:

$$H_{KPM}(\{\sigma\}) = - \sum_{i \neq j} A_{ij} \delta(\sigma_i, \sigma_j). \quad (2.3)$$

另外, 常见的社团探测方法还包括模拟退火算法 (SA) [16]、外部优化算法 [2]、最大化期望值算法 [3], Bayes 推理方法和变分 Bayes 方法等 (参见文献 [17, 20]).

3 问题描述

3.1 基本定义

首先我们提出一些基本定义:

- (1) 参数 c 、 n 和 m 分别代表网络中社团、节点和边的数量;
- (2) l_μ^{in} 、 l_μ^{out} 和 n_μ 分别代表社团 μ 团内边、团间边和点的数量;
- (3) p_μ^{in} 代表社团 μ 团内边的数量与所有可能的边的数量的比值, p_μ^{out} 代表社团 μ 团间边的数量与所有可能的边的数量的比值.

易得

$$p_\mu^{\text{in}} = \frac{2l_\mu^{\text{in}}}{n_\mu(n_\mu - 1)}, \quad p_\mu^{\text{out}} = \frac{l_\mu^{\text{out}}}{n_\mu(n - n_\mu)}. \quad (3.1)$$

一个节点可以以不同的概率分属不同的社团. 考虑一个由 n 个节点、 m 条边和 c 个社团构成的网络, 本文引入向量 $X_i(t) = [x_{i\mu}(t)]$ ($\mu = \{1, 2, \dots, c\}$), 其中 $x_{i\mu}(t)$ 表示在时刻为 t 时, 节点 i 属于社团 μ 的可能性. 易得

$$0 \leq x_{i\mu}(t) \leq 1 \quad \text{且} \quad \sum_{\mu=1}^c x_{i\mu}(t) = 1, \quad \forall i \in V.$$

集合 $X(t) = [x_i(t)]$ ($i = \{1, 2, \dots, n\}$) 为在 t 时刻网络 G 的社团归属 (community configuration).

定义 1 (硬划分) 在硬划分中, 点 i 在 t 时刻仅属于特定社团 c , 向量 $X_i(t)$ 的第 c 个元素为 1, 其余值都等于 0. 在硬划分中, 不存在重叠节点 (overlapping nodes).

定义 2 (节点 - 社团度) 我们将节点 i 在社团 μ 中的节点 - 社团度表示为 $k_{i\mu}$, 它定义为节点 i 连接到社团 μ 的边数. 它的数学化定义为 $k_{i\mu} = \sum_{\alpha} x_{\alpha\mu} A_{\alpha i}$.

利用社团归属 $X(t)$, 一些常用概念 n_μ 、 l_μ^{in} 、 l_μ^{out} 、 p_μ^{in} 和 p_μ^{out} 可以重新表述为

$$n_\mu(t) = \sum_i x_{i\mu}(t), \quad (3.2)$$

$$l_\mu^{\text{in}} = \frac{1}{2} \sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}, \quad (3.3)$$

$$l_\mu^{\text{out}} = \sum_i \sum_{j \neq i} x_{i\mu}(t) (1 - x_{j\mu}(t)) A_{ij}, \quad (3.4)$$

$$\begin{aligned} p_\mu^{\text{in}} &= \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{\sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t)} = \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{\sum_i x_{i\mu}(t) (n_\mu(t) - x_{i\mu}(t))} \\ &= \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t) A_{ij}}{n_\mu^2(t) - \sum_i x_{i\mu}^2(t)}, \end{aligned} \quad (3.5)$$

$$\begin{aligned} p_\mu^{\text{out}} &= \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) (1 - x_{j\mu}(t)) A_{ij}}{\sum_i \sum_{j \neq i} x_{i\mu}(t) (1 - x_{j\mu}(t))} = \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) (1 - x_{j\mu}(t)) A_{ij}}{\sum_i \sum_{j \neq i} x_{i\mu}(t) - \sum_i \sum_{j \neq i} x_{i\mu}(t) x_{j\mu}(t)} \\ &= \frac{\sum_i \sum_{j \neq i} x_{i\mu}(t) (1 - x_{j\mu}(t)) A_{ij}}{(n - 1)n_\mu(t) - (n_\mu^2(t) - \sum_i x_{i\mu}^2(t))}. \end{aligned} \quad (3.6)$$

3.2 基于线性迭代的动态系统

相比大多数经典算法, 如层次分析法和启发式算法, 利用动态系统更新归属向量 (即 $X = \{x_{i\mu}\}$) 到一个理想的固定值是非常简单的, 其计算复杂度非常小, 便于在大规模网络上进行计算和分析. 虽然有一些其他的动态系统, 如随机游走和协同系统, 也能够将归属矩阵收敛到一个稳定的状态, 但是它们仅是对归属矩阵进行迭代而无法保证收敛到最优值 (最优化特定的指标函数 Q). 为了准确而快速地探测社团结构, 本文提出一种新型的基于动态系统的线性迭代算法. 这里, 算法首先给出离散时间下的动态系统:

$$x_{i\mu}(t+1) = \frac{x_{i\mu}(t)e^{Q_{i\mu}(x_{*\mu}(t))}}{\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t)e^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))}}, \quad (3.7)$$

其中 $Q_{i\mu}(x_{*\mu}(t)): \mathbb{R}^n \rightarrow \mathbb{R}$ 是节点 i 对社团 μ 归属程度的适应度函数 (质量函数), $x_{*\mu}$ 是一个 n 维向量, 第 i 维是 $x_{i\mu}$, 函数 Q 满足约束条件

$$\frac{\partial Q_{i\mu}}{\partial x_{i\mu}} = 0. \quad (3.8)$$

这里, 我们推断出了几个关于动态系统的重要性质.

命题 1 离散时间的动态系统 (3.7) 有两类平凡稳定点 (trivial fixed points), 分别为

- (1) 硬社团结构的社团归属, 即 $x_{i\mu}(t) = 1$ 且 $x_{i\tilde{\mu}}(t) = 0, \mu \neq \tilde{\mu}$;
- (2) 均匀的社团归属, 即 $x_{i\mu}(t) = \frac{1}{c}, i = \{1, 2, \dots, n\}$ 和 $\mu = \{1, 2, \dots, c\}$.

证明 当 $x_{i\mu} = 1$ 和 $x_{i\tilde{\mu}} = 0$ 时, 对于 $\forall \tilde{\mu} \neq \mu$, 我们可以得到

$$x_{i\mu}(t+1) = \frac{x_{i\mu}(t)e^{Q_{i\mu}(x_{*\mu}(t))}}{\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t)e^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))}} = \frac{x_{i\mu}(t)e^{Q_{i\mu}(x_{*\mu}(t))}}{x_{i\mu}(t)e^{Q_{i\mu}(x_{*\mu}(t))} + \sum_{\tilde{\mu} \neq \mu} (0 \times e^{Q_{i\tilde{\mu}}(x_{*\tilde{\mu}}(t))})} = 1. \quad (3.9)$$

同理, 对于所有的 i , 有

$$x_{i\mu}(t) = \frac{1}{c},$$

这意味着 $x_{i\mu}(t+1) = \frac{1}{c}$ 对于所有的 i 都成立. $\square$

为了方便阐述, 此后用 $Q_{i\mu}(t)$ 代替 $Q_{i\mu}(x_{*\mu}(t))$.

定理 1 考虑一个离散 (连续) 时间的动态系统, x^* 为相应的动态系统结果曲线的稳定值, 有下列性质:

- (1) 当且仅当 Jacobi 矩阵 $\frac{\partial F}{\partial x}(x^*)$ 的所有的特征值 λ_i 满足 $|\lambda_i| < 1$ ($\text{Re}(\lambda_i) < 0$) 时, x^* 是渐近稳定的.
- (2) 当且仅当至少存在一个 Jacobi 矩阵 $\frac{\partial F}{\partial x}(x^*)$ 的特征值满足 $|\lambda_i| > 1$ ($\text{Re}(\lambda_i) > 0$) 时, x^* 是不稳定的.

接下来证明利用动态系统 (3.7) 可以最优化适应度函数 Q .

定理 2 对于任意的硬社团归属 $X(t)$, 我们记 μ_i^* 为归属值 $x_{i\mu_i^*}(t) = 1$ 时的社团. 那么, 当

$$\forall i, \quad \mu \neq \mu_i^*, \quad Q_{i\mu}(t) < Q_{i\mu_i^*}(t), \quad (3.10)$$

动态系统 (3.7) 在 $X(t)$ 有一个渐近式的稳定点. 如果对于特定节点 i 和社团 $\mu, \mu \neq \mu_i^*$, 我们可以得到 $Q_{i\mu}(t) > Q_{i\mu_i^*}(t)$, 则此稳定点不固定.

证明 为了确定动态系统在 $X(t)$ 处的 Jacobi 矩阵, 系统 (3.7) 的 $n \times K$ 个状态变量和相应的导

数计算如下:

$$\begin{aligned} \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)} &= \frac{e^{Q_{i\mu}(t)} \sum_{\tilde{\mu} \neq \mu} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)}}{(\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)})^2}, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)} \Big|_{x_{i\mu}(t)=1} &= 0, \end{aligned} \quad (3.11)$$

$$\begin{aligned} \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\mu}(t)} \Big|_{x_{i\mu}(t)=1, x_{i\tilde{\mu}}(t)=0} &= \frac{e^{Q_{i\mu}(t)}}{e^{Q_{i\tilde{\mu}}(t)}}, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\tilde{\mu}}(t)} &= -\frac{x_{i\mu}(t) e^{Q_{i\mu}(t)} e^{Q_{i\tilde{\mu}}(t)}}{(\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)})^2}, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\tilde{\mu}}(t)} \Big|_{x_{i\mu}(t)=1} &= -\frac{e^{Q_{i\tilde{\mu}}(t)}}{e^{Q_{i\mu}(t)}}, \end{aligned} \quad (3.12)$$

$$\begin{aligned} \frac{\partial x_{i\mu}(t+1)}{\partial x_{i\tilde{\mu}}(t)} \Big|_{x_{i\mu}(t)=0} &= 0, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)} &= \frac{x_{i\mu}(t) e^{Q_{i\mu}(t)} \left(\frac{Q_{i\mu}(t)}{x_{j\mu}(t)} \right) \sum_{\tilde{\mu} \neq \mu} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)}}{(\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)})^2}, \quad i \neq j, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)} \Big|_{x_{i\mu}(t)=1} &= 0, \end{aligned} \quad (3.13)$$

$$\begin{aligned} \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\mu}(t)} \Big|_{x_{i\mu}(t)=0} &= 0, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\tilde{\mu}}(t)} &= \frac{x_{i\mu}(t) e^{Q_{i\mu}(t)} \left(\frac{Q_{i\mu}(t)}{x_{j\tilde{\mu}}(t)} \right) x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)}}{(\sum_{\tilde{\mu}} x_{i\tilde{\mu}}(t) e^{Q_{i\tilde{\mu}}(t)})^2}, \quad i \neq j, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\tilde{\mu}}(t)} \Big|_{x_{i\mu}(t)=1} &= 0, \\ \frac{\partial x_{i\mu}(t+1)}{\partial x_{j\tilde{\mu}}(t)} \Big|_{x_{i\mu}(t)=0} &= 0. \end{aligned} \quad (3.14)$$

对所有的 μ 和 $\tilde{\mu}$, 假设 J 是一个包含 $\frac{x_{i\tilde{\mu}}(t+1)}{x_{i\mu}(t)}$ 的 $K \times K$ 的矩阵块 (matrix block). 接着, 在方程 (3.11)–(3.14) 中运用合适的排列和替换, 我们可以得到 J_{ii} 在 $X(t)$ 的值为

$$J_{ii} = \left[\begin{array}{cccc} 0 & -\frac{e^{Q_{i1}(t)}}{e^{Q_{i\mu_i^*}(t)}} & \cdots & -\frac{e^{Q_{ik}(t)}}{e^{Q_{i\mu_i^*}(t)}} \\ 0 & \frac{e^{Q_{i2}(t)}}{e^{Q_{i\mu_i^*}(t)}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{e^{Q_{ik}(t)}}{e^{Q_{i\mu_i^*}(t)}} \end{array} \right] \Big|_{X(t)}. \quad (3.15)$$

在硬划分的稳定点下的整个 Jacobi 矩阵可以由下面的矩阵来表示:

$$J = \left[\begin{array}{ccc} J_{11} & 0 & 0 \\ 0 & J_{ii} & 0 \\ 0 & 0 & J_{nn} \end{array} \right] \Big|_{X(t)}. \quad (3.16)$$

很显然, 在硬划分的稳定点下的 Jacobi 矩阵 J 的特征值为

$$\lambda_{i\mu_i^*} = 0, \quad \lambda_{i\mu} = \frac{e^{Q_{i\mu}(t)}}{e^{Q_{i\mu_i^*}(t)}}, \quad i = 1, 2, \dots, n, \quad \mu \neq \mu_i^*. \quad (3.17)$$

基于定理 1, 动态系统 (3.7) 在 $X(t)$ 有一个渐近式的稳定点, $\forall i, \mu, |\lambda_{i\mu}| < 1$ 已经对所有 $\lambda_{i\mu^*}$ 都满足, 而且剩余也满足的话需满足下面条件:

$$\forall i, \quad \mu \neq \mu_i^*, \quad |\lambda_{i\mu}| = \left| \frac{e^{Q_{i\mu}(t)}}{e^{Q_{i\mu_i^*}(t)}} \right| < 1, \quad (3.18)$$

可以得到

$$\forall i, \quad \mu \neq \mu_i^*, \quad e^{Q_{i\mu}(t)} < e^{Q_{i\mu_i^*}(t)} \Rightarrow Q_{i\mu}(t) < Q_{i\mu_i^*}(t). \quad (3.19)$$

另一方面, 如果某个点对 $i\mu, \mu \neq \mu_i^*$, 我们可以得到 $Q_{i\mu}(t) > Q_{i\mu_i^*}(t)$, 那么这个稳定点不固定. $\square$

定理 2 直观地说明了函数值 $Q_{i\mu}(t)$ 是“节点 i 属于社团 μ_i^* ”的适应程度值. 因此, 如果硬社团归属的稳定值是固定的, 那么所有节点的相关适应程度值都应该是最大的. 相应地, 社团探测问题可以简化为设计合适的 Q 函数, 用来衡量单个点在所有社团的归属程度.

到目前为止, 我们介绍了一种基于迭代的探测方法以及分析了动态系统收敛的具体条件. 接下来, 我们为算法设计一种具有代表性的一般化社团探测指标. 为此, 让我们考虑方程

$$H(X(t)) = \frac{1}{2} \sum_{\mu} \sum_{i=1}^n x_{i\mu}(t) Q_{i\mu}(t). \quad (3.20)$$

接下来证明最大化方程 H 会得到动态系统 (3.7) 的固定的稳定值.

定理 3 如果记 $X(t)$ 为社团归属向量, 那么最大化方程 H 得到的社团归属向量 $X(t)$ 是动态系统 (3.7) 的固定的稳定值.

证明 如果 $X(t)$ 最大化方程 H 得到的社团归属向量. 假设 $X(t)$ 是动态系统 (3.7) 的不固定的稳定值, 那么会有

$$\exists i, \quad \exists \tilde{\mu} \neq \mu_i^*, \quad Q_{i\tilde{\mu}}(t) > Q_{i\mu_i^*}(t), \quad (3.21)$$

其中 μ_i^* 是使得 $x_{i\mu_i^*}(t) = 1$ 的社团标号. 如果 $x_{i\mu_i^*}(t)$ 和 $x_{i\tilde{\mu}}(t)$ 的值互换, 那么会有 $x_{i\mu_i^*}(t) = 0$ 和 $x_{i\tilde{\mu}}(t) = 1$, 则方程 H 的值会有如下相应变化:

$$H(X(t)) \big|_{x_{i\mu_i^*}(t)=0, x_{i\tilde{\mu}}(t)=1} - H(X(t)) \big|_{x_{i\mu_i^*}(t)=1, x_{i\tilde{\mu}}(t)=0} = Q_{i\tilde{\mu}}(t) - Q_{i\mu_i^*}(t) > 0. \quad (3.22)$$

上式意味着, 如果节点属于社团 $\tilde{\mu}$ 而不是 μ_i^* , 方程 H 的值反而更高, 这是矛盾的. 因此, 我们开始的假设是无效的, 进而可以验证最大化方程 H 得到的社团归属向量 $X(t)$ 是动态系统 (3.7) 的固定的稳定值. $\square$

3.3 方程 Q 的一般化形式和扩展形式

我们发现, 对于许多社团探测算法, 优化目标函数具有以下的一般形式:

$$E(t) = -\frac{1}{2} \sum_{\mu} \sum_{i=1}^n \left(\sum_{j=1}^n f_{\mu}^{+} A_{ij} x_{i\mu}(t) x_{j\mu}(t) - \sum_{j=1}^n f_{\mu}^{-} (1 - A_{ij}) x_{i\mu}(t) x_{j\mu}(t) \right) + \sum_{\mu} R_{\mu}, \quad (3.23)$$

其中 f^{+} 为奖励项参数, f^{-} 为惩罚项参数, R_{μ} 为一般正则项 (冗余项). 方程 (3.23) 可以表述为

$$\begin{aligned} E(t) &= -\frac{1}{2} \sum_{\mu} \left[\frac{2 \sum_{j=1}^n x_{j\mu}(t)}{l_{\mu}(t)} R_{\mu} + \sum_{i=1}^n \left(\sum_{j=1}^n f_{\mu}^{+} A_{ij} x_{i\mu}(t) x_{j\mu}(t) - \sum_{j=1}^n f_{\mu}^{-} (1 - A_{ij}) x_{i\mu}(t) x_{j\mu}(t) \right) \right] \\ &= -\frac{1}{2} \sum_{\mu} \sum_{i=1}^n x_{i\mu}(t) \left(\sum_{j=1}^n f_{\mu}^{+} A_{ij} x_{j\mu}(t) - \sum_{j=1}^n f_{\mu}^{-} (1 - A_{ij}) x_{j\mu}(t) + \frac{2}{l_{\mu}(t)} R_{\mu} \right), \end{aligned} \quad (3.24)$$

其中 $l_{\mu}(t) = \sum_{j=1}^n x_{j\mu}(t)$ 是 t 时刻社团 μ 的规模 (节点个数).

事实上, 我们可以将不同类型的指标函数解释为其具有不同的奖励参数 f^{+} 、惩罚参数 f^{-} 和 R_{μ} , 如模块度函数和 Hamilton 函数. 每一种不同参数组合的选择都有其内在原因、优势和缺点. 定义如下方程:

$$Q_{i\mu}(t) = \sum_{j=1}^n f_{\mu}^{+} A_{ij} x_{j\mu}(t) - \sum_{j=1}^n f_{\mu}^{-} (1 - A_{ij}) x_{j\mu}(t) + R_{i\mu}, \quad (3.25)$$

选择 $R_{i\mu}$ 使得 $\partial R_{i\mu} / \partial x_{i\mu}(t) = 0$ 和 $R_u = \sum_{i=1}^n R_{i\mu}$, 例如, $R_{i\mu} = \frac{2}{l_{\mu}} R_u$, 那么方程 (3.23) 完全符合方程 (3.20) 所定义的能量函数形式.

有趣的是, 对于特定的社团归属 $x_{i\mu}(t)$, 方程 (3.25) 所定义的 $Q_{i\mu}$ 可以引申到几乎所有的著名指标函数, 因此, 它是具有代表性的一般化指标:

(1) Hofman-Wiggins 模型^[20]

$$f_{\mu}^{+} = \log \frac{p^{\text{in}}}{p^{\text{out}}}, \quad f_{\mu}^{-} = \log \frac{1 - p^{\text{out}}}{1 - p^{\text{in}}}, \quad R_{\mu} = l_{\mu} \log \pi_{\mu}. \quad (3.26)$$

(2) Ronhovde-Nussinov 模型^[22]

$$f_{\mu}^{+} = 1, \quad f_{\mu}^{-} = \min_{\mu} p_{in, \mu}, \quad R_{\mu} = 0. \quad (3.27)$$

(3) RB Potts 模型 (Erdos-Renyi 空模型)^[18]

$$f_{\mu}^{+} = 1 - \gamma_{\text{RB}} p, \quad f_{\mu}^{-} = \gamma_{\text{RB}} p, \quad R_{\mu} = 0. \quad (3.28)$$

(4) RB Potts 模型 (configuration 空模型)^[18]

$$f_{\mu}^{+} = 1 - \frac{\gamma_{\text{RB}}}{2m}, \quad f_{\mu}^{-} = \frac{\gamma_{\text{RB}}}{2m}, \quad R_{\mu} = \sum_{i>j} \frac{\gamma_{\text{RB}}}{2m} (k_i k_j - 1) x_{i\mu}(t) x_{j\mu}(t), \quad (3.29)$$

其中 k_i 是节点 i 的度, m 是网络中的边数.

(5) 模块度 (modularity)^[9]

$$f_{\mu}^{+} = 1, \quad f_{\mu}^{-} = \frac{k_i k_j}{2m}, \quad R_{\mu} = \sum_{i>j} \frac{1}{2m} (k_i k_j - 1) x_{i\mu}(t) x_{j\mu}(t), \quad (3.30)$$

其中 k_i 是节点 i 的度, m 是网络中的边数.

(6) 标号传播 (Label propagation) [21]

$$f_{\mu}^{+} = 1, \quad f_{\mu}^{-} = 0, \quad R_{\mu} = 0. \quad (3.31)$$

算法 1 社团探测算法

输入: 给定具有 n 个点和 m 条边的网络 G , 及算法的参数 f_{μ}^{+} 、 f_{μ}^{-} 、 R_{μ} 和社团个数 c ;

输出: 社团归属矩阵 X .

- (1) 重复.
- (2) 初始化 $X(0)$.
- (3) 重复.
- (4) 根据 (3.25) 计算 $Q_{i\mu}(t)$.
- (5) 根据 (3.7) 迭代更新归属矩阵 $X(t)$.
- (6) 直到超出预设的最大迭代次数.
- (7) 根据 (3.20) 计算 $H(X(t))$.
- (8) 如果 $H(X(t)) > H^{\text{best}}$,
设定 $H^{\text{best}} = H(X(t))$ 和 $X^{\text{best}} = X(t)$,
直到超出预设的最大迭代次数.

4 社团探测算法描述

4.1 算法复杂性分析

在算法 1 中, 每一步计算方程 H 的值会花费 $O(mc)$ 的时间, 其中 m 是网络中边的数目, c 是预设的社团的可能数目. 动态方程 (3.7) 中的指数项, 会使得收敛速度非常快. 算法总的复杂性与迭代次数和网络的初始性质有关系, 一些具有强社团结构的网络会消耗较少的时间. 总而言之, 对于大规模稀疏网络来讲, 算法的复杂性是 $O(m)$, 由于网络的边数 m 与节点规模 n 呈线性关系, 因此, 算法是非常快速的 (一些著名算法的运行时间情形见表 1). 图 2 展示了六种算法在不同规模网络下的运行时间情形, 即 Newman 快速算法 (NF) [8]、Girvan-Neman (GN) 算法 [15]、Louvain 算法 [13]、Danon 算法 [23]、SA 算法 [16] 和本文的算法. 实验的运行环境是一个具有 2GHz CPU 和 2GB 内存的台式电脑, 操作系统为 Windows 7, 模拟软件是 Matlab 2013b. 我们利用 Lancichinetti-Fortunato-Radicchi (LFR) 基准模型 [11, 24] 构建一个包含社团结构的稀疏网络模型, 网络中包含 n 个点和 $m = O(n)$ 条边. 从图中可以看出, (1) 我们的算法比其他算法要快很多, 证明了算法的高效性; (2) 我们算法的运行时间与网络规模基本呈正比关系, 因此, 非常适合于大规模社会及生物网络的分析计算.

表 1 在稀疏网络中, 著名算法计算复杂度的比较

算法	复杂性
CNM [14]	$O(n \log^2 n)$
DA [25]	$O(n^2 \log n)$
Louvain [13]	$O(n \log n)$
OCR-HK [26]	$O(n^2)$
Bayes 推理 [17]	$O(n \log^{\alpha} n)$
变分 Bayes [20]	$O(n^{1.44})$
Reichardt-Bornholdt 模型 [22]	$O(n^{1.3})$
标签传播模型 [21]	$O(n + m)$

图 2 不同规模网络下, 各个算法运行时间的比较

4.2 社团配置初始化

在算法 1 的 (2) 中, 社团归属向量应该首先被初始化. 我们用以下方式初始化向量 $X(0)$, 首先,

$$\begin{aligned} x_{i\mu}(0) &= \frac{1}{c} + y_{i\mu}, \quad i = \{1, 2, \dots, n\}, \quad \mu = \{1, 2, \dots, c\}, \\ \text{s.t. } \sum_{\mu=1}^c y_{i\mu} &= 0, \end{aligned} \quad (4.1)$$

其中 $y_{i\mu}$ 是一个小的 Gauss 噪声. 这种初始化策略会避免命题 1 中的平凡解. 在初始化后的步骤中, 我们提出一种有效策略来优化先前步骤中的最优解 (见算法 1),

$$\begin{aligned} x_{i\mu}(0) &= x_{i\mu}^{\text{best}} + y_{i\mu}, \quad i = \{1, 2, \dots, n\}, \quad \mu = \{1, 2, \dots, c\}, \\ \text{s.t. } \sum_{\mu=1}^c y_{i\mu} &= 0, \end{aligned} \quad (4.2)$$

其中 $x_{i\mu}^{\text{best}}$ 是 X_{best} 的第 (i, μ) 个元素. 换言之, 该迭代过程尝试以累积的方式提高计算的性能. 值得注意的是噪声的指数项 $y_{i\mu}$, 这一步必须大于第一步, 否则系统的动态过程会收敛到上一次的结果.

4.3 确定最优社团个数

在很多算法中, 社团的最优数目并不是一个已知的先验信息, 但是会极大地影响社团划分的结果, 如 Clauset-Newman-Moore (CNM) 算法^[14]、Louvain 算法^[13]、SA 算法^[16]和 Duch-Arenas (DA) 算法^[25]. 本文算法需要在初始设定最优的社团数目, 因此, 我们利用 Markov 过程, 借助谱理论给出最优社团数目的严格推导, 并能够计算社团结构显著性等隐藏特性. 对于遍历式的 Markov 过程, P^t 表示经过 t 步后的转移概率矩阵. 为了方便利用谱性质, 我们将 P^t 扩展到对称的形式 $G^{(t)} = (P^t + (P^t)^T)/2$. 假设网络被划分成 c 个社团或子图 $V_\mu \subset V$, $\mu \in \{1, 2, \dots, c\}$, 社团的规模为 $n_\mu = |V_\mu|$. 社团 μ 的归属向量为 x_μ , 表示每个节点在社团 μ 中的归属程度. 根据硬划分的特性, 向量 x_l 和 x_s 应该是正交的.

给定社团个数 c , 我们可以通过最大化下面的方程来寻找最优社团结构^[11,12]

$$J_c^{(t)} = \sum_{\mu=1}^c \sum_{i,j \in V_\mu} \frac{[G^{(t)}]_{i,j}}{n_\mu} = \sum_{\mu=1}^c \frac{x_\mu^T G^{(t)} x_\mu}{x_\mu^T x_\mu}, \quad (4.3)$$

其中 $x_\mu \in \{0,1\}$, 如果 $l \neq s$, 则 $x_l^T x_s = 0$. 通过将向量 x_μ 整合成矩阵 $X = (x_1, x_2, \dots, x_c)$, 方程 (4.3) 可以重写为矩阵的形式:

$$J_c^{(t)} = \text{tr}\{(X^T X)^{-1} X^T G^{(t)} X\} = \text{tr}\{(X^T X)^{-1/2} X^T G^{(t)} X (X^T X)^{-1/2}\}, \quad (4.4)$$

其中矩阵 $X^T X$ 是正交的. 利用简单的替换 $Y = X(X^T X)^{-1/2}$, 可以得到 $J_c^{(t)} = \text{tr}\{Y^T G^{(t)} Y\}$. 因为

$$Y^T Y = (X^T X)^{-1/2} (X^T X) (X^T X)^{-1/2} = I_c,$$

因此, 向量 $y_\mu \subset Y$ 具有单位长度并互相正交. 根据上面的替换, 最优化问题可以表示为 Y 的形式

$$\max_{Y^T Y = I} \text{tr}\{Y^T G^{(t)} Y\}. \quad (4.5)$$

基于对称矩阵的 Rayleigh-Ritz 定理^[27], 该最大化问题的最优解为 $G^{(t)}$ 的前 c 个最大特征值对应的特征向量 $U = \{u_1, \dots, u_c\}$. 我们可以得到个数 μ 的特定社团强度为特征值的 t 次方

$$S_\mu^{(t)} \approx \frac{u_\mu^T G^{(t)} u_\mu}{u_\mu^T u_\mu} = \lambda_\mu^t \frac{u_\mu^T u_\mu}{u_\mu^T u_\mu} = \lambda_\mu^t. \quad (4.6)$$

从 (4.6) 可以看出, 随时间逐渐消失的小特征值表明, 一些相对较弱的社团结构被同化, 而这些社团特征对应于不同的特征向量. 为了能够划分社团结构, 我们对中间时刻的状态更为感兴趣. 如果我们计划划分出 z 个社团, 我们期望在一个特定时刻 t , 矩阵 P^t 的特征值 λ_μ^t 只在区间 $\mu = 1, \dots, z$ 内显著大于 0, 这可以通过确定时刻 t , 使得 $|\lambda_\mu|^t \approx 0$. 从另一个角度来说, t 时刻的社团强度 λ_μ^t 可以看成是一种度量指标, 特征值差 $\lambda_{\mu-1}^t - \lambda_\mu^t$ 可以认为是 t 时刻 μ 社团状态转移到 $\mu - 1$ 社团状态的难度. 我们可以通过计算最大的特征差值, 得出特定时刻 t 最适合的社团个数 $\Lambda(t)$,

$$\Lambda(t) = \arg \left[\max_{\mu} (\lambda_{\mu-1}^t - \lambda_\mu^t) \right]. \quad (4.7)$$

从全局的角度来看, 如果社团个数 $\Lambda(t)$ 在最长的时间内没有改变, 我们可以把它当作整个网络的最优社团个数. 另外, 在一定程度上, 我们还可以找出每个时刻 t 整个网络的显著性, 定义为该时刻最优社团个数下社团结构特征差值, 即 $\Omega(t) = \lambda_{\Lambda(t)-1}^t - \lambda_{\Lambda(t)}^t$.

5 数值实验

根据算法 1 中, 为了确定函数 Q , 我们需要输入 f_μ^+ 、 f_μ^- 和 R_{mu} 的具体值. 在本文实验中, 为了测试算法的高效性, 我们选择 $f_\mu^+ = 1$ 、 $f_\mu^- = \min_{\mu} \frac{\langle k \rangle_\mu}{n_\mu}$ 和 $R_{mu} = 0$ 来进行测试, 其中 $\langle k \rangle_\mu$ 是社团 μ 的平均度, n_μ 是社团 μ 的规模.

首先, 我们在两个著名的人工网络, 即 GN 网络和 LFR 网络上测试算法的性能. 这里使用标准化互信息值 (NMI) 去测试社团划分的精确度. NMI 值处于区间 $[0, 1]$ 之间, 用来指示真实的划分和实际划分之间的重叠程度. 特别地, 当实际划分和真实划分完全一致时, NMI 的值等于 1; 相反, 如果真实划分和实际划分的重叠程度越低, NMI 的值越小. 真实划分 X 和实际划分 Y 的互信息值可以用下列数学公式定义:

$$I(x, y) = \frac{-2 \sum_{i=1}^{c_X} \sum_{j=1}^{c_Y} N_{ij} \ln(n N_{ij} / N_i N_j)}{\sum_{i=1}^{c_X} N_i \ln(N_i / n) + \sum_{j=1}^{c_Y} N_j \ln(N_j / n)}, \quad (5.1)$$

其中 c_X 和 c_Y 分别表示划分 X 和 Y 中社团的个数. N 表示混同矩阵, 其中 N_{ij} 表示真实社团 i 和实际社团 j 重叠节点的个数, N_i 和 N_j 分别表示矩阵 N 第 i 行和第 j 列的和.

图 3 在 GN 网络^[9]上比较了不同算法的性能. GN 网络由 Newman 等提出, 共包含 $n = 128$ 个节点, 分成 4 个社团, 每个社团包含 32 个节点. 假设属于相同社团的节点对以概率 p_{in} 相连接, 而属于不同社团的节点对以概率 p_{out} 相连接. 因此对于每一个节点, 团内部的期望值和团间边的期望值分别为 $Z_{in} = 31p_{in}$ 和 $Z_{out} = 96p_{out}$. 网络中点的平均节点度固定为 $\langle k \rangle = 16$, 可以得到 p_{in} 和 p_{out} 有关系 $31p_{in} + 96p_{out} = 16$. 可以看出, 当团间边数目 $Z_{out} \leq 7.5$ 时, 所有算法基本上都能找出正确的社团归属 (互信息值 $NMI \geq 0.99$). 但是随着团间边比例 Z_{out} 的增加, 大多数算法都不能高效地找出正确的社团划分, 而我们的算法直到 Z_{out} 接近 8 时仍然拥有最高的效率, 验证了其划分的高效性.

进一步, 我们在 LFR 网络上比较了不同社团的效率, 结果如图 4 所示. LFR 网络模型由 Lancichinetti 等^[24]提出, 具有无标度特征的度分布和社团规模分布, 因此更加接近现实中社会和生物系统. LFR 网络模型的生成由一些参数控制: 节点数目 n 、节点的平均度 $\langle k \rangle$ 、最大节点度 $\max_k$ 、混合比例参数 θ (mixing parameter, 即每个节点与其他社团的节点共享 θ 比例的边)、最小社团规模 $\min_c$ 和最大社团规模 $\max_c$. θ 的变化范围为 $[0, 1]$, 用来调整网络的模糊性, 较大的 θ 代表更加弱的社团结构. 在本次实验中, 每个社团的规模有大小之分, 具体来讲, 在 $n = 1,000$ 的网络中, 小 (大) 社团的数目为 100 (50), 而在 $n = 5,000$ 的网络中, 小 (大) 社团的数目为 300 (200). 从图 4 可以看出, 我们的算法在各种网络配置中, 都几乎拥有最高的准确性, 展示了其有效性和普遍适用性.

另外, 除了 GN 网络和 LFR 网络, 我们进一步在拥有 1,000 个团、每个团拥有 10 个节点的团环网络 (ring of cliques networks) 上进行验证. 团环网络由文献 [19] 引入, 用来说明模块度指标 (modularity) 具有分辨率限制问题 (resolution limit problem)^[19]. 团环网络将 k 个完全图用少量边连接起来, 因此

图 3 GN 网络中, 各个算法性能的比较, 其中每个点都是 50 次结果的平均

图 4 在不同 LFR 网络各个算法的性能比较. (a) $n = 1,000$ 个节点的小规模社团 LFR 网络; (b) $n = 1,000$ 个节点的大规模社团 LFR 网络; (c) $n = 5,000$ 个节点的小规模社团 LFR 网络; (d) $n = 5,000$ 个节点的大规模社团 LFR 网络, 图中每个点都是 50 次实验结果的平均

拥有 2 个参数, 即团的数目和每个团的规模. 结果显示, 我们的算法能够完美地发现每个团, 而不具有分辨率限制问题, 进一步验证了其有效性.

5.1 揭示多层次社团交互模式

为了进一步验证算法在实际网络中的有效性, 我们利用 Krebs 构建的政治书共购买网络 (political book co-purchasing network) 来开展实验. 为了揭示社团结构中交互模式, 我们定义社团交互强度 C_{ks} 为社团 k 和 s 之间相互连边的比率. 另外, 社团 μ 的重要性可以定义为其与各个社团交互强度的加和, 即 $I_{\mu} = \sum_s C_{\mu s}$. 划分结果如图 5(a) 所示, 其中节点代表在亚马逊网络上出售的美国政治书, 边代表两本书会经常一起被购买. 另外, 书也根据其内容描述被标上“自由主义”、“中立主义”或者“保守主义”的标签^[28]. 从图 5(b) 可以看出, $\tau = 1$ 时 7 个社团对应的 $\max \Theta(7) = 0.689$, 大于 $\tau = 6$ 是 2 个社团对应的 $\max \Theta(2) = 0.483$, 这意味着 7 和 2 是最优的社团个数.

通过在政治书共购买网络上运行算法, 我们发现了一个双层社团结构. 图 6(a) 展示了第一个层次的社团交互图, 其中不同的社团用不同的标号表示, 社团的大小表示其重要性. 社团交互强度如果高于平均阈值则用粗箭头显示出来. 通过在社团交互图中识别交互规则, 我们发现了一些有趣的模式, 例如, 包含两个枢纽 (社团 2 和 7)、一个离散组织 (社团 4) 和两个羽翼 (社团 1, 2, 3 和 5, 6, 7). 在无向网络中, 枢纽 (社团 2 和 7) 可以看作是两个羽翼的重叠部分. 值得注意的是, 虽然两个羽翼中较远社团的政治书可能不会同时购买, 但是通过中间的枢纽社团, 他们可以拥有相同的标号, 例如, 都是“自由主义”. 进一步在第二个层次中, 算法识别了一个宏观的交互模式, 如图 6(b) 所示. 有趣的是, 根据包含的书的类型, 左右的一些社团可以在全局上认为是“左羽翼 (社团 1, 2, 3)”和“右羽翼 (社团 5,

(a)

(b)

图 5 政治书共购买网络. (a) 网络中 105 个政治书的共购买关系情形, 不同社团的节点用不同深度的颜色来表示, 其中节点的大小于节点的度呈正比; (b) $\Lambda(\tau)$ 和 $\Theta(\tau)$ 的值与时间 τ 的对应情形, 从中可以看出, 最优的社团数目为 7 和 2

图 6 政治书共购买网络中的多层次社团交互模式. (a) 第一个层次中, 7 个社团的交互示意图; (b) 第二个层次中, 2 个社团的交互示意图

6, 7)”. 这种全部的模式可以被认为是全局性的: 具有共同类型的书, 被共购买的几率会非常高 (大约 24%), 而不同类型的书几乎不可能被共购买 (几率小于 1%).

当社会结构被放大到第一层次时, 我们可以根据社团交互矩阵 C_{ks} 得到一个 7×7 的关系. 例如, 我们降序列举了社团 2 的交互强度 (见定义 4): $\langle C_{23} = 0.58 \rangle$, $\langle C_{21} = 0.31 \rangle$, $\langle C_{24} = 0.11 \rangle$, $\langle C_{27} = 0.04 \rangle$, $\langle C_{25} = 0 \rangle$, $\langle C_{26} = 0 \rangle$. 这样的宏观交互模式可以帮助书商适当地调整书的销售策略和书的购入数量等. 例如, 社团 2 中节点代表的政治书如果销售一空, 他们可以同时增加社团 2 和 3 中节点代表的政治书的数目. 同时, 书商还可以减少社团 5 和 7 中节点代表的政治书的数目, 因为它们和社团 2 节点代表的政治书几乎不可能会被同时购买.

6 结论

本文提出了一种新颖的动态社团探测算法, 利用迭代技术高效而准确地揭示网络中的社团结构. 具体来讲, 我们引入一种新型的基于离散时间的动态系统, 来描述社团归属的从随机状态到最优划分的演化轨迹, 并利用严格的数学分析找出了社团归属动态收敛到最优的条件. 另外, 本文还创新性地提出了一种划分指标函数的一般化形式, 通过选择不同的参数, 可以拓展到很多著名的划分指标函数. 本算法非常高效, 计算复杂度分析显示算法需要的时间与稀疏网络的节点数量呈线性关系. 为了使得算法具有更加广泛的应用, 一些问题值得我们进行更深入地探讨, 例如, (1) 如何进一步加快算法速度, 使得其能够在超大规模的网络上准确划分社团, 如包含几亿到几十亿节点的微博网络和生物网络上; (2) 由于现实世界并不都是双向关系, 网络中会出现单向或者混合方向的边, 如何利用算法高效地处理这种混杂网络, 是一个挑战.

参考文献

- 1 Folino F, Pizzuti C. An evolutionary multiobjective approach for community discovery in dynamic networks. *IEEE Trans Knowl Data Eng*, 2013, 26: 1838–1852
- 2 Gong M G, Cai Q, Chen X W, et al. Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. *IEEE Trans Evol Comput*, 2014, 18: 82–97
- 3 Tang L, Liu H, Zhang J P. Identifying evolving groups in dynamic multimode networks. *IEEE Trans Knowl Data Eng*, 2012, 24: 72–85
- 4 Liu Y, Moser J, Aviyente S. Network community structure detection for directional neural networks inferred from multichannel multi-subject EEG data. *IEEE Trans Biomed Eng*, 2014, 61: 1919–1930
- 5 Wang C D, Lai J H, Yu P S. Neiwalk: Community discovery in dynamic content-based networks. *IEEE Trans Knowl Data Eng*, 2014, 26: 1734–1748
- 6 Tremblay N, Borgnat P. Graph wavelets for multiscale community mining. *IEEE Trans Signal Process*, 2014, 62: 5227–5239
- 7 Rolland T, Tasan M, Charlotiaux B, et al. A proteome-scale map of the human interactome network. *Cell*, 2014, 159: 1212–1226
- 8 Newman M E J. Fast algorithm for detecting community structure in networks. *Phys Rev E* (3), 2004, 69: 066133
- 9 Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Phys Rev E* (3), 2004, 69: 026113
- 10 Zhang X S, Wang R S, Wang Y, et al. Modularity optimization in community detection of complex networks. *Europhys Lett EPL*, 2009, 87: 38002
- 11 Li H J, Wang Y, Wu L Y, et al. Potts model based on a Markov process computation solves the community structure problem effectively. *Phys Rev E* (3), 2012, 86: 016109
- 12 Li H J, Zhang X S. Analysis of stability of community structure across multiple hierarchical levels. *Europhys Lett EPL*, 2013, 103: 58002

- 13 Blondel V D, Guillaume J, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*, 2008, 2008: 10008
- 14 Clauset A, Newman M E J, Moore C. Finding community structure in very large networks. *Phys Rev E* (3), 2004, 70: 066111
- 15 Girvan M, Newman M E J. Community structure in social and biological networks. *Proc Nat Acad Sci India Sect A*, 2002, 99: 7821–7826
- 16 Guimera R, Nunes Amaral L A. Functional cartography of complex metabolic networks. *Nature*, 2005, 433: 895–900
- 17 Hastings M B. Community detection as an inference problem. *Phys Rev E* (3), 2006, 74: 035102
- 18 Reichardt J, Bornholdt S. Statistical mechanics of community detection. *Phys Rev E* (3), 2006, 74: 016110
- 19 Fortunato S, Barthélemy M. Resolution limit in community detection. *Proc Nat Acad Sci India Sect A*, 2007, 104: 36–41
- 20 Hofman J M, Wiggins C H. Bayesian approach to network modularity. *Phys Rev Lett*, 2008, 100: 258701
- 21 Raghavan U N, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E* (3), 2007, 76: 036106
- 22 Ronhovde P, Nussinov Z. Local resolution-limit-free Potts model for community detection. *Phys Rev E* (3), 2010, 81: 046114
- 23 Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification. *J Stat Mech Theory Exp*, 2005, 2005: 09008
- 24 Lancichinetti A, Fortunato S, Radicchi F. Benchmark graphs for testing community detection algorithms. *Phys Rev E* (3), 2008, 78: 046110
- 25 Duch J, Arenas A. Community detection in complex networks using extremal optimization. *Phys Rev E* (3), 2005, 72: 027104
- 26 Boccaletti S, Ivanchenko M, Latora V, et al. Detecting complex network modularity by dynamical clustering. *Phys Rev E* (3), 2007, 75: 045102
- 27 Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Trans Pattern Anal Mach Intell*, 2001, 22: 888–905
- 28 Knuth D E. *The Stanford GraphBase: A Platform for Combinatorial Computing*. Reading: Addison-Wesley, 1993

A linear community detection algorithm based on dynamical system in networks

LI HuiJia, YAN Guan, LIU ZhiDong, LI GuiJun & ZHANG XiangSun

Abstract Detection of communities or clustering is particularly valuable for understanding, analyzing and optimizing many natural and engineering complex networks, such as gene regulatory networks, smart grid and transportation networks. Present techniques relies heavily on the optimization or heuristic methods, which cannot balance the computational efficiency and accuracy. In this paper, we propose an iterative algorithm to realize the exact detection of network communities, by using a novel method based on dynamical system. We first introduce a discrete-time dynamical system that characterizes the evolutionary iteration of community membership from a random configure to the optimal one, and then specify the conditions that can direct the trajectory of this dynamical system to the convergence, which reveals the community label of each node. The computational complexity analysis shows the high-performance of our algorithm: The required computational time is linearly dependent on the total number of nodes in a sparse network. Analyzing the eigenvalue gap of the Markovian transition matrix, a rigorous theory is provided to find the optimal number of communities divided from a network. We also show that the new algorithm can be generalized to unify the conventional algorithms that are widely used. Finally, we perform extensive simulations using both synthetic and real-world benchmark networks to illustrate the nice performance of our method.

Keywords community detection, dynamical systems, hierarchical structure, linear time, hidden features

MSC(2010) 05C82, 35R02, 05C72

doi: 10.1360/N012015-00284