Chapter 4 of Network Science
Degree Correlation and the Community Structure

Presented By LI Zaile

Fudan University
zaiyueli19@fudan.edu.cn

Wang Xiaofan, Li Xiang, Chen Guanrong (2012). Network Science: An Introduction

November 5, 2020

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

What we have learned from the last chapter

- 复杂网络的连通性
- 节点的度与稀疏性
- 平均路径长度与直径
- 聚类系数
- 度分布（频率的视角）
- 幂律分布
- 零阶度分布特性：$\langle k \rangle = 2M/N$
- 一阶度分布特性：$P(k) = n(k)/N$，$n(k)$是网络中度为k的节点数；一阶度分布特性包含了平均度的信息：

$$\langle k \rangle = \sum_{k=0}^{\infty} kP(k)$$

What we will learn from this chapter

一阶性质并不能唯一地刻画一个网络：

- 刻画二阶分布特性（度相关性）的几种不同的方法
  - 一般但复杂的联合概率分布
  - 简洁但不宜比较的条件概率和余平均度
  - 可以定量刻画相关性但过于粗略的相关系数
- 社团结构

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

高阶度分布的引入

具有完全相同的度分布的网络可能具有完全不同的性质或行为，如图4-1

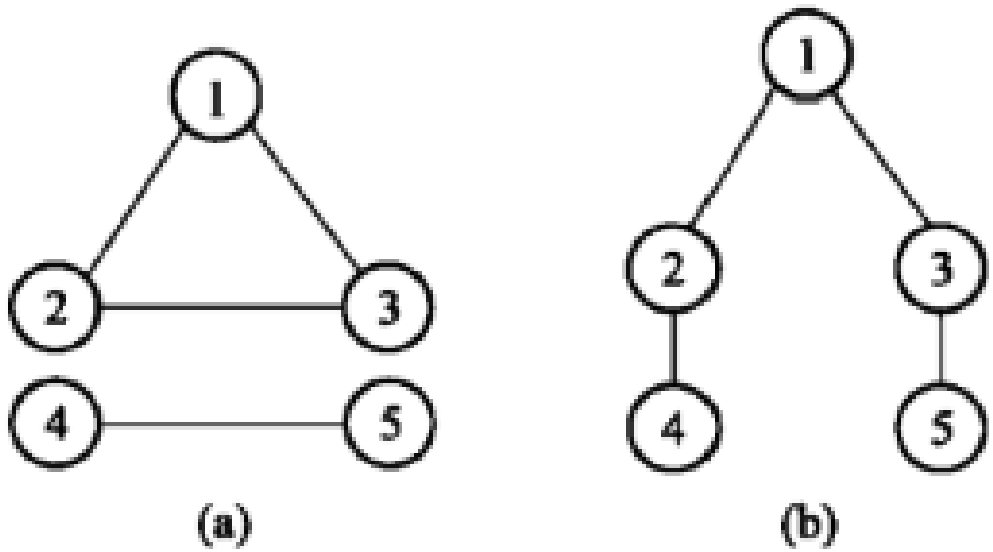


图 4-1　两个具有相同度序列的简单网络

联合概率分布1：定义与性质

定义

随机选取一条边的两端点的度分别为j和k的概率（频率），

$$P(j,k) = \frac{m(j,k)\mu(j,k)}{2M}$$

其中，$m(j,k)$ 是度为 j 的节点和度为 k 的节点之间的连边数；如果 $j = k$，那么 $\mu(j,k) = 2$，否则 $\mu(j,k) = 1$。

性质

- $P(j,k) = P(k,j), \quad \forall j,k$
- $\sum_{j,k=k_{\min}}^{k_{\max}} P(j,k) = 1$
- 余度分布：$P_n(k) = \sum_{j=k_{\min}}^{k_{\max}} P(j,k)$，即 网络中随机选取的一个节点随机选取的一个邻居节点的度为k的概率。

$P_n(k)$与$P(k)$的关系

一般来说，当网络中存在孤立节点时，两者是不相同的，在这种情形下，定义 $P_n(0) \equiv 0 < P(0)$。考虑A-B C的三节点图：

- $P(0) = \frac{1}{3}, P(1) = \frac{2}{3}$
- $P_n(1) = \frac{2}{2}, P_n(0) \equiv 0$

二阶分布特性包含一阶分布特性

$$P_n(k) = \frac{n(k) \times k}{2M} = \frac{N \times P(k) \times k}{2M}$$

为了叙述方便，定义一些记号（度分布、余度分布、联合概率）

$$p_k \triangleq P(k), \quad q_k \triangleq P_n(k), \quad e_{jk} \triangleq P(j,k)$$

$$p_k = \frac{\langle k \rangle}{k} q_k = \frac{\langle k \rangle}{k} \sum_{j=k_{\min}}^{k_{\max}} e_{jk}.$$

联合概率分布2：相关性与同配性

“度不相关”的定义

$$e_{jk} = q_j q_k, \quad \forall j, k$$

网络不具有度相关性 or 网络是中性的：网络中随机选择的一条边的两个端点的度是完全随机的。

度相关网络的分类：正相关与负相关

- 同配 (Assortative)：度大的节点倾向于连接度较大的节点
- 异配 (Disassortative)：度小的节点倾向于连接度小的节点

(a) 同配网络　(b) 中性网络　(c) 异配网络

图 4-2　具有相同度序列和不同同配性质的 3 个网络

同配网络的社会性和异配网络的自然性

现实社会中的同配网络与自然和技术中的异配网络

- 蛋白质交互网络、（真）神经网络、万维网
- 科研人员合作网络、电影演员网络等（人以类聚、职业合作、不可替代、同一领域、共同兴趣，组织）

在线社交平台对社交网络同配性的影响

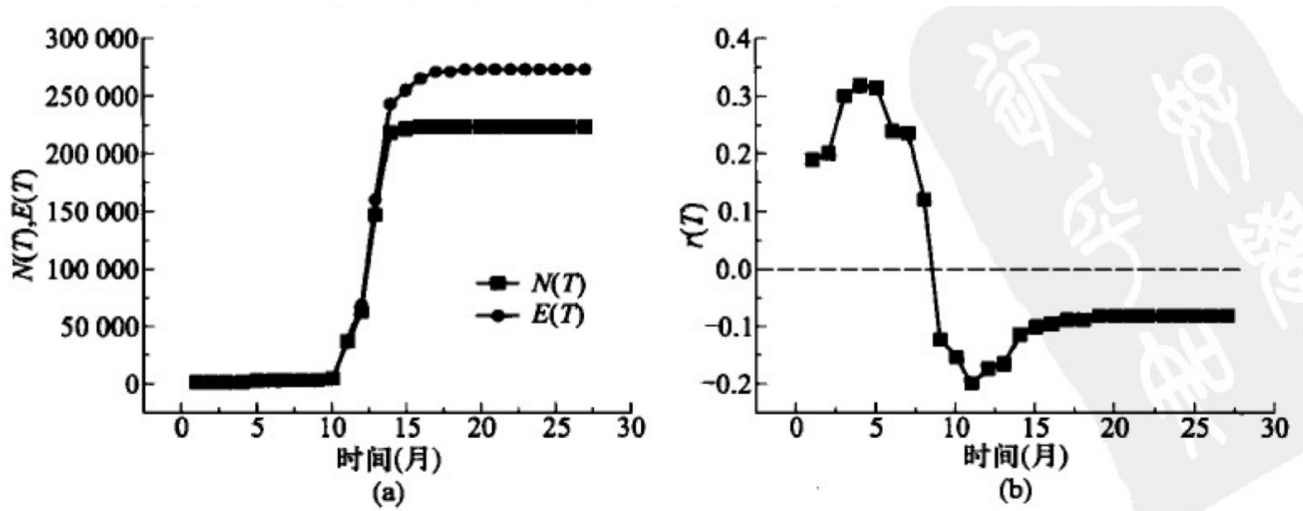


图 4-6　一个在线社会网络的演化（取自文献[8]）

从同配到异配：熟人社交->名人社交（关注、点赞等）

接下来要考虑的几个问题

同配 or 异配的意义

已有研究表明,网络的同配或异配对网络结构和行为如鲁棒性和传播等可 能有显著的影响。

如何判断是否为度相关

- 度相关的定义
- ?

判断度相关性的类别

判断度相关性的强弱系数（网络之间可比较）

判断是否存在度相关性：条件概率

网络中随机选取一个度为 k 的节点的一个邻居的度为 j 的概率

$$P_c(j \mid k) = \frac{P(j,k)}{P_n(k)} = \frac{e_{jk}}{q_k}$$

根据“中性”的定义，当 $\frac{e_{jk}}{q_k} = q_j$ 与k无关，即 $P_c(j \mid k)$ 与k无关，网络为中性；如果与k有关，则为度相关。

判断度相关性的类别：余平均度

某节点i的余平均度$\langle k_{nn} \rangle_i$：邻居节点的平均度

假设节点 i 的 k_i 个邻居节点的度为 $k_{i_j}, j = 1, 2, \ldots, k_i$ ，则该节点的 余平均度为：$\langle k_{nn} \rangle_i = \frac{1}{k_i} \sum_{j=1}^{k_i} k_{i_j}$

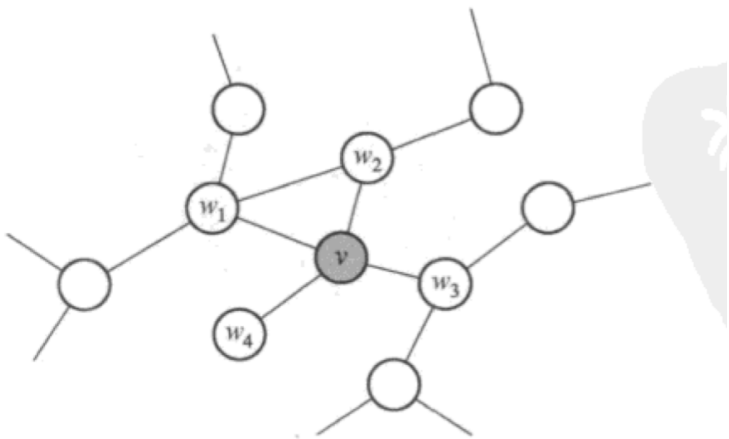


图 4-4　余平均度计算示意图

$$\langle k_{nn} \rangle_v = \frac{4+3+3+1}{4} = \frac{11}{4}$$

k-余平均度：度为k的所有节点的余平均度的平均

假设网络中度为 k 的节点为 $v_1, v_2, \ldots, v_{i_k}$，k-余平均度为：

$$\langle k_{nn} \rangle (k) = \frac{1}{i_k} \sum_{i=1}^{i_k} \langle k_{nn} \rangle_{v_i}$$

k-余平均度作为一个期望：与条件概率的关系

条件概率：随机一个度为k 的节点的一个邻居的度为j的概率

$$\langle k_{nn} \rangle (k) = \sum_{k'=k_{\min}}^{k_{\max}} k' P_c \left(k' \mid k \right) = \frac{1}{q_k} \sum_{=k_{\min}}^{k_{\max}} k' e_{k'k}$$

判别规则

- 同配：$\langle k_{nn} \rangle (k)$ 是 k 的增函数
- 异配：$\langle k_{nn} \rangle (k)$ 是 k 的减函数

不具有度相关性时的期望余平均度

网络不具有度相关性,那么 $\langle k_{nn} \rangle (k)$ 是一个与 k 无关的常数：

$$\langle k_{nn} \rangle (k) = \frac{1}{q_k} \sum_{=k_{\min}}^{k_{\max}} k' e_{k'k} = \frac{\sum_j j q_j q_k}{q_k} = \sum_j j q_j = \sum_j j \frac{j p_j}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle}$$

无法判断同配的程度，也无法对不同网络（或同一网络在不同的时间）进行比较

来自Facebook的例子：友谊悖论（Friendship paradox）

- 除非你的朋友数目超过700，否则你的朋友的朋友比你的朋友多
- 对于条件概率分布，当k比较大时，曲线峰值右移，同样说明同配

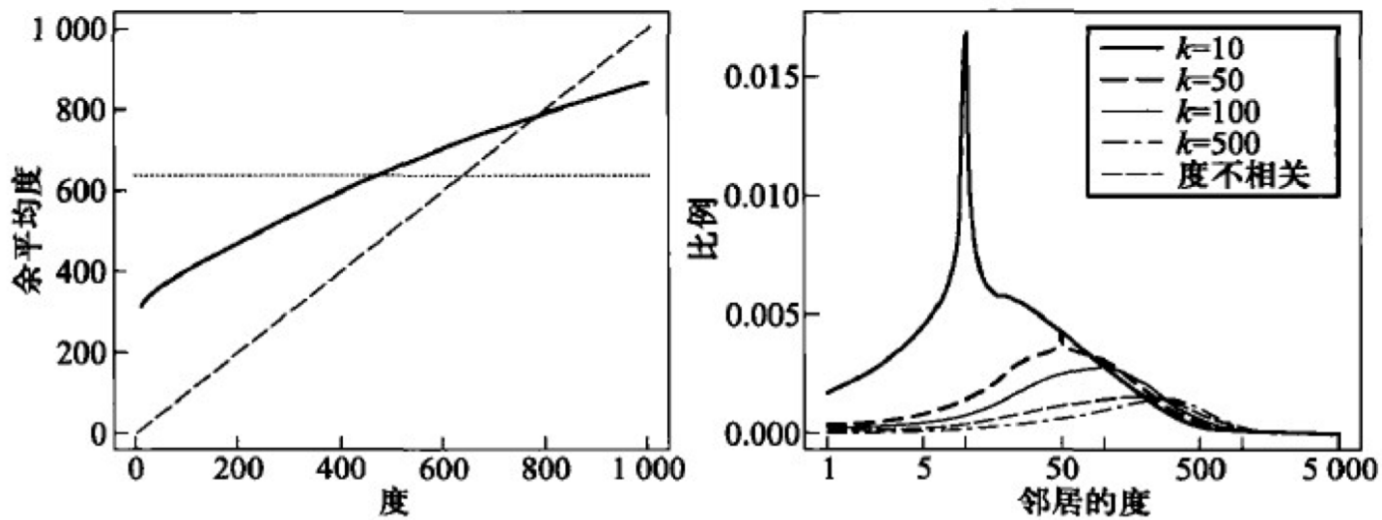


图 4-5　Facebook 网络的余平均度和条件概率分布(取自文献[5])

同配系数

度相关系数

在度相关的情况下，有$e_{jk} - q_j q_k \neq 0$ for some j and k，于是定义

$$\langle jk \rangle - \langle j \rangle \langle k \rangle = \sum_{j,k} jk \left(e_{jk} - q_j q_k \right)$$

作为一个与j和k相关的量，度相关系数大小与网络的规模相关。

同配系数：归一化后的度相关系数

$$r = \frac{1}{\sigma_q^2} \sum_{j,k} jk \left(e_{jk} - q_j q_k \right)$$

其中σ_q^2是可能的最大度相关系数，即网络完全同配：$e_{jk} = q_k \delta_{jk}$ $\sigma_q^2 = \sum_k k^2 q_k^2 - \left[\sum_k k q_k \right]^2$。 $r \in [-1, 1]$ 。如果 $r > 0$, 那么网络是同配的。

同配概念的一般化

度的同配以及社交网络分析中的属性-同质性（homophily）

- 同配：属性相近的节点倾向于互相连接

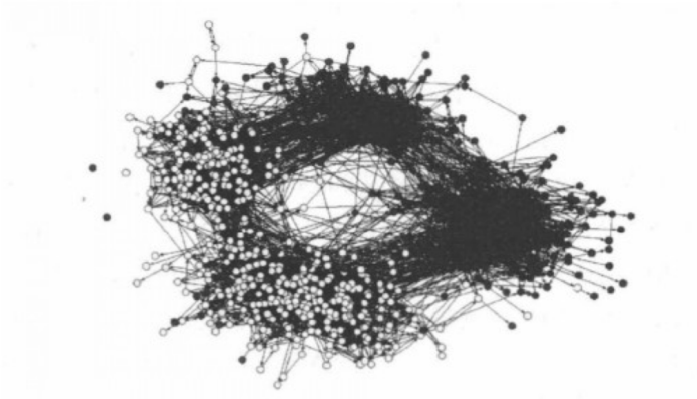


图 4-7　一个中学生朋友关系网络

图片来源：http://www-personal.umich.edu/~mejn/networks/

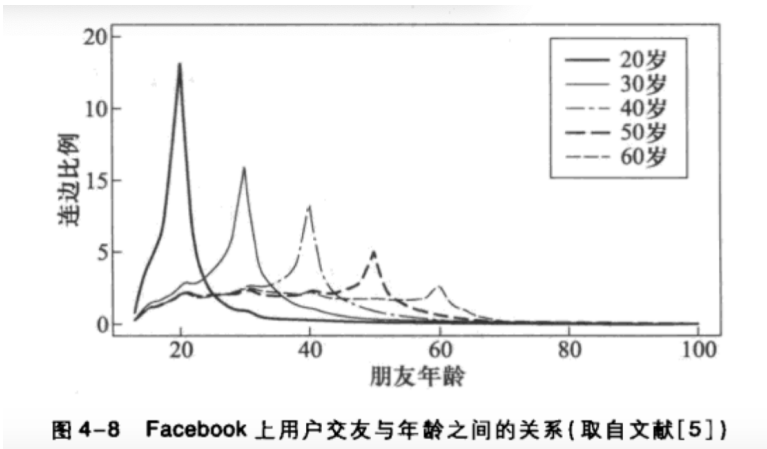


图 4-8　Facebook 上用户交友与年龄之间的关系（取自文献[5]）

- 社会网络同质性：选择（人以类聚）V.S. 影响（近朱者赤）
- 时变数据+演化建模：音乐+电影共同品味更容易成为朋友

度同配系数在其他属性上的推广*

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

社团结构的描述

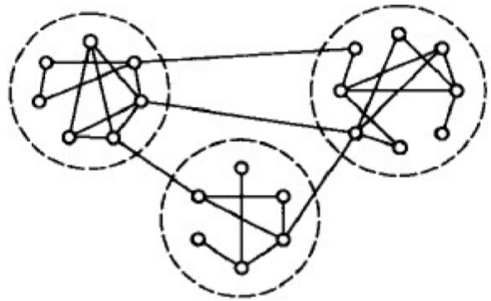


图 4-9　一个小型的具有社团结构性质的网络

图分割与并行计算

n个互相通信的程序在k个处理器上运行， 如何分配这 n 个程 序到 g 个处理器上, 使得每个处理器负载近似相等,同时处理器之间连接的边数最少,从而使得各个处理器之间的通信量最小。

社会学中的分层聚类*

分级聚类是寻找社会网络中社团结构的一类传统算法。 它基于各个节点之 间连接的相似性或者强度,把网络自然地划分为各个子群。 根据向网络中添加 边还是从网 络中移除边,该类算法又可以分为两类：凝 聚方 法（Agglomerative method) 和分裂方法(Divisive method)

衡量社团划分质量：模块度

一个网络相应的零模型（null model）

- 零模型：与该网络具有某些相同性质的随机图
- 零阶零模型：与原网络具有相同边数，且具有均匀度分布
- 一阶零模型：具有相 同度序列的随机图

Q值：所有社团内部的边数的总和

$$Q_{\text{real}} = \frac{1}{2}\sum_{i} a_{ij}\delta(C_i, C_j) \quad Q_{\text{null}} = \frac{1}{2}\sum_{ij} p_{ij}\delta(C_i, C_j)$$

$A = (a_{ij})$ 是实际网络的邻接矩阵，C_i为i节点所属社团，δ为一示性函数。 p_{ij} 是零模型中节点 i 和节点 j 之间的边数的期望值。

（无向无权网络的）模块度

$$Q = \frac{Q_{\text{real}} - Q_{\text{null}}}{M} = \frac{1}{2M}\sum_{ij}(a_{ij} - p_{ij})\,\delta(C_i, C_j)$$

考虑一个与原网络具有相同度序列但不具有度相关性的零模型-配置模型，有 $p_{ij} = k_i k_j/(2M)$，

$$Q = \frac{1}{2M}\sum_{ij}\left(a_{ij} - \frac{k_i k_j}{2M}\right)\delta(C_i, C_j) = \frac{1}{2M}\sum_{ij} b_{ij}\delta(C_i, C_j)$$

$$b_{ij} = a_{ij} - \frac{k_i k_j}{2M}$$

$\boldsymbol{B} = (b_{ij})_{N\times N}$ 也称为模块度矩阵(Modularity matrix)。

基于边数据的模块度的计算

- e_{vw} 为社团 v 和社团 w 之间的连边占整个网络边数的比例

$$e_{vw} = \frac{1}{2M} \sum_{ij} a_{ij} \delta\left(C_i, v\right) \delta\left(C_j, w\right)$$

- a_v 为一端与社团 v 中节点相连的连边的比例（重要！）

$$a_v = \frac{1}{2M} \sum_{i} k_i \delta\left(C_i, v\right)$$

注意到，

$$\delta\left(C_i, C_j\right) = \sum_{v} \delta\left(C_i, v\right) \delta\left(C_j, v\right)$$

于是有：

$$Q = \sum_{v} \left[e_{vv} - a_v^2\right] \quad \left(Q = \frac{Q_{\text{real}}：\text{社团内部边数} - Q_{\text{null}}}{M}\right)$$

一些说明

$$
\begin{aligned}
Q &= \frac{1}{2M}\sum_{ij}\left(a_{ij}-\frac{k_i k_j}{2M}\right)\sum_{v}\delta\left(C_i,v\right)\delta\left(C_j,v\right) \\
&= \sum_{v}\left[\frac{1}{2M}\sum_{ij}a_{ij}\delta\left(C_i,v\right)\delta\left(C_j,v\right)-\right. \\
&\left.= \frac{1}{2M}\sum_{i}k_i\delta\left(C_i,v\right)\frac{1}{2M}\sum_{j}k_j\delta\left(C_j,v\right)\right] \\
&= \sum_{v}\left[e_{vv}-a_v^2\right]
\end{aligned}
$$

或者，

$$
Q=\sum_{v=1}^{n_l}\left[\frac{l_v}{M}-\left(\frac{d_v}{2M}\right)^2\right]
$$

其中 n_c 是社团的数量，l_v 是社团 v 内部所包含的边数，d_v 是社团 v 中所有节点的 度值之和。

关于模块度

- 把整个网络视为一个社团,对应的模块度为零：$Q_{\text{real}} = Q_{\text{null}}$
- 每一个节点视为一个社团,模块度为负：$Q_{\text{real}} = 0, Q_{\text{null}} \neq 0$
- 最优分割：模块度值 $Q = Q_{\max}, 0 \leqslant Q_{\max} < 1$

$$Q_{\text{null}} \neq 0 => Q = \frac{Q_{\text{real}} - Q_{\text{null}}}{M} < 1$$

- 一般来说，模块度都在0.3-0.7之间
- 模块度的缺陷： 注意到规模较大的网络所对应的 $Q_{\max}$ 通常也较大, 因此不能简单通过模块度的 大小来比较不同规模网络的社团划分的质量。

加权和有向网络的模块度

加权网络

$$Q_w = \frac{1}{2W}\sum_{ij}\left(w_{ij} - \frac{s_i s_j}{2W}\right)\delta\left(C_i, C_j\right) = \sum_{c=1}^{n}\left[\frac{W_c}{W} - \left(\frac{S_c}{2W}\right)^2\right]$$

有向网络

$$Q_d = \frac{1}{M}\sum_{ij}\left(a_{ij} - \frac{k_i^{out} k_j^{in}}{M}\right)\delta\left(C_i, C_j\right)$$

加权有向网络

$$Q_{wd} = \frac{1}{W}\sum_{ij}\left(w_{ij} - \frac{s_i^{\text{out}}\ s_j^{in}}{W}\right)\delta\left(C_i, C_j\right)$$

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

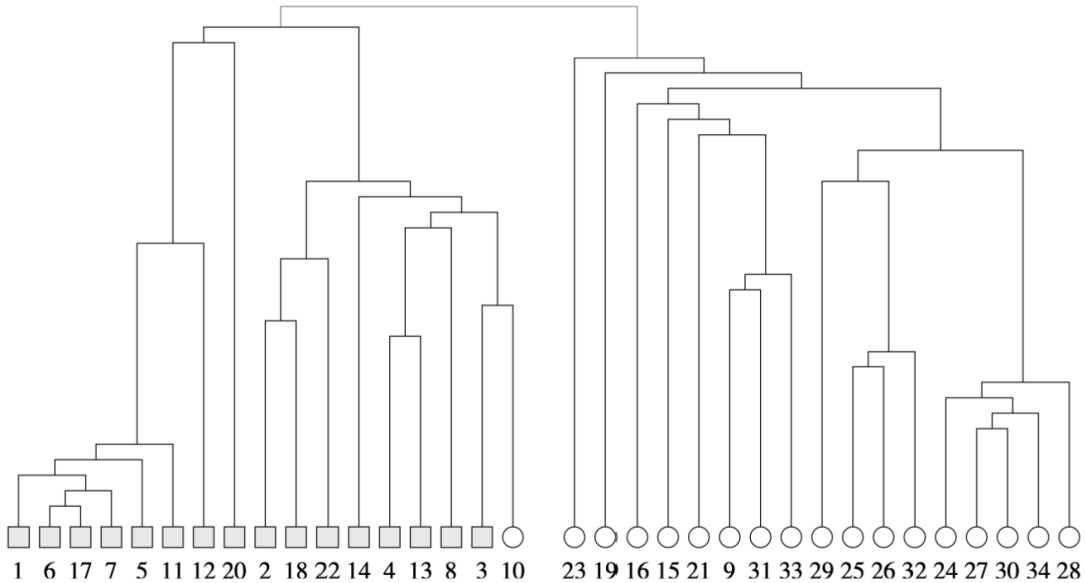
Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

A Naive Greedy Algorithm: Newman (Phys. Rev. E, 2004)

Agglomerative hierarchical clustering methods

Dendrogram



Idea of Newman's Algorithm

- Initially set every node as a community
- Iteratively join a pair of communities with the biggest ΔQ

The Newman's Algorithm

Since the joining of a pair of communities between which there are no edges at all can never result in an increase in Q, we need only consider those pairs between which there are edges, of which there will at any time be at most m, where m is again the number of edges in the graph. The change in Q upon joining two communities is given by $\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j)$, which can clearly be calculated in constant time. Following a join, some of the matrix elements e_{ij} must be updated by adding together the rows and columns corresponding to the joined communities, which takes worst-case time $\mathrm{O}(n)$. Thus each step of the algorithm takes worst-case time $\mathrm{O}(m + n)$. There are a maximum of $n-1$ join operations necessary to construct the complete dendrogram and hence the entire algorithm runs in time $\mathrm{O}((m+n)n)$, or $\mathrm{O}(n^2)$ on a sparse graph. The algorithm has the added advantage of calculating the value of Q as it goes along, making it especially simple to find the optimal community structure.

It is worth noting that our algorithm can be generalized trivially to weighted networks in which each edge has a numeric strength associated with it, by making the initial values of the matrix elements e_{ij} equal to those strengths, rather than just zero or one; otherwise the algorithm is as above and has the same running time. The networks studied in this paper however are all unweighted.

The Newman's Algorithm

$$a_v = \frac{1}{2M}\sum_i k_i \delta(C_i, v) \quad e_{vw} = \frac{1}{2M}\sum_{ij} a_{ij}\delta(C_i, v)\,\delta(C_j, w)$$

$$Q = \sum_v \left[e_{vv} - a_v^2\right]$$

- 将网络每个节点看作一个社团，标记为$1, 2, 3\cdots, N$
- 初始的网络模块度为0；计算初始社团间和社团特征(e_{ij}, a_i):

$$e_{ij} = \begin{cases} 1/(2M), & \text{如果节点（社团） } i \text{ 和 } j \text{ 之间有边相连} \\ 0, & \text{其他} \end{cases}$$

$$a_i = \sum_j e_{ij} = k_i/(2M)$$

- 如果将社团i和社团j聚合，产生的社团记为j'：

$$\begin{aligned}\Delta Q &= \sum_{v \neq i,j} \left[e_{vv} - a_v^2\right] + \left[e_{j'j'} - a_{j'}^2\right] - \sum_{v} \left[e_{vv} - a_v^2\right] \\ &= \left[e_{j'j'} - a_{j'}^2\right] - e_{ii} - e_{jj} + a_i^2 + a_j^2 \\ &= e_{ii} + e_{jj} + e_{ij} + e_{ji} + (a_i + a_j)^2 - e_{ii} - e_{jj} + a_i^2 + a_j^2 \\ &= 2e_{ij} - 2a_i a_j\end{aligned}$$

if社团i和社团j无连接，不考虑聚合这一pair（增量小于0）

- 模块度增量矩阵元素：

$$\Delta \boldsymbol{Q}_{ij} = \begin{cases} 2e_{ij} - 2a_i a_j, & \text{如果节点 } i \text{ 和 } j \text{ 相连} \\ 0, & \text{其他} \end{cases}$$

- 选取最大增量pair进行聚合；迭代更新（上面数据也需更新）：

$$\boldsymbol{a}'_j = \boldsymbol{a}_i + \boldsymbol{a}_j, \quad a'_i = 0$$

$$\boldsymbol{Q} = \boldsymbol{Q} + \Delta \boldsymbol{Q}_{ij}$$

Complexity of the Newman's Algorithm: $\mathrm{O}((m+n)n)$

- m (number of edges) pairs to compare
- Following a join, some e_{ij} must be updated: O(n)
- At most n (number of nodes) iterations to stop the algorithm

CNM（Clauset, Newman, Moore）算法: $O(md\log n)$

提升来自于两方面：

- 寻找最大模块度增量：平衡树存储数据，最大堆寻找最大值
- 更新社团特征数据以及模块度增量：

合并社团i 和 j后，标记为j；并更新模块度增量矩阵 $\Delta \boldsymbol{Q}_{ij}$：

- 删除第i行和第i列的元素,更新第 j 行和第 j 列的元素, 得到

$$\Delta \boldsymbol{Q}'_{jk} = \begin{cases} \Delta \boldsymbol{Q}_{ik} + \Delta \boldsymbol{Q}_{jk}, & \text{社团 } k \text{ 与社寸 } i \text{ 和社团 } j \text{ 都相连} \\ \Delta Q_{ik} - 2a_j a_k, & \text{社团 } k \text{与社团} i \text{ 相连,不与社团 } j \text{ 相连} \\ \Delta Q_{jk} - 2a_i a_k, & \text{社团 } k \text{与社团} j \text{ 相连,不与社团 } i \text{ 相连} \end{cases}$$

在算法整个过程中,模块度 Q 仅有一个最大的峰值。当模块度增量矩阵中 最大的元素都小于零以后, Q 值就只可能一直下降了。所以,只要模块度增量矩 阵中最大的元素由正变为负, 就可以停止合并,并认为此时的结果就是网络的社 团结构。

来自亚马逊的例子

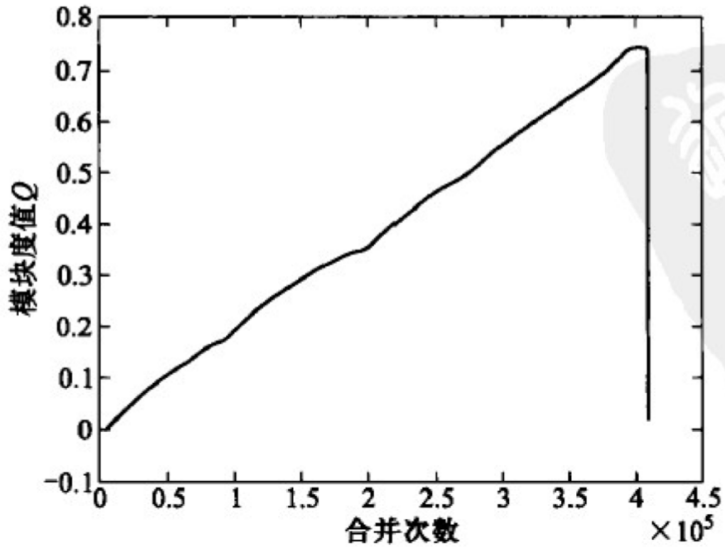
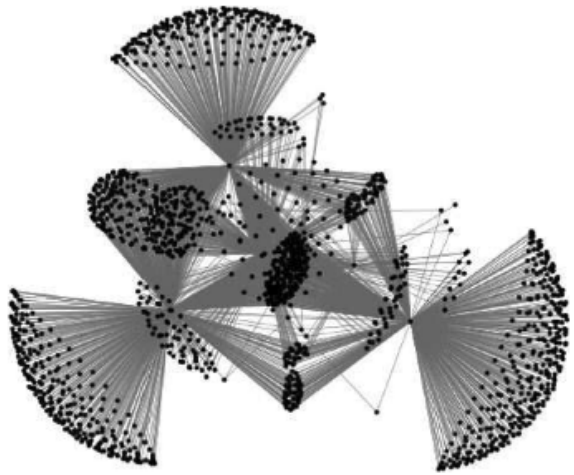


图 4-10　CNM 算法节点合并过程中的模块度的变化(取自文献[15])



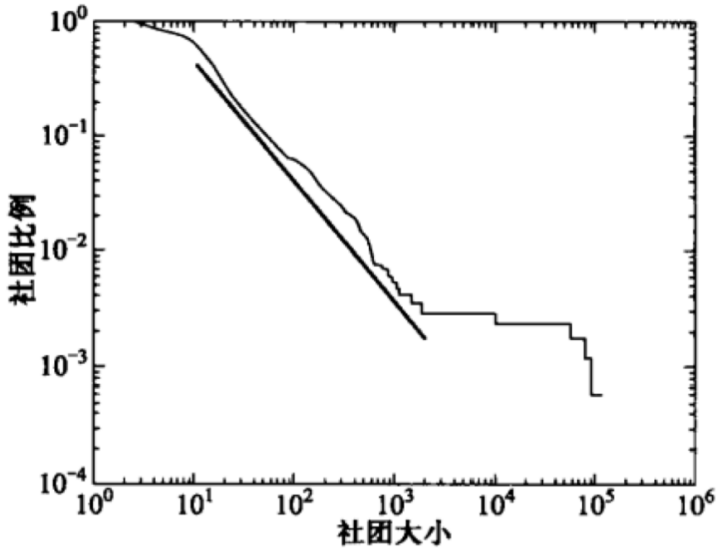


图 4-11　社团规模的累积分布(取自文献[15])

层次化社团检测—BGLL算法

基于模块度的加权网络的层次化社团结构分析的聚类算法

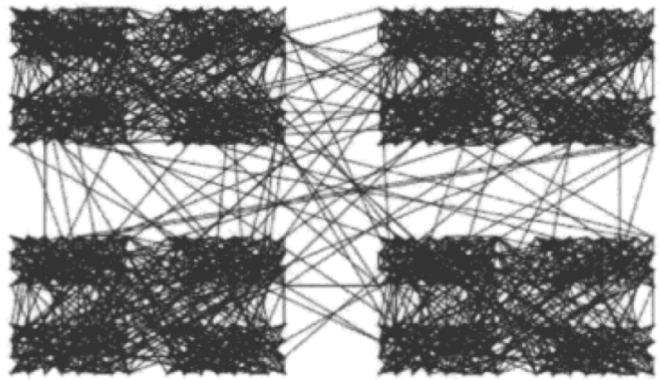


图 4–12　层次化社团结构示意

阶段1：基层聚类

初始每个节点都是一个独立的社团。对任意相邻节点i和节点j计算将节点i加入其邻居节点j所在社团C时对应的模块度增量ΔQ:

$$\Delta Q = \left[\frac{W_c + s_{i,in}}{2W} - \left(\frac{S_c + s_i}{2W}\right)^2\right] - \left[\frac{W_c}{2W} - \left(\frac{S_c}{2W}\right)^2 - \left(\frac{s_i}{2W}\right)^2\right]$$

$s_{i,in}$是节点i与C内其他节点所有边权和。取最大正增量处聚合。

阶段2：基层社团的聚类

构造一个新网络，其中的节点是前一阶段划分出的社团,节点之间连边的权重是两个社团之间所有连边的权重和。然后再利用阶段1对新网络进行社团划分，得到第二层社团结构。以此类推。

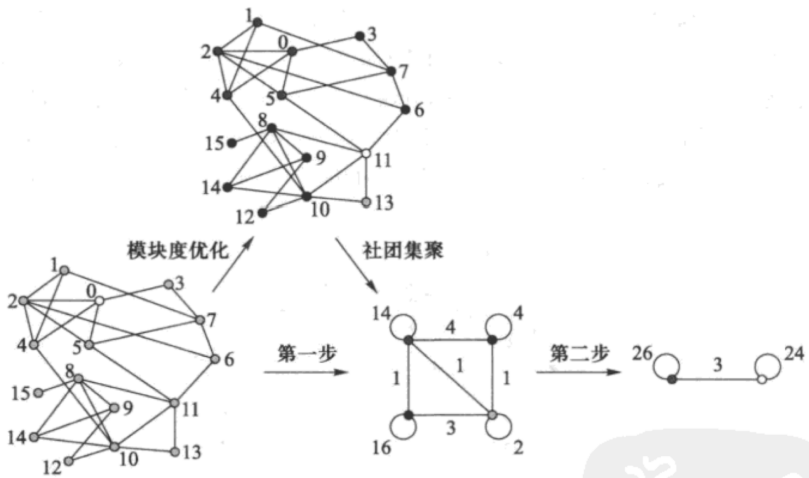


图 4-13　BGLL 算法的处理过程(取自文献[16])

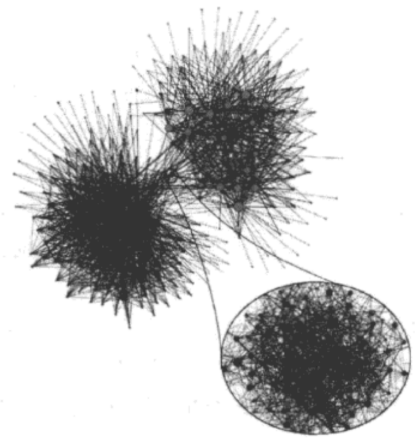


图 4-14　基于 BGLL 算法的实际网络社团分析结果(取自文献[16])

多片网络社团检测

多片网络（Multislice network）

- Time-dependent network
- Multiplex network：多种连接形式
- Multiscale network：不同尺度社团结构

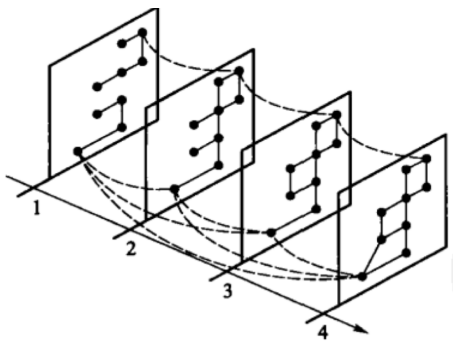


图 4-15　多片网络结构(取自文献[17])

- 各片之间有先后次序关系的多片网络：美国参议员关系网络
- 各片之间并无先后次序关系的多片网络：基于不同的关系类型定义而得到的不同的关系网络

多片加权网络的模块度公式

- w_{ijp}：第p片上节点i与节点j之间的连接权重
- c_{ipq}：节点i在第p片与第q之间的连接的权重
- 片上强度 $s_{ip} = \sum_j w_{ijp}$
- 片间强度 $c_{ip} = \sum_q c_{ipq}$
- 总强度 $w_{ip} = s_{ip} + c_{ip}$
- p 片上所有节点的强度之和：$W_p = \sum s_{i_p}$
- 所有片上的所有节点的总强度之和：$2\mu = \sum_{ip} w_{ip}$

模块度为：

$$Q_{\text{mulitilice}} = \frac{1}{2\mu} \sum_{ijpq} \left[\left(w_{ijp} - \gamma_p \frac{s_{ip} s_{jp}}{2W_p} \right) \delta_{pq} + c_{jpq} \delta_{ij} \right] \delta\left(C_{ip}, C_{jq}\right)$$

γ_p 是用来控制各片网络内社团划分规模和数量的分辨率系数。

空间网络社团检测

现实中的空间聚类

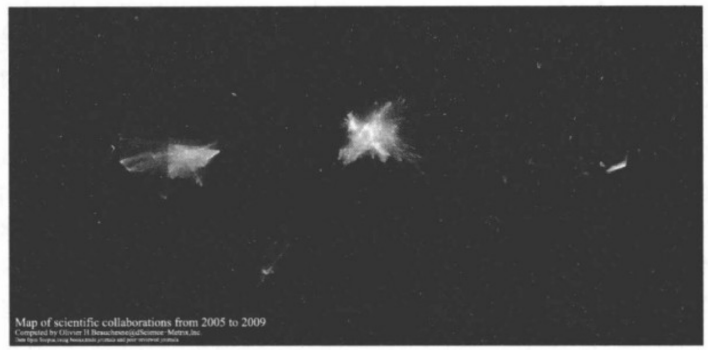


图 1-18　全球科研人员合作网络

图片来源：http://scimaps. org/submissions/7-digital_libraries/maps/thumbs/002_LG. jpg

考虑距离的零模型

$$p_{ij} = k_i k_j/(2M) => p_{ij}^{sa} = N_i N_j f\left(d_{ij}\right)$$

N_i度量节点i的重要性，d_{ij}为节点i和节点j之间的物理距离。对于零模型，相隔一定距离的节点之间的总的权值应保持不变，

$$\sum_{ij|d_{ij}=d} p_{ij}^{spa} = \sum_{ij|d_{ij}=d} a_{ij} => f(d) = \frac{\sum_{ij|d_{ij}=d} a_{ij}}{\sum_{ij|d_{ij}=d} N_i N_j}$$

模块度的局限性

最大模块度$0 \leqslant Q_{\max} < 1$社团划分的质量无法保证

- 任意随机产生的一个网络都具有正的 $Q_{\max}$
- 一种合理方式：把一个网络与该网络的随机化模型做对比
  - 随机重连方式生成许多具有相同度序列的随机化网络
  - 随机网络的模块度均值和方差,分别记为 $\langle Q \rangle_{NM}$ 和 δ_{ℓ}^{NM}
  - 统计重要性： $z_Q = \frac{Q_{\max} - \langle Q \rangle_{NM}}{\delta_Q^{NM}}$

分辨率限制

基于模块度优化的算法用于实际网络就很有 可能无法识别出许多实际存在的小规模的社团。 引入分辨率参数也不work，因为这个参数未知。

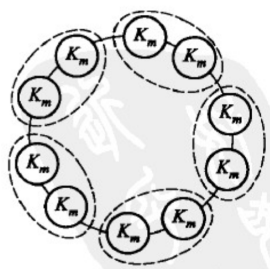


图 4-16 模块度优化的分辨率限制

社团检测算法的评价标准

- 计算复杂性
- 社团划分的performance（无监督问题）

基准图方法：公认图方法

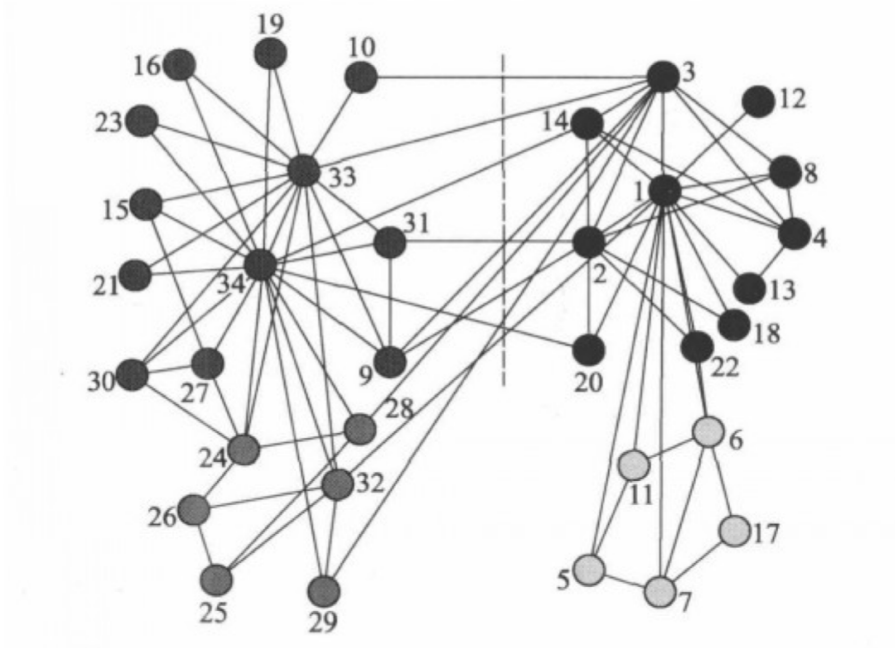


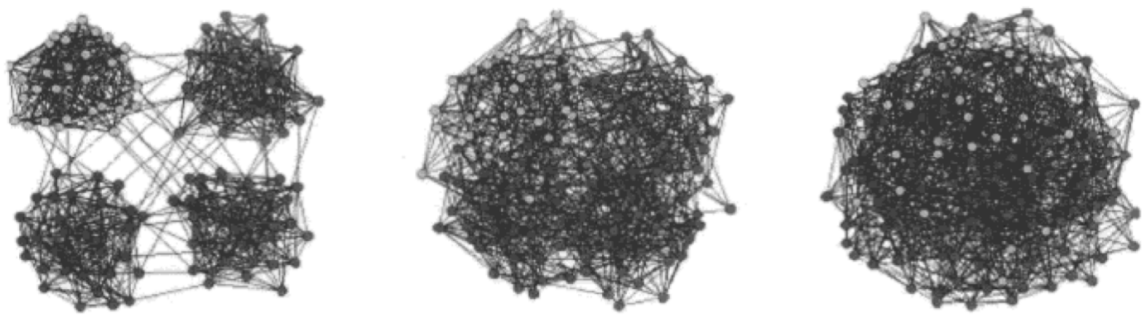
图 4-27　Zachary 空手道俱乐部网络

基准图方法：（生成）预设l-划分模型

- set $N = g \cdot l$个节点,分为l组,每组包含g个节点
- 随机连接：同组两节点连接概率p_{in}，异组两节点概率p_{out}
- 组内为一ER随机图，度值分布均匀
- 网络平均度：$\langle k \rangle = p_{in}(g-1) + p_{\text{out}} g(l-1)$
- $p_{\text{in}} - p_{\text{out}} > 0$ 意味着社团结构
- 主要参数：组内平均度+组外平均度

$$z_{in} = p_{in}(g-1), \quad z_{\text{out}} = p_{\text{out}} g(l-1)$$

- 缺陷：均匀度分布、三角形等模块较少，不符合实际网络



(a) Z_{in}=15　(b) Z_{in}=11　(c) Z_{in}=8

图 4-28　三种不同参数下的基准图

元数据方法

关于节点的描述：

- 构造网络的信息：i.e. 如何连接节点
- （部分）节点（共有）的元数据：e.g. 图书分类、标签

基于元数据的社团划分指标

- 社团质量—节点对相似度的富裕度

$$\frac{\langle \mu(i,j) \rangle_{\text{同一社团中所有的}i,j}}{\langle \mu(i,j) \rangle_{\text{网络中所有节点对}i,j}}$$

$\mu(i,j)$是基于元数据的节点i和j之间的相似度

- 重叠质量：节点所属的社团数目和元数据中的重叠信息之间的交互信息
- 社团覆盖：属于非平凡社团的节点所占的比例
- 重叠覆盖（重叠社团检测）：每个节点所属的非平凡社团的数的平均值
- 复合性能（Composite performance）：四个指标归一化之和

Outline

Motivation and Introduction

Degree Correlation and Assortative Structure

Modularity and the Community Structure

Community Detection Algorithms based on Modularity

Other Community Detection Algorithms

派系过滤算法CPM：clique perlocation method

层次结构之外：骑墙节点与社团的重叠

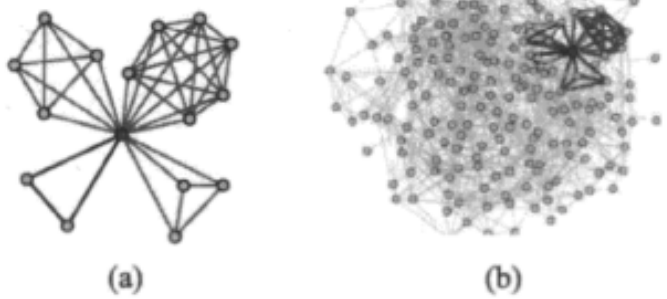


图 4-17　具有重叠性的社团结构示意

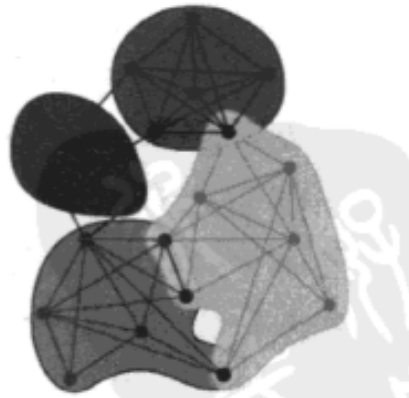


图 4-18　重叠的 4-派系社团
（取自文献[21]）

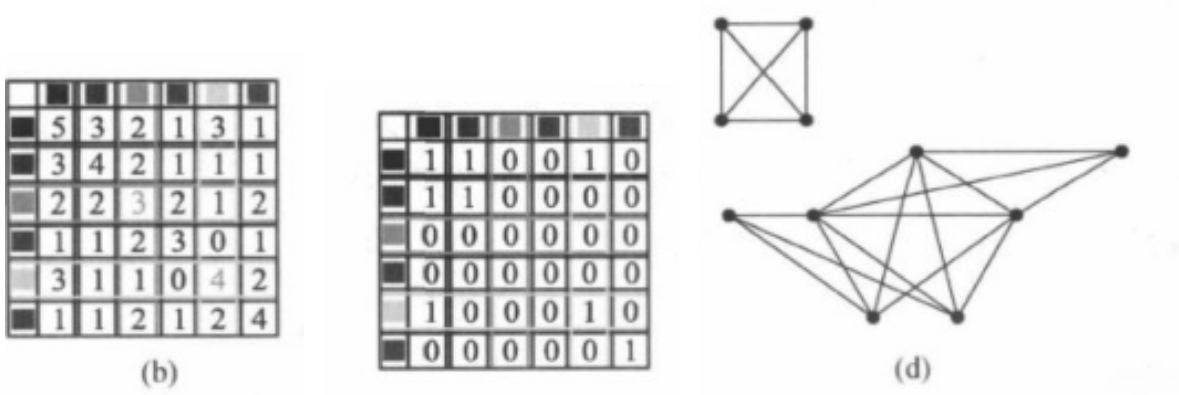
派系过滤算法的notation

- k-派系：包含k个节点的全耦合子图
- k-派系是相邻的：两个k-派系有$k-1$个公共节点
- 两个k-派系连通：可以通过若干个相邻的k-派系到达
- k-派系社团：彼此联通的k—派系构成的集合
- k-派系社团的重叠：节点属于多个不相邻的k-浪系

派系过滤算法step 1：直接搜索网络中的派系

1. 确定可能存在的最大全耦合子图的大小$s = k_{\max} + 1$
2. 选择某一节点，搜索所有s-全耦合子图
   A为包括节点v在内的两两相连的所有节点集合， B为与A中各节点都相连的节点的集合：
   - (1) 初始集合 $A = \{v\}, B = \{v$ 的邻居 $\}$;
   - (2) 从集合 B 中移动一个节点到集合 A, 同时删除集合 B 中不再与集合 A 中 所有节点相连的节点 ;
   - (3) 如果在集合 A 的大小未达到 s 之前,集合 B 已为空集 ,或者集合 A 和 B 为 已有的一个较大的派系中的子集,则停止计算。
   - (4) 当集合 A 的大小达到 s,就得到一个新的派系,记录该派系,然后返回上一步,继续寻找包含 门点 v 的新的派系。
3. 删除上述节点及其边，搜索所有s-全耦合子图，直至完成
4. s=s-1，重复2-3步骤，直到所有派系搜索完成

派系过滤算法step 2：利用派系寻找k-派系社团



- 构造派系重叠矩阵：
  - 每一行（列）对应一个派系，对角线元素代表派系大小
  - 非对角线元素代表两个派系的公共节点数目
- 从派系重叠矩阵中得到k-派系社团邻接矩阵：
  对角线上小于k而非对角线上小于$k-1$的元素置为0,其他元素设置为1
- 科学家合作网络与移动手机用户网络： 大的社团如果内部人员动态变化,反而能够使社团维持更长的时间，小社团则不然。

连边社团检测算法：重叠性+层次性

从连点社团到连边社团

- 连点社团的缺陷：重叠性使得无法用单个连点树图（dendrogram）表达网络层次结构

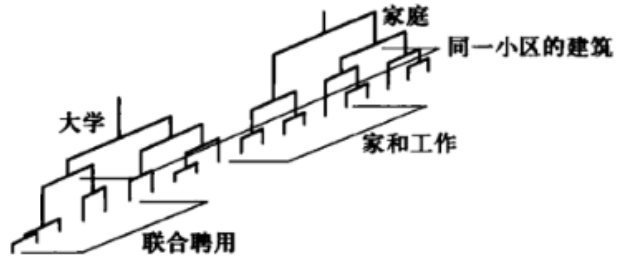


图 4-21　单个点树难以完整表示网络的层次结构(取自文献[23])

- 连边社团的好处：
  - 一条边只能属于一个社团，因此连边树图（dendrogram）可以表达网络层次结构
  - 重叠节点问题fixed
  - 通过不同阈值分割dendrogram，可以得到层次化社团结构

基于边相似度合并的连边社团检测

- 两个边相似的前提：存在公共节点，记边为e_{ik}, e_{jk}
- 基于节点对i与j的相似度的边的相似度定义：
  共同邻居相对数量：

$$S(e_{ik}, e_{jk}) = \frac{|n_+(i) \cap n_+(j)|}{|n_+(i) \cup n_+(j)|}$$

  $n_+(i)$ 为节点 i 及其所有邻居节点的集合。
- 分级聚类思想进行边的合并：
  1. 计算所有相连的连边对的相似度，降序排序
  2. 按排序次序依次合并相应边，表达为dendrogram的形式

算法结果：社团划分、dendrogram、连边相似度矩阵

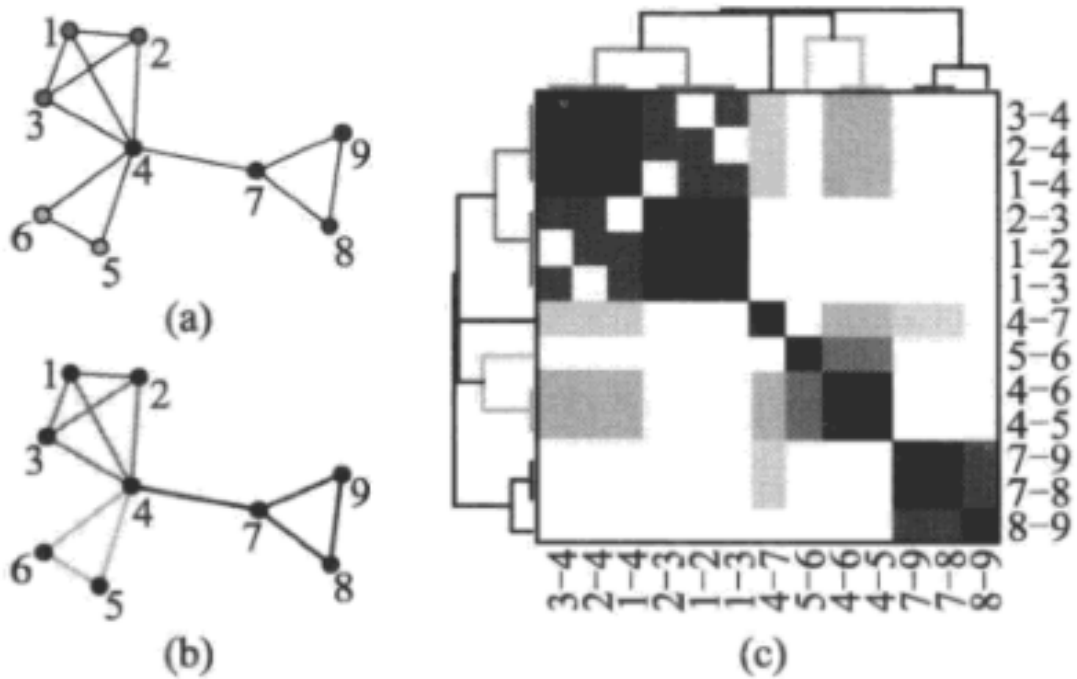


图 4-23　一个简单网络的连边社团检测(取自文献[23])

寻找分割树图的最佳分割位置

- 定义目标函数—划分密度D：
  - 网络中边树为M，划分为C个社团$\{P_1, P_2, \ldots, P_C\}$
  - 社团 P, 包含 m_c 条连边和 n_c 个节点
  - 归一化密度：

$$D_c = \frac{m_c - (n_c - 1)}{n_c (n_c - 1)/2 - (n_c - 1)}$$

  - 整个网络的划分密度：

$$D = \frac{1}{M} \sum_c m_c D_c$$

  - 在层次聚类过程中跟踪D，选取使D最大的层次划分

英语单词关联网络的社团结构

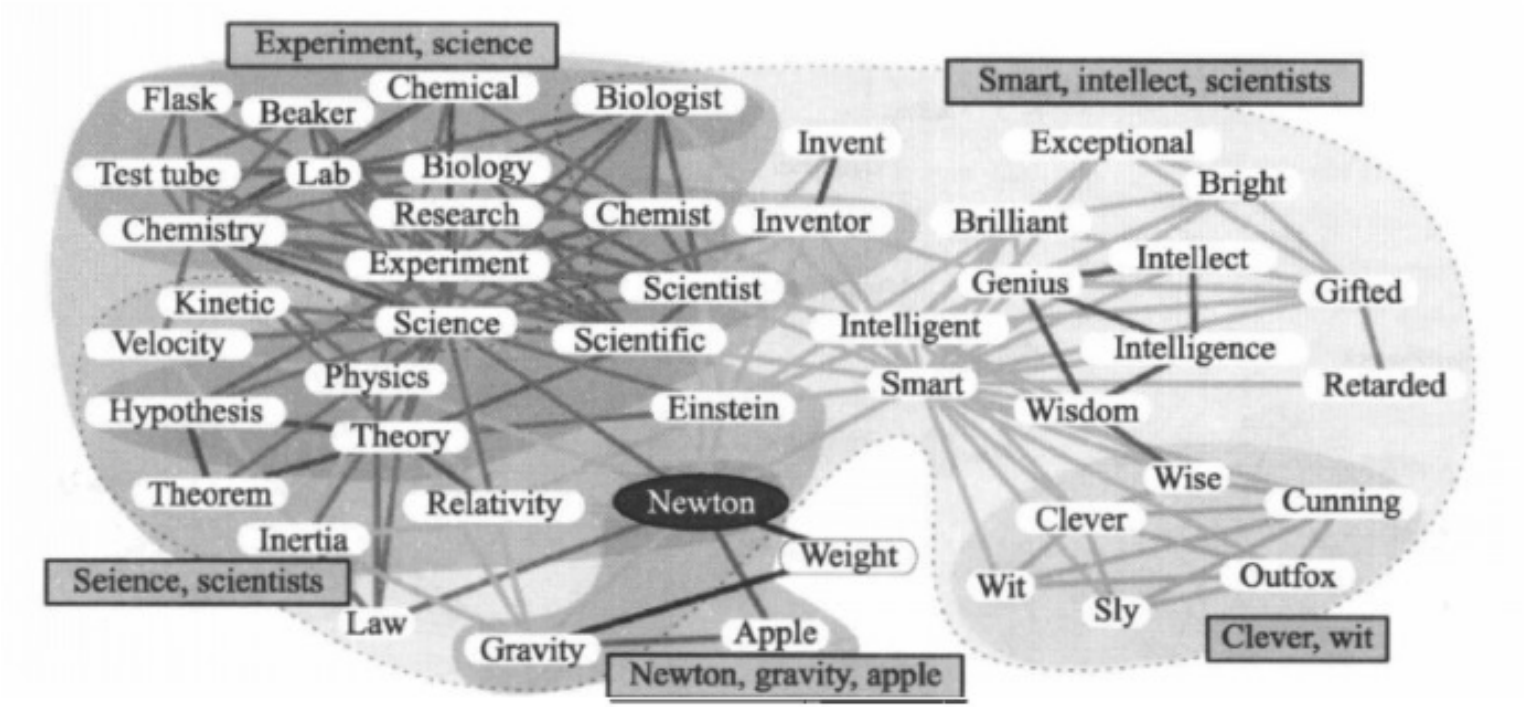


图 4-25　英语单词关联网络的社团结构(取自文献[23])

Thank you!