社会计算

第八节

社团挖掘

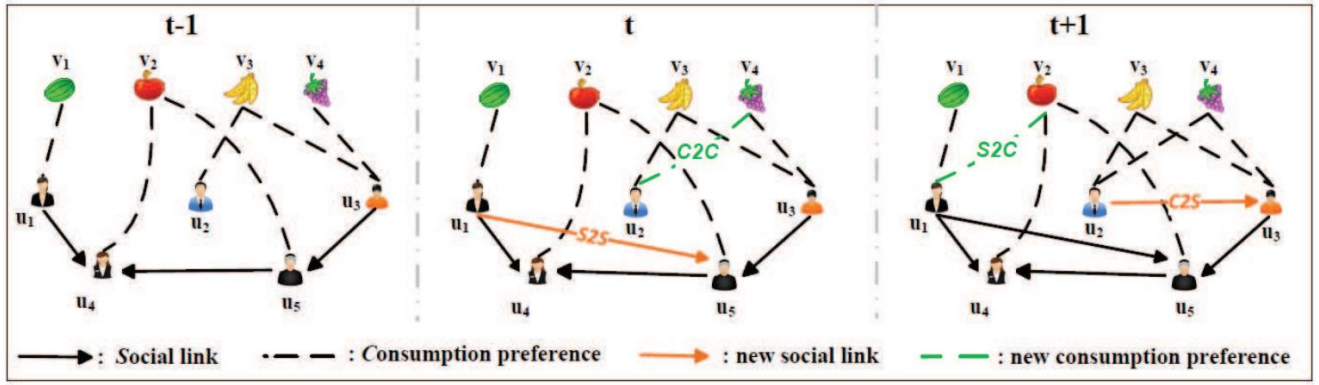
徐童

2024.4.22

上节回顾

- **社交推荐：社交与行为的联合演化**

- 先前所介绍的所有社交推荐问题，其本质都是建立在固定的社交网络上
  - 然而现实中，社交网络的结构在不断发生演化
  - 社交影响与社交选择互为因果，从而不断强化节点同质性
    - 这一现象是否能够解释社会网络中各种稠密子结构的来源?



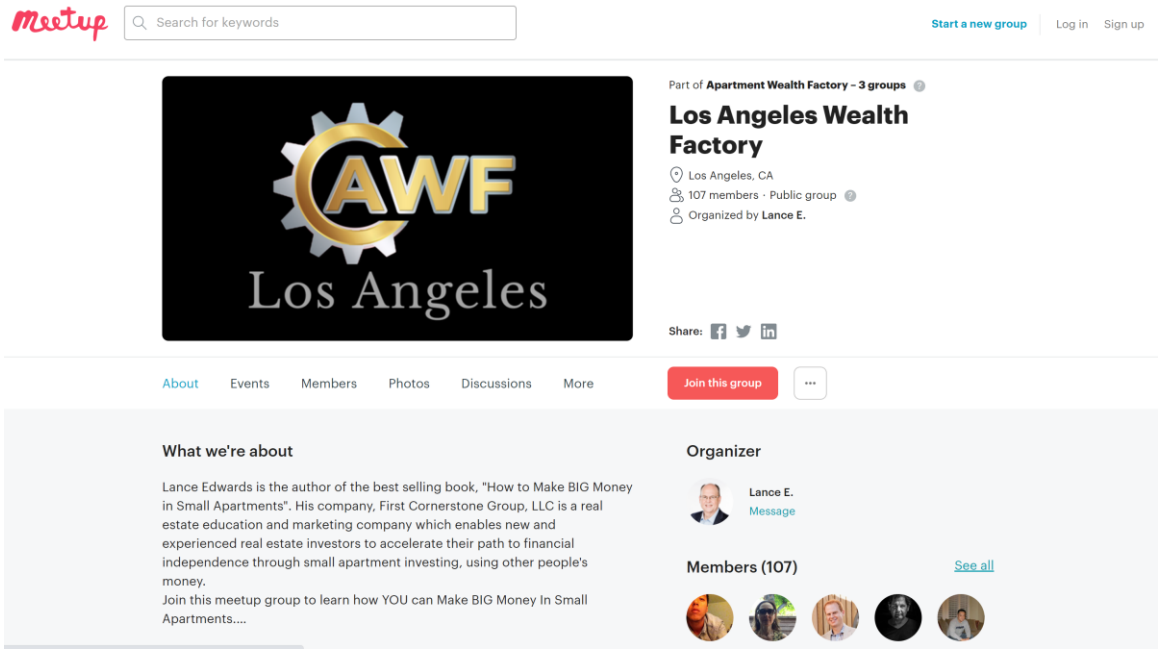
Explanations of A2B

A2B	Explanations
C2C	Collaborative Filtering
S2C	Social Influence
S2S	Node Proximity
C2S	Homophily Effect

本节主旨

- **社会网络中的那些子结构**

- 社会网络的发展，更导致了社交方式的改变
  - 传统的熟人社交演化为今天的陌生人社交，兴趣在社交中的作用日益重要
  - 志同道合的用户形成兴趣社团，*就共同关心的话题交换了意见*



本节主旨

- **围绕社团的那些问题**

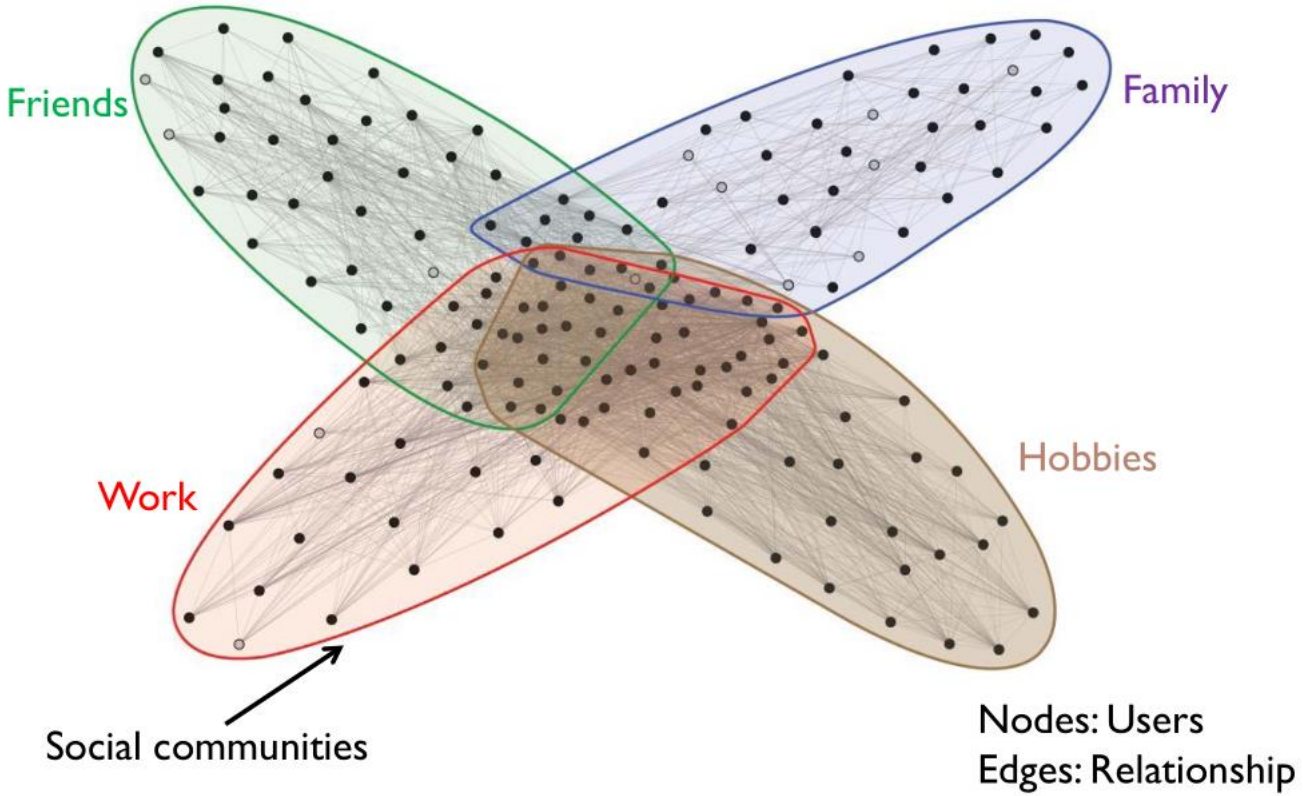
- 回到我们课程开始的那句话：有人的地方，就有江湖
- 然而，江湖如何成型，又如何影响我们的决策？
  - 问题一：如何挖掘网络中的社团？
  - 问题二：基于成型社团，人们如何实现团体决策？
  - 问题三：借助众人的智慧，我们能做些什么？

本节目录

- **社团挖掘**
  - 层次聚类技术
  - 划分聚类技术
- 社团决策
- 众包与群体智能概述

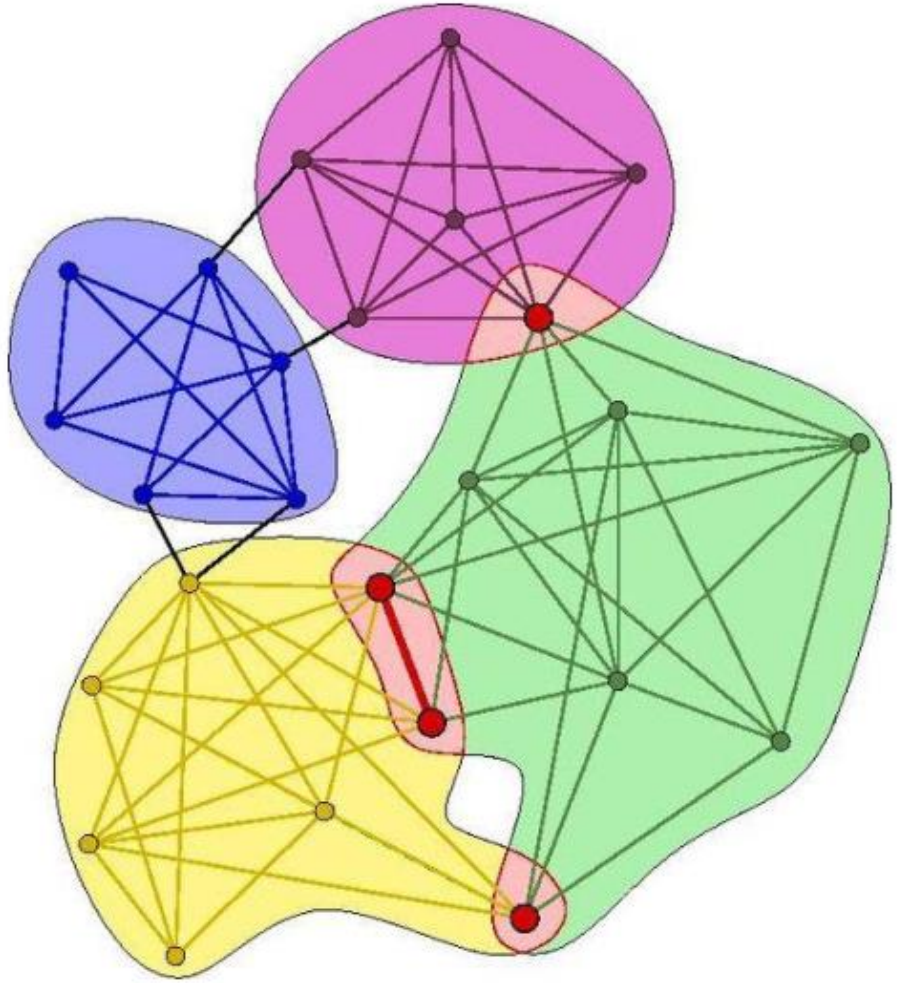
- **社团的广泛存在**

- 社团结构广泛存在于我们的日常生活中，我们在不同社团中扮演着不同角色



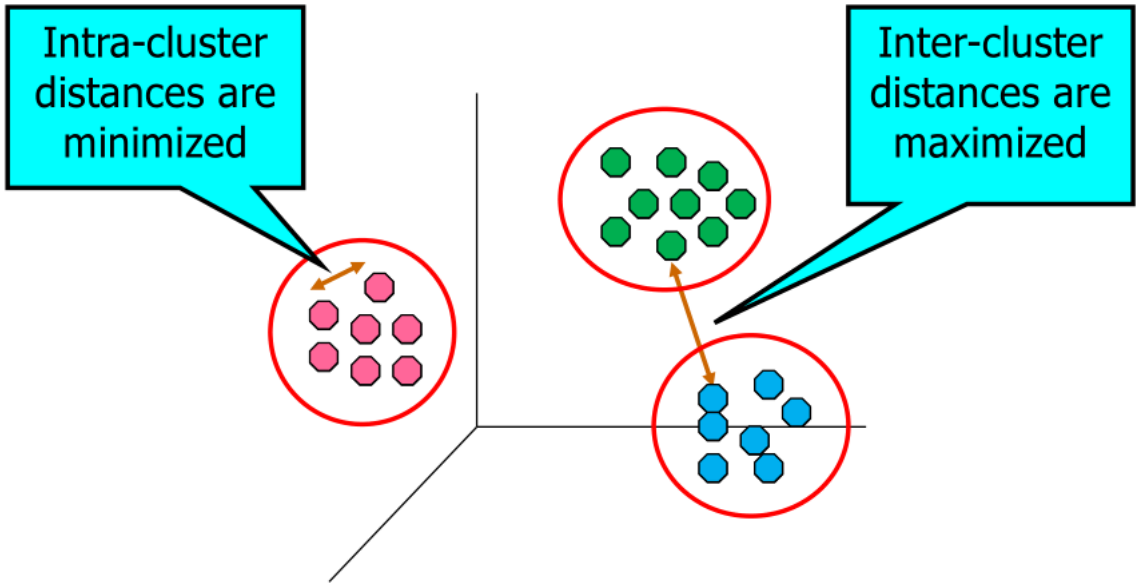
- **如何挖掘社团?**

- 我们凭借何种技术实现社团挖掘?
  - 为简便问题考虑，我们在这里仅考虑无向图（双向关系）上的社团挖掘问题
  - 一个基本的思路：社团的来由是“物以类聚，人以群分”
  - 如何描述这种自发的聚集性?



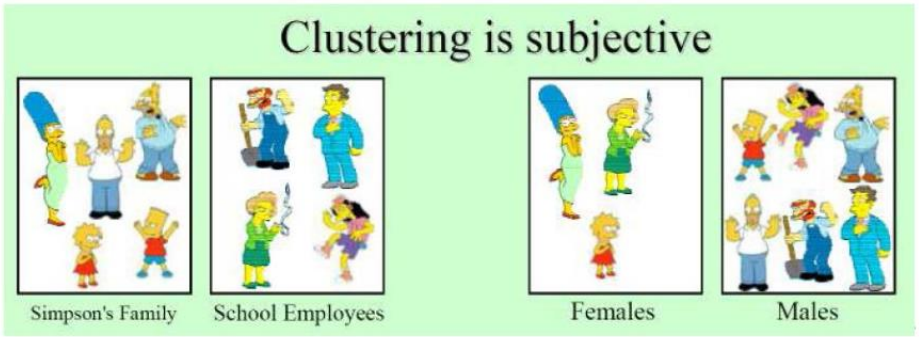
- **他山之石：聚类问题**

- 聚类问题的目的，在于将样本自发地聚集形成若干“簇”的结构
  - 簇的特点：簇内相似（距离较近），簇间相异（距离较远）
  - 对应着社团内高度的趋同性和一致性



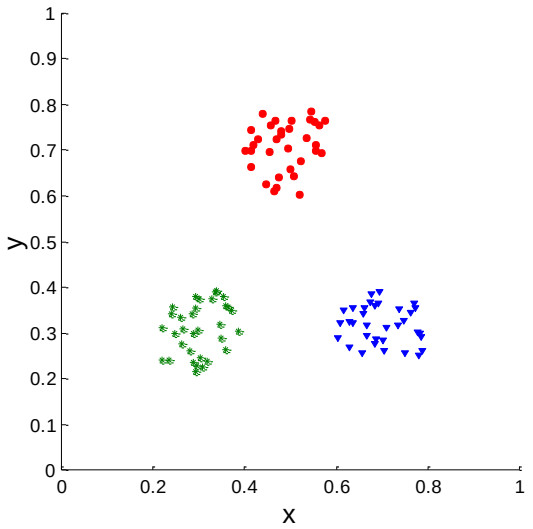
- **值得思考的问题：何以为簇?**

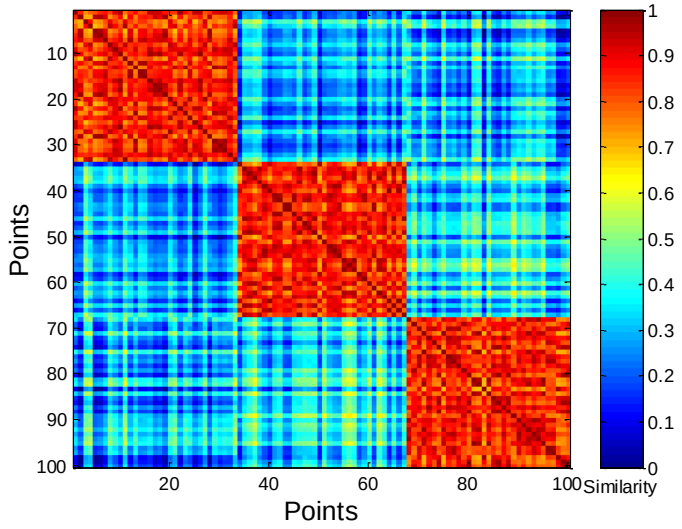
- 基于不同的“群体性”立场，可以得到不同的簇
  - 因此，聚类是具有一定主观性的，其主观性来自于聚类依据的选择
  - 在进行聚类时，需要根据问题目的选择合适的聚类依据
- 对于社团挖掘问题而言，网络关系是社团判定的首要准则
  - 先“社交”，再“社团”
  - 其它因素（如兴趣、关系强度等）可以通过网络权重加以体现



• 什么样的社团挖掘结果是好结果

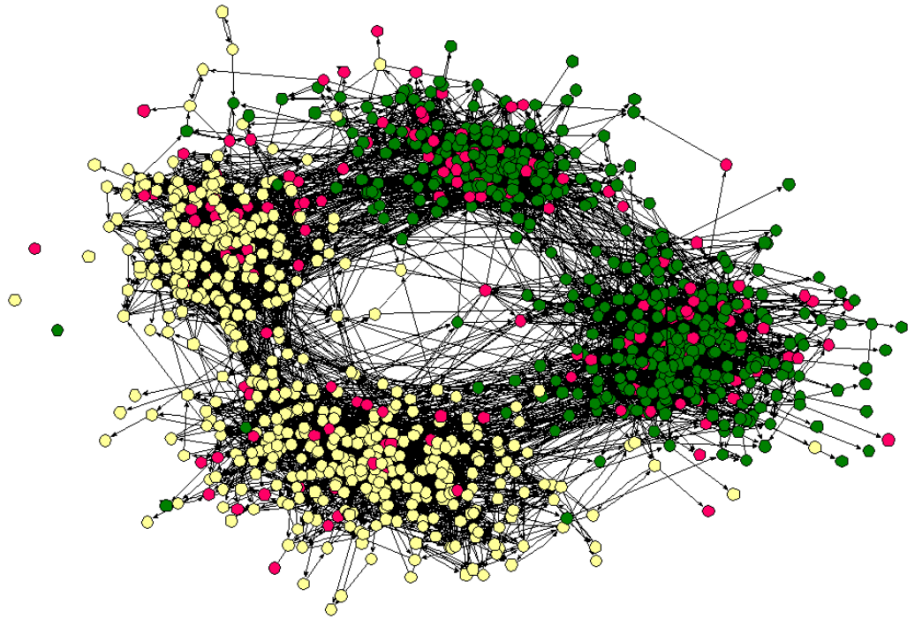
- 如前所述，理想的聚类应该是“簇内相似，簇间相异”
  - 在社交网络的视角下，就是社团内网络链接密度高，社团间弱连接稀疏
  - 通过行列变换，观察邻接矩阵是否体现出对角模式，可以大致判断社团的情况





- **什么样的社团挖掘结果是好结果**

  - 然而，现实中的网络规模宏大且结构复杂，难以采用观察或简单规则判断
  - 我们需要更为有效处理大规模数据的聚类方法，用于支撑社团挖掘！

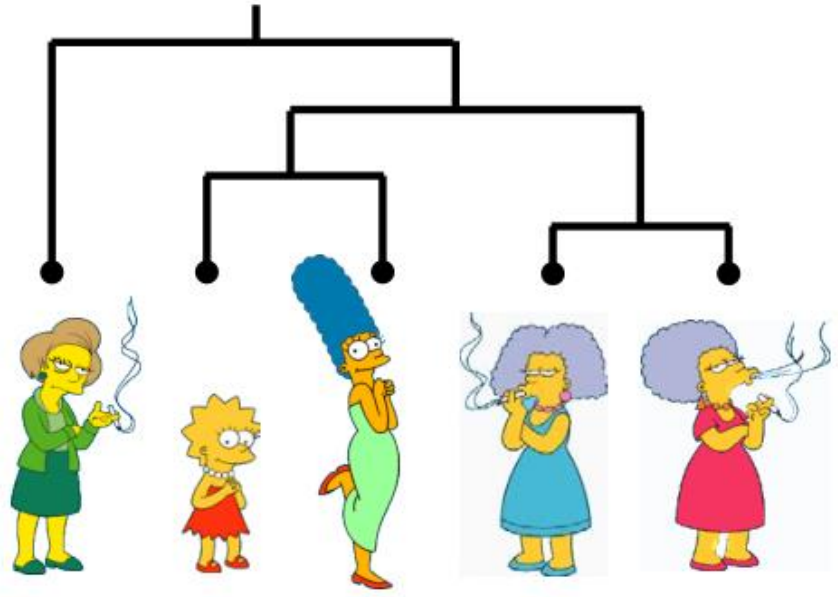


- **基于聚类的社团挖掘**

- 既然说到聚类，就不得不提到聚类的两类常见算法了

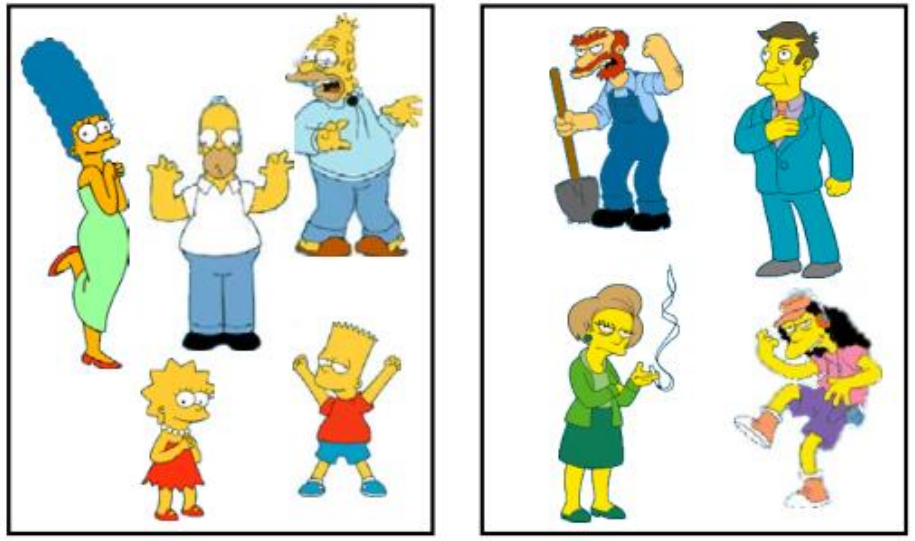
层次聚类

树状形式的嵌套簇集合



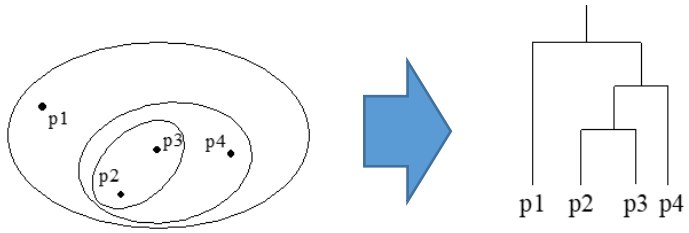
划分聚类

简单将样本划分为不重叠的簇



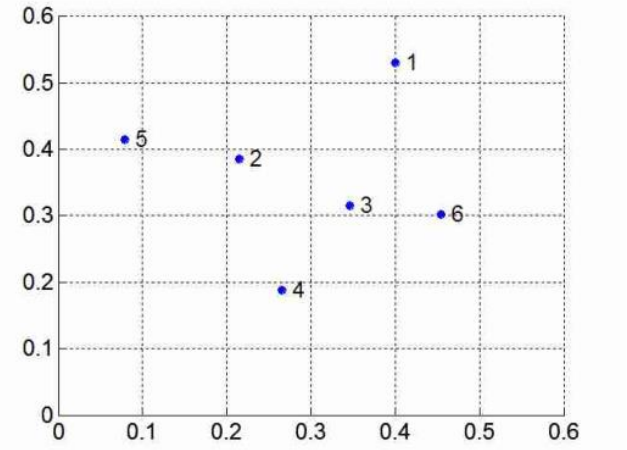
• 基于层次聚类的社团挖掘

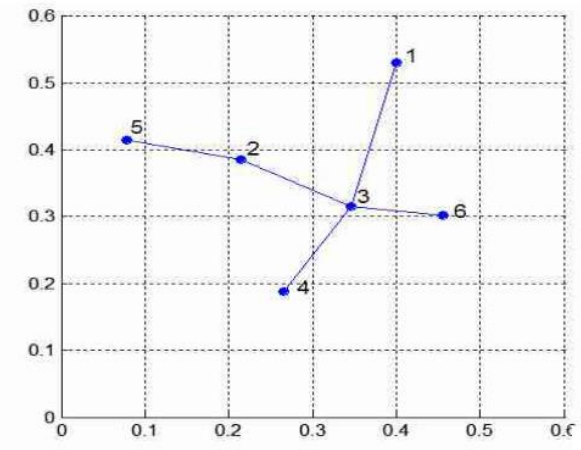
- 我们先来看基于层次聚类的社团挖掘思路
- 通常而言，层次聚类具有以下两种基本形式
  - 凝聚式聚类（Agglomerative，自下而上）
    - 将所有样本视作个体簇，逐步合并最接近的两个簇
  - 分裂式聚类（Divisive，自上而下）
    - 从包含所有样本的完整簇开始，每一步分裂一个簇
  - 一般而言，凝聚式聚类更为常见



- **基于层次聚类的社团挖掘**

- 两种层次聚类算法是否都可以利用与社团挖掘？我们首先来看分裂式聚类
  - 分裂式聚类的代表性算法：最小生成树聚类
  - 每次从图中断开一条边，就可以将原先的簇一分为二，形成两个新簇
    - 事实上，每次断开一条边，就是一个社团一分为二的过程





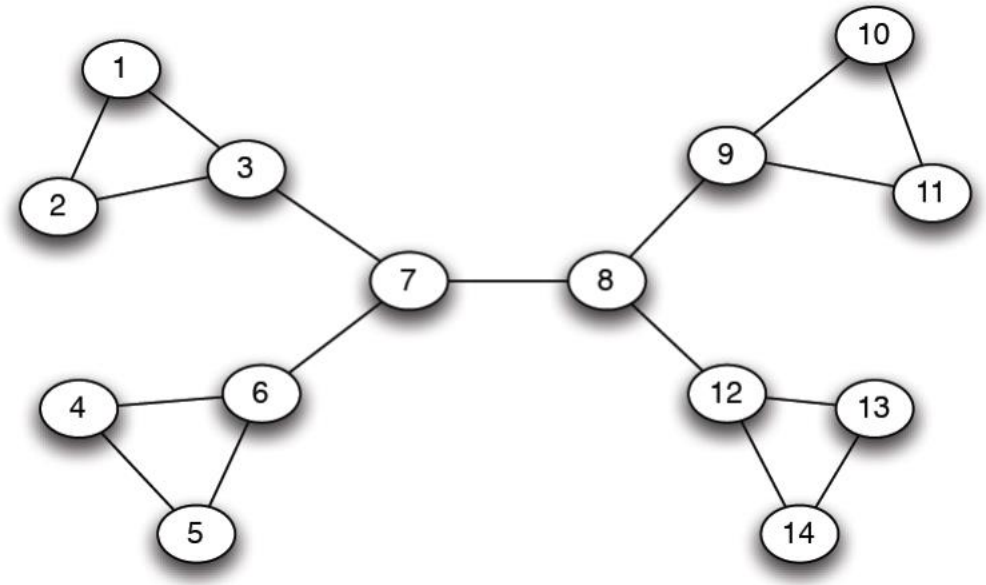
上古回顾

• 基于层次聚类的社团挖掘

- 社团一分为二的过程有没有觉得眼熟？邓巴数与150原则
  - 150原则：人类智力所允许拥有稳定社交网络的规模大约是150人。
  - 该原则由英国牛津大学的人类学家罗宾·邓巴（Robin Dunbar）在20世纪90年代提出，源于对于动物族群的观察
    - 这个过程，无疑是连通组件拆分为更多子连通组件的过程
    - 拆分之后我们可以发现，原先一个完整的社会网络，衍生出了两个相互连接不多的“社团”结构

- **基于层次聚类的社团挖掘**

- 事实上，分裂式聚类的社团挖掘，本质上就是逐步去掉弱连接的过程
  - 将那些沟通不同连通组件的桥接（Bridge）逐步删除
  - 现在的问题在于：如何找出这些桥接，按什么顺序删除?

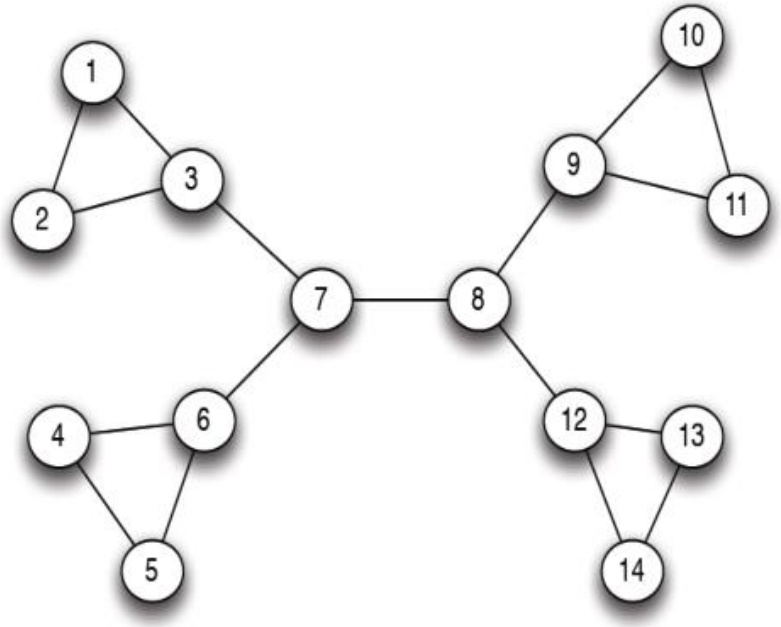


- **基于层次聚类的社团挖掘**

- 在之前介绍结构洞时我们曾提过，结构洞作为沟通各方的桥梁，往往在通讯中居于“垄断”地位。而垄断的原因，就是这种沟通必须经过结构洞
  - 换言之，有多少条路径通过某条边，可以反应这条边的中介作用强度
  - 一个基本指标：边介数（Edge Betweeness）
    - 定义为网络中任意两点间的最短路径，有多少条会通过这条边

基于层次聚类的社团挖掘

- 边介数的一个计算实例
  - 7-8这条边，任意由[1-7]和[8-14]的点对之间最短路径必经过这条边，介数为7×7=49
  - 3-7这条边，任意由[1-3]和[4-14]的点对之间最短路径必经过这条边，介数为3×11=33
    - 6-7，8-9，8-12与之类似



  - 1-3这条边，任意由[1]和[3-14]的点对之间最短路径必经过这条边，介数为1×12=12
    - 2-3，4-6，5-6，9-10，9-11，12-13，12-14与之类似
  - 1-2这条边，与其他最短路径无关，介数为1。4-5，10-11，13-14与之类似

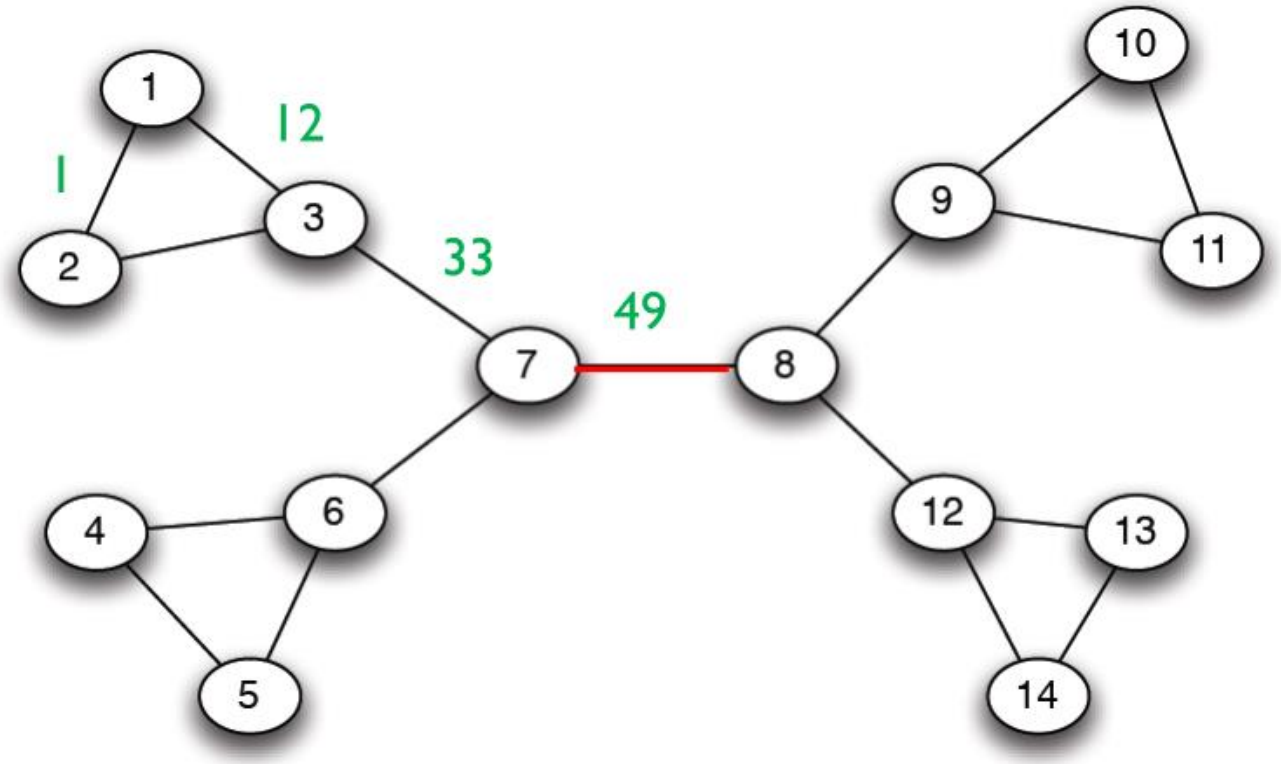
- **基于层次聚类的社团挖掘**

- 由此，2002年，Girvan与Newman提出了Girvan-Newman算法
  - 基于边的边介数，实现节点的层次聚类，具体流程如下：
    1. 计算网络中所有边的边介数；
    2. 去除边介数最高的边；
    3. 重新计算去除边后的网络中所有边的边介数；
    4. 跳至步骤2，重新计算，直至网络中没有边存在。

Girvan, M., and Newman, M. E. (2002). Communitystructure in social and biological networks. Proceedings of theNational Academy of Sciences, 99(12), 7821-7826.

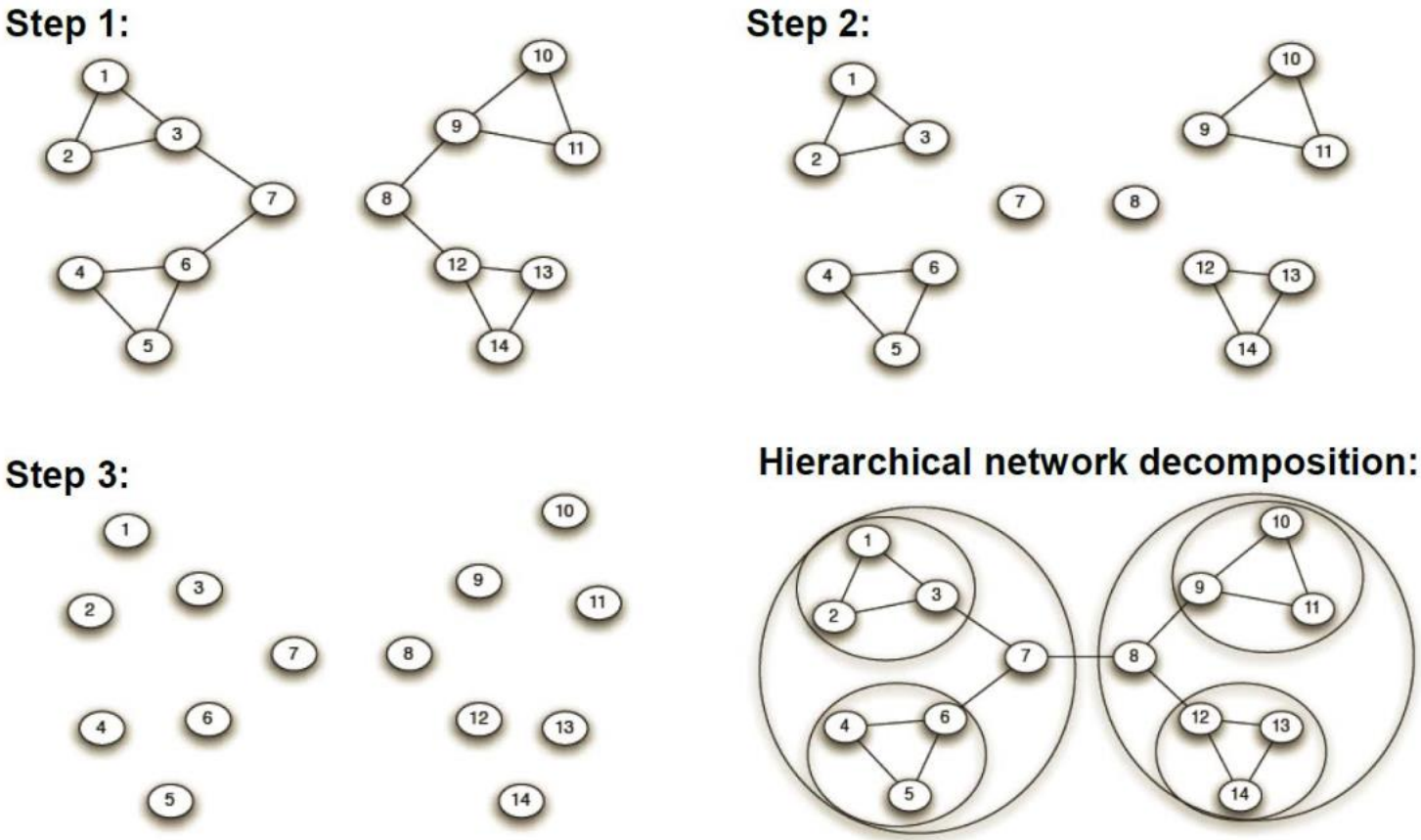
- **基于层次聚类的社团挖掘**

- Girvan-Newman算法的操作实例



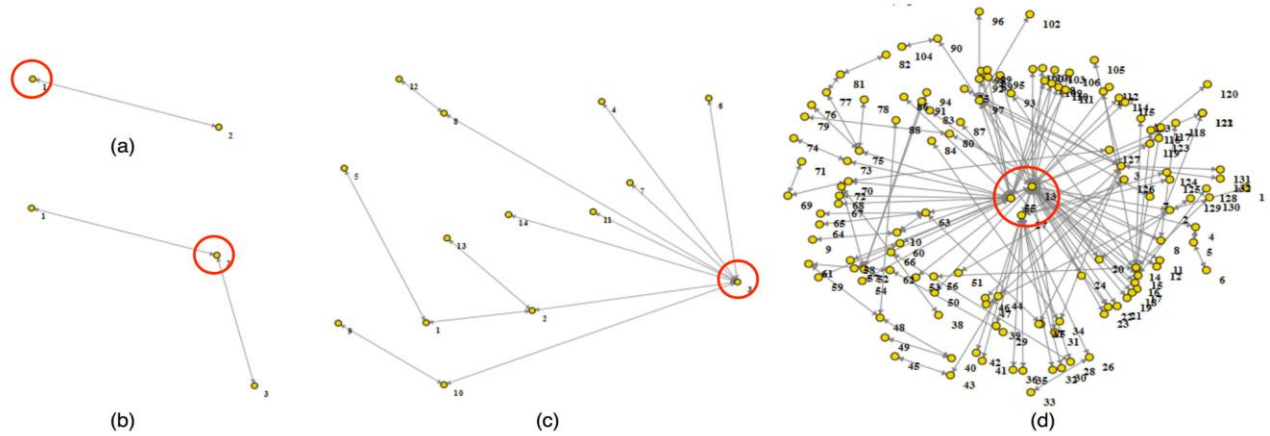
- **基于层次聚类的社团挖掘**

- Girvan-Newman算法的操作实例



- **基于层次聚类的社团挖掘**

- 另一类社团挖掘过程，是自下而上的“凝聚式”聚类
  - 从本质上说，它描述了个体节点如何逐渐聚集起来形成社团的过程
  - 回忆：先前提到过的“传销式”网络生成过程及其 “星形网络”特性



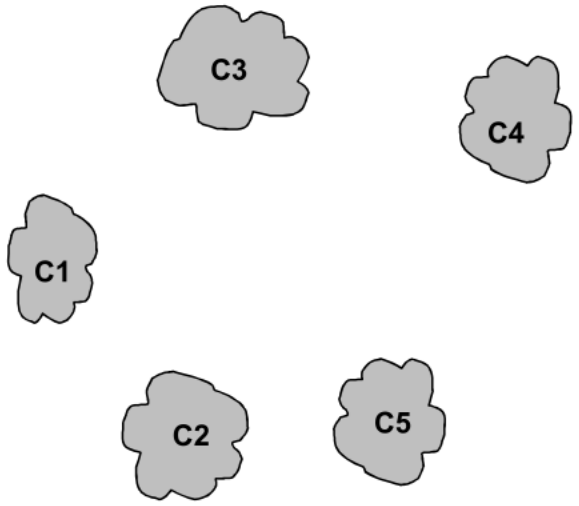
Chunyan Wang, et al., From Face-to-Face Gathering To Social Structure, CIKM-2012

基于层次聚类的社团挖掘

- 为实现凝聚式聚类，需要引入**邻近度矩阵**的概念
  - 用于存储两两簇之间的邻近度
  - 在社团挖掘任务中，“邻近度” 需要结合网络结构进行定义！
    - 例如，带权网络的权重作为邻近度，如果两个点不相连则邻近度最低（0或负无穷）
- 凝聚式聚类的基本流程非常直观，主要迭代以下两步，直到仅剩一个簇
  1. 合并邻近度最高的两个簇
  2. 基于更新的簇重新计算邻近度，更新邻近度矩阵
- 不同的凝聚式聚类方法，区别主要在于不同的**邻近度定义**

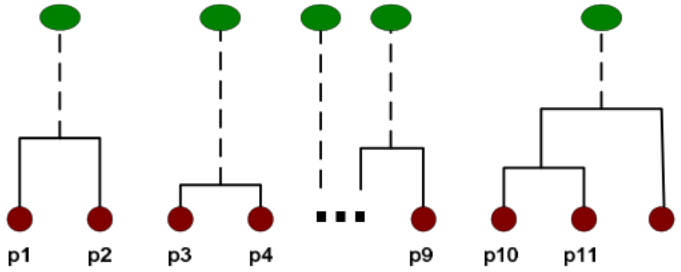
- **基于层次聚类的社团挖掘**

- 凝聚式聚类的实例
  - 如下图所示，我们已经获得了五个簇，并得到了相应的邻近度矩阵



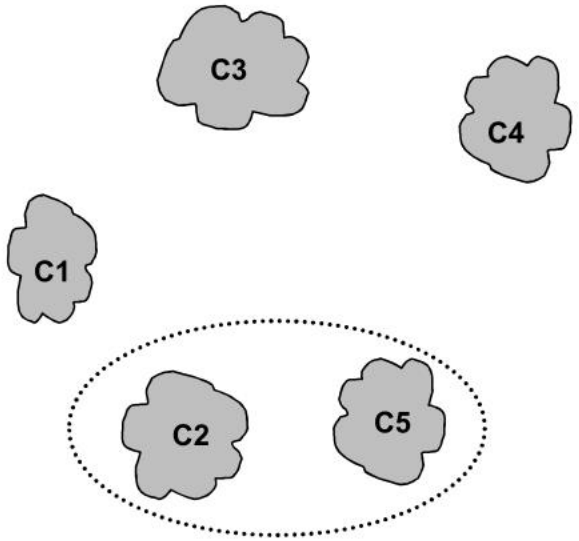
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

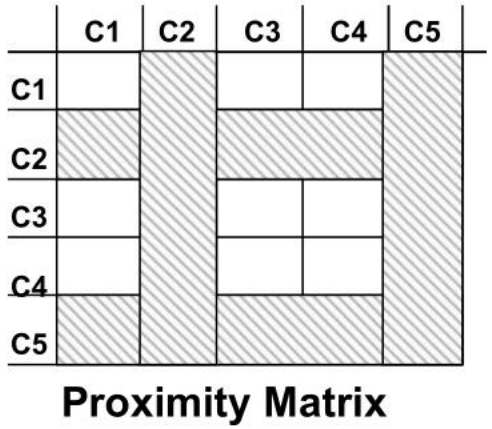
Proximity Matrix

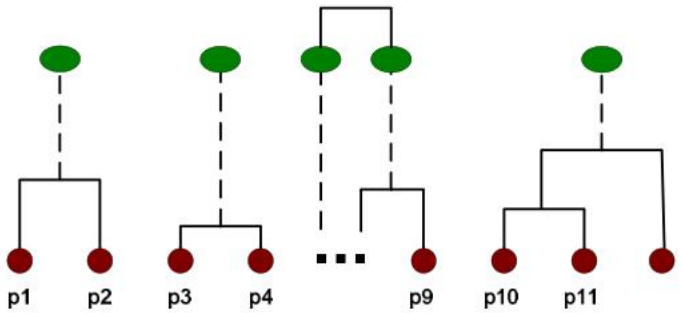


- **基于层次聚类的社团挖掘**

- 凝聚式聚类的实例
  - 通过邻近度矩阵，我们发现C2与C5之前的邻近度最高，可以进行合并

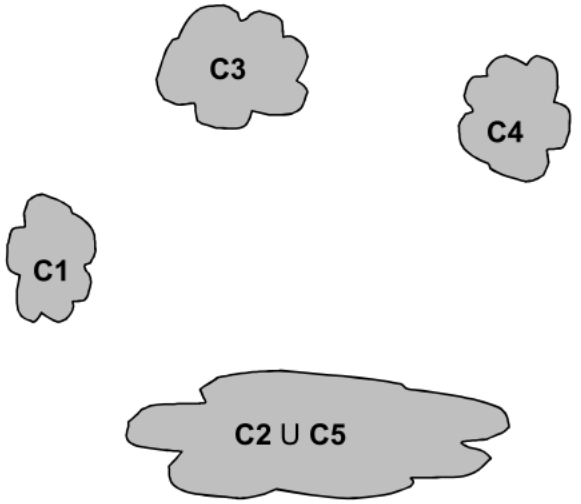






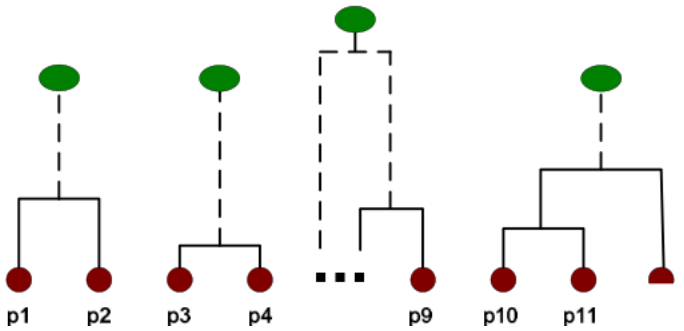
- **基于层次聚类的社团挖掘**

- 凝聚式聚类的实例
  - 基于合并结果，重新计算两两簇之间的邻近度并更新邻近度矩阵



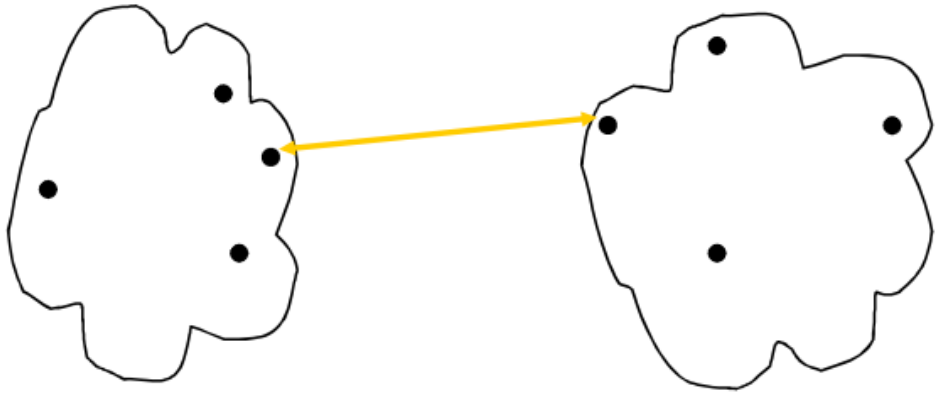
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



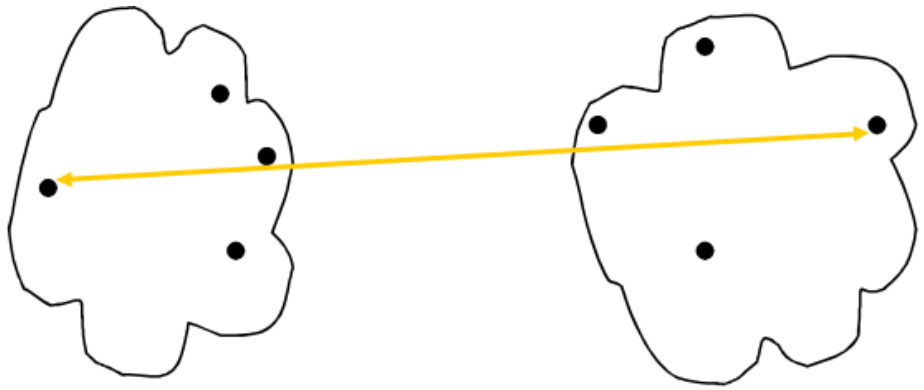
- **基于层次聚类的社团挖掘**

- 凝聚式聚类的核心问题在于**邻近度**的计算，不同聚类方法计算方式不同
- 常见的邻近度定义包括：
  - 单链（Single Link），也可表示为MIN，指不同簇最近的点之间的邻近度
    - 很好理解，两个强关系的伙伴可以成为联结两大社团的纽带



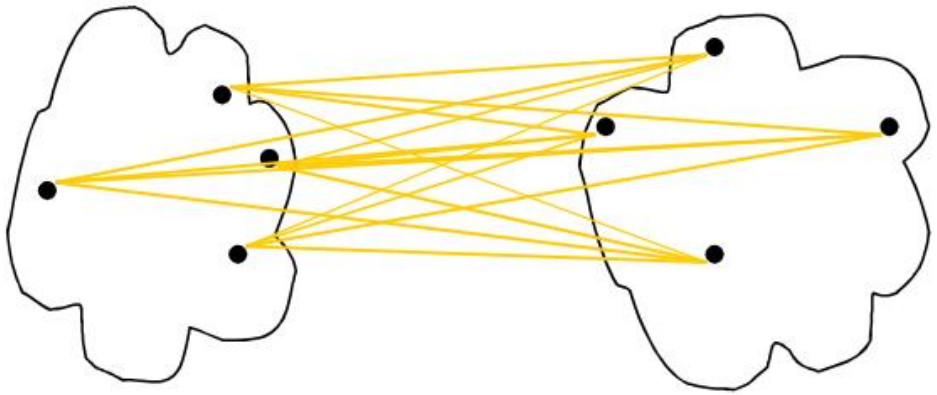
- **凝聚式聚类的邻近度定义**

- 凝聚式聚类的核心问题在于**邻近度**的计算，不同聚类方法计算方式不同
- 常见的邻近度定义包括：
  - 全链（Complete Link），也可表示为MAX，指不同簇最远的点之间的邻近度
    - 如何解决两个点因不存在社交关系导致邻近度为0的干扰？能否简单一删了之？



- **凝聚式聚类的邻近度定义**

- 凝聚式聚类的核心问题在于**邻近度**的计算，不同聚类方法计算方式不同
- 常见的邻近度定义包括：
  - 组平均（Group Average），指所有来自不同簇的两点之间的平均邻近度
    - 同样，不相邻节点是否需要统计在内？



- **基于层次聚类的社团挖掘**

- 层次聚类算法的优点之一：任意数量的社团结果都可以获得
- 然而，优点同样带来局限性：无法确定合适的社团数
  - 事实上，这是所有聚类方法的通病
    - 层次聚类是将所有可能全部展示出来供挑选，但挑选仍需标准
    - 一种思路是，采用层次聚类的评价指标进行选择
      - 如，社团的凝聚度与分离度
    - 另一种思路：社团结构有多鲜明？和随机打乱的网络比一下就知道

- **基于层次聚类的社团挖掘**

- 2004年，Newman为衡量G-N算法效果提出了Q-模块度（Modularity）的指标
  - 由于该指标的科学性，迅速成为一般性的社区划分算法的通用标准
  - Q-模块度的计算公式 $$Q = \sum_{ij} \left[\frac{A_{ij}}{2m} - \frac{k_i * k_j}{(2m)(2m)} \right] \delta(c_i, c_j)$$
  - 其中，$\delta(c_i, c_j)$ 仅当 i，j 两个节点属于同一社团时为1，否则为0
  - A为网络对应的邻接矩阵，A_{ij} = 1意味着 i，j 两个节点之间有一条边
  - k表示节点的度，m 为边的数量（为什么要乘以2？）
    - 提示：这里的图为无向图（A矩阵的对称性）

- **基于层次聚类的社团挖掘**

- 2004年，Newman为衡量G-N算法效果提出了Q-模块度（Modularity）的指标

  - Q-模块度的计算公式 $$Q = \sum_{ij} \left[\frac{A_{ij}}{2m} - \frac{k_i * k_j}{(2m)(2m)} \right] \delta(c_i, c_j)$$

  - 其中括号的部分，表达的就是节点之间的实际连边概率高于期待值的程度

    - 其中，$\frac{k_i}{2m} * \frac{k_j}{2m}$ 表示如果节点度数不变，而把所有的边全部打乱重来，恰好在 i 和 j 节点之间仍然存在一条边的概率

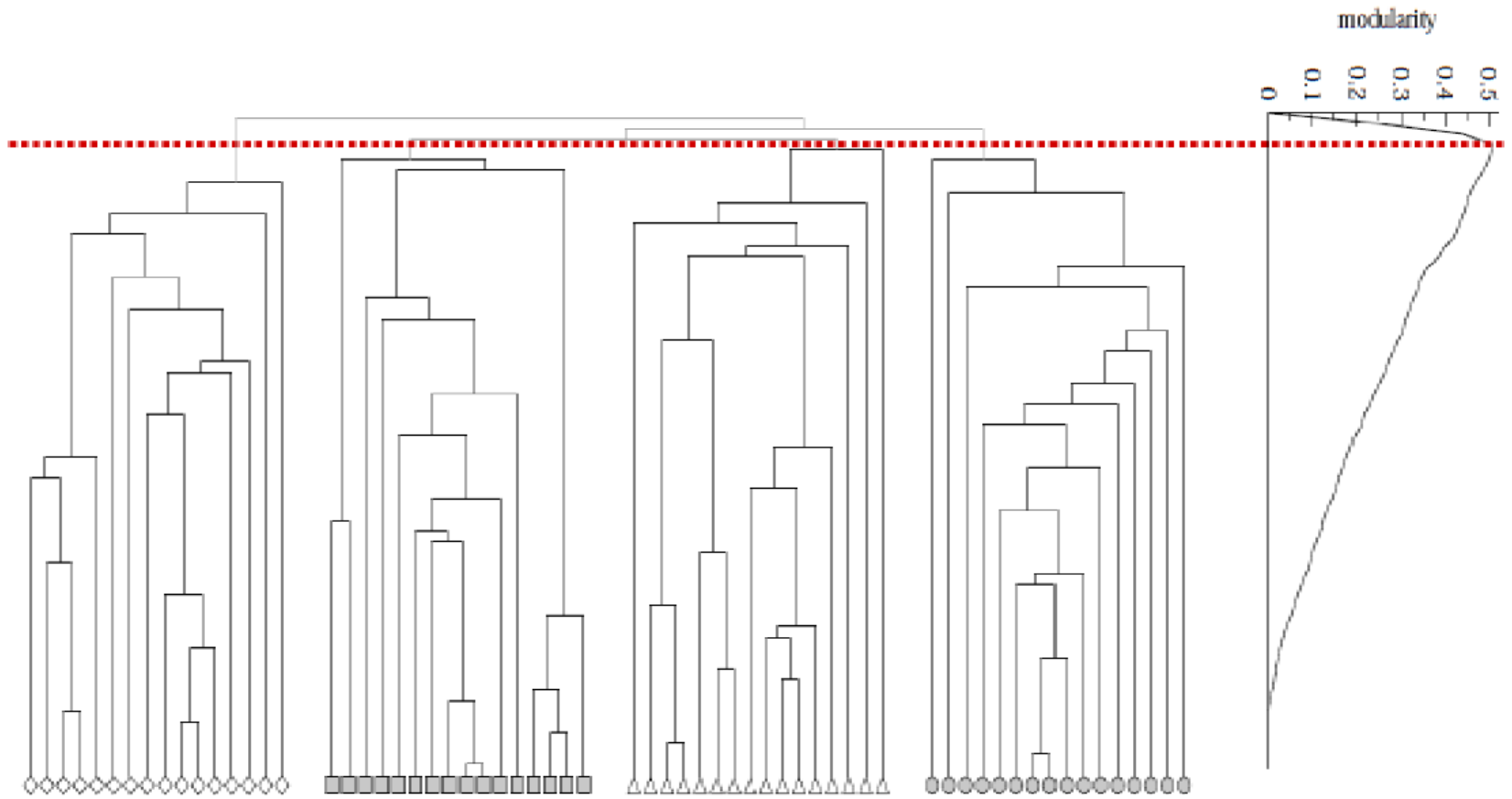
    - 而相应的，$\frac{A_{ij}}{2m}$ 表示当前情况下两节点之间存在边的实际概率

    - 主要就是为了判断相比于随机情况，现有社团的内聚性有多好

  可参考：http://www.yalewoo.com/modularity_community_detection.html

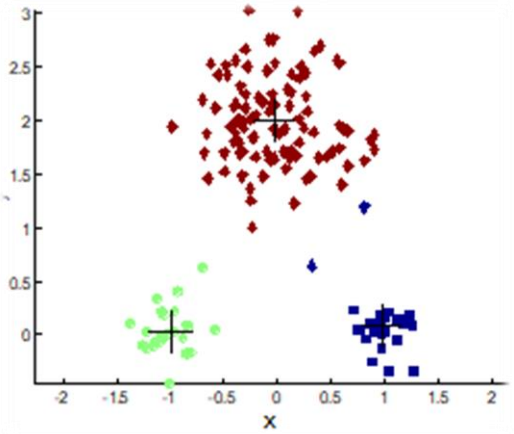
- **基于层次聚类的社团挖掘**

- 基于Q-模块度指标，可以实现对于社团个数的挑选



- **基于划分聚类的社团挖掘**

- 另一方面，划分聚类也可有效用于社团挖掘
  - 例如，采用划分聚类的代表性算法：K均值聚类
- K均值（K-means）聚类的思想，就是通过设定K个中心，来形成K个簇，
  - 然后，通过不断更新簇中心的向量，来更新聚类的结果，直至收敛
  - 簇中心的更新，依赖于对当前簇中样本的算术平均
  - 簇中心更新后，根据距离将样本重新分配至不同的簇
  - **收敛**：所有样本的聚类结果不再更新，停止迭代

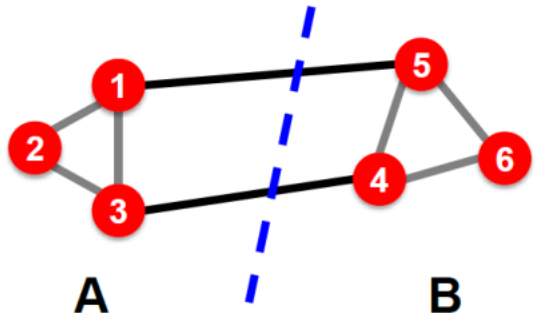


• 基于划分聚类的社团挖掘

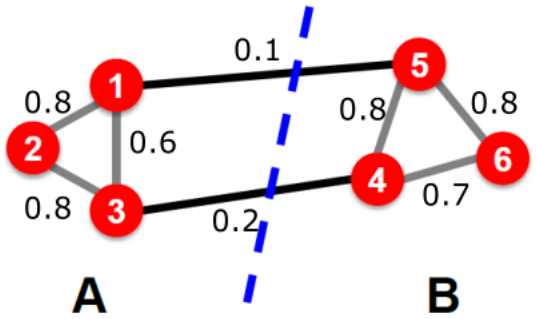
- 另一方面，划分聚类也可有效用于社团挖掘
  - 例如，采用划分聚类的代表算法：K均值聚类
    - K均值聚类依赖样本属性的相似度进行聚类，节点的属性是什么？
    - 一种办法是引入额外信息，如节点的属性，但并不普遍
    - 另一种方法与前面类似：直接采用邻接矩阵作为节点的属性
      - 例如，n个节点，每个节点一个n维0-1向量，1表示邻居
      - 基于这种思路，可以采用各种聚类算法加以实现

- **基于划分聚类的社团挖掘**

- 另一方面，划分聚类也可以用于社团挖掘
  - 另一种代表性算法：谱聚类（Spectral Clustering）
    - 基于最小割算法，面向图结构的聚类技术



$cut(A,B) = 2$

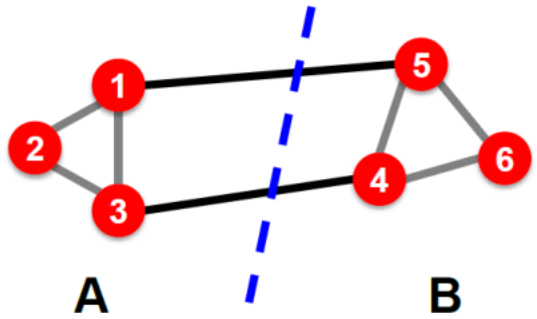


$cut(A,B) = 0.3$

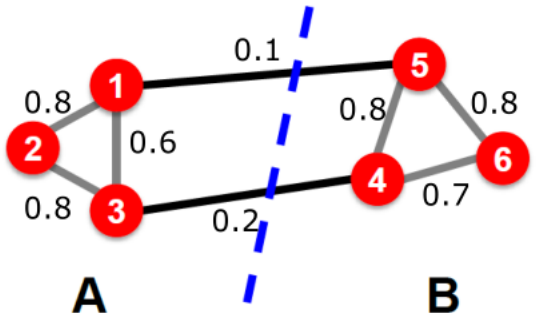
- **基于划分聚类的社团挖掘**

- 谱聚类的目的，在于找到最小的“割”
  - “割”：边的集合，每一条边的两个端点分别属于不同的节点集合

$$cut(A,B) = \sum_{i \in A, j \in B} w_{ij}$$



$cut(A,B) = 2$

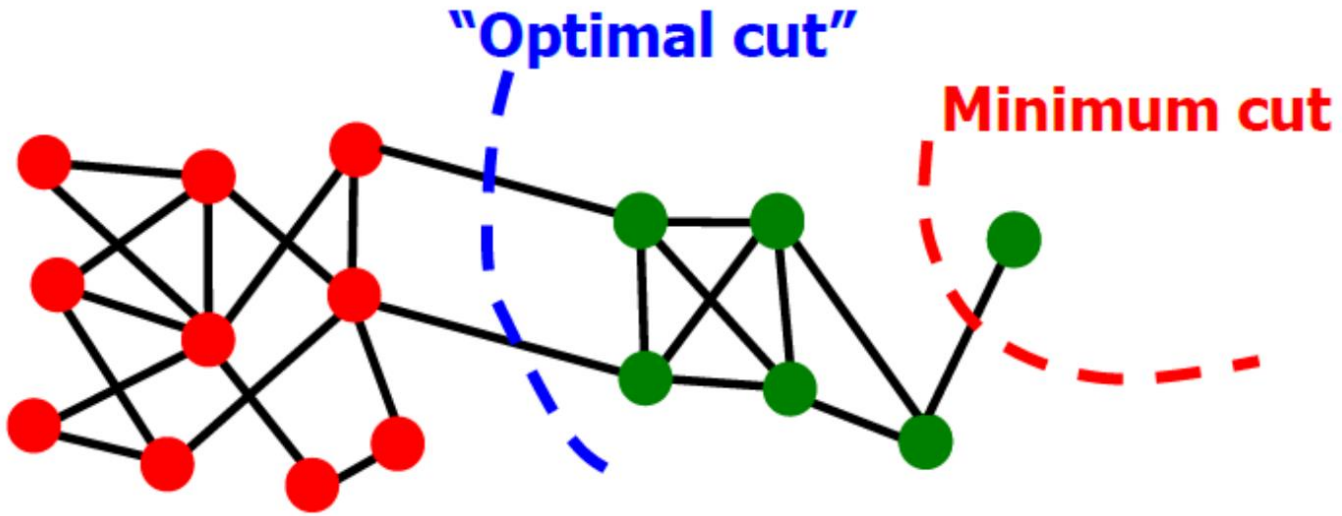


$cut(A,B) = 0.3$

如果不记得最小割算法了，可参考：https://blog.csdn.net/jteng/article/details/49590069

- **基于划分聚类的社团挖掘**

- 谱聚类的目的，在于找到最小的“割”
  - 一些值得注意的问题：基本的目标函数忽视了社团规模带来的影响
    - 倾向于切分出来一些规模分布严重不均的社团，而非最优切分方式



- **基于划分聚类的社团挖掘**

- 谱聚类的目的，在于找到最小的“割”
  - 通过对目标函数（Cut）加以约束，使之所得社团趋于均衡

Balanced min-cut:

$$\min cut(A,B) \quad \text{subject to} \quad |A| = |B|$$

Ratio cut:

$$RatioCut(A,B) = cut(A,B)(\frac{1}{|A|} + \frac{1}{|B|})$$

约束点集的规模

Normalized cut:

$$Ncut(A,B) = cut(A,B)(\frac{1}{vol(A)} + \frac{1}{vol(B)})$$

约束点集和边连接的情况

注：Vol(A)指A中所有节点的度之和，避免生成过于稀疏的社团

本节目录

- 社团挖掘
  - 层次聚类技术
  - 划分聚类技术
- **社团决策**
- 众包与群体智能概述

从社团到社团决策

- 社团形成后，社团成员就面临着共同进退的问题
  - 此时，个性化的决策和偏好需要让步于以社团为单位的统一决策
  - 统一决策的结果，实际上体现着群体对于事物的判断
- 此时，问题来了：如何决策？两个基本问题
  - 如何进行决策?
  - 如何确定结果?

- **社团决策的基本机制：集体表决**

- 给定备选项$A = \{A1, A2, A3, \ldots, An\}$，通过社团成员投票方式加以判断

任务一：如何进行决策？

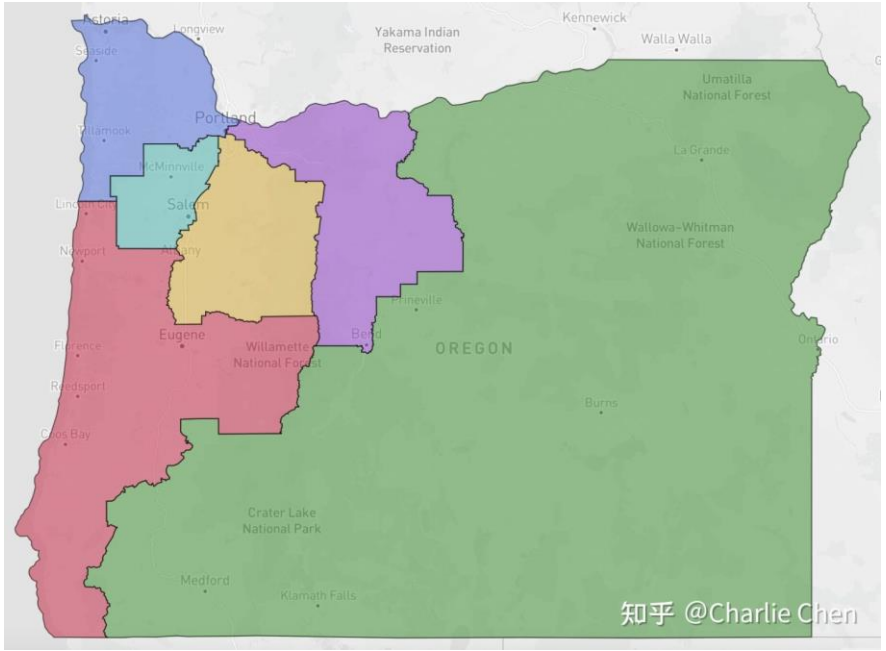
- 同意/不同意（O/X）
- 对A中的元素进行排序
- 对A的每个元素打分
- ……

任务二：如何确定结果？

- 少数服从多数
- 按一定比例（1/2或2/3等）阈值进行通过
- 去掉最高分/最低分
- 集体排序
- ……

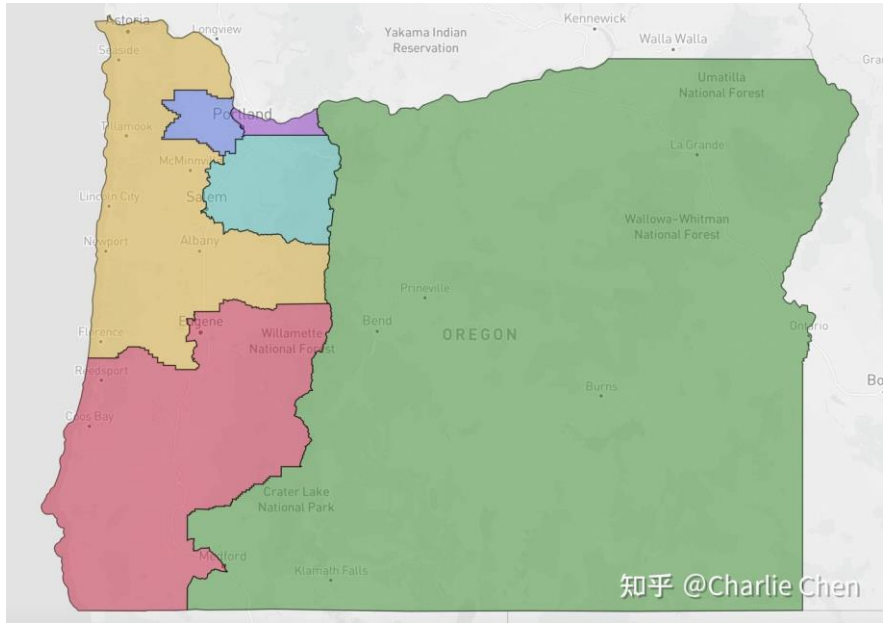
- **投票机制重要吗？很重要**

- 通过修改机制，同样的选票可能获得截然不同的结果



←民主党

共和党→



美国两党关于俄勒冈州选区划分的争议

- **投票机制重要吗？很重要**

- 通过修改机制，同样的选票可能获得截然不同的结果
- 历史教训：法共的议会斗争史
  - 1946年法国议会选举，法共为第一大党，得票率28.26%，在议会中斩获182个席位，议会总席位为627个，占比29.03%
  - 戴高乐的应对方式：修改选举法，将“比例代表制”改为“单记名多数两轮投票制”。同时配合选区划分，以往的十几个大选区被细分为577个小选区
  - 1958年，法共得票率降为18.9%，而席位更降为仅占1.83%

年度	得票数	得票率	获得席位数	获得席位占比	总席位数
1924	885,993	9.82%	26	4.48%	581
1928	1,066,099	11.26%	11	1.82%	604
1932	796,630	8.32%	10	1.65%	607
1936	1,502,404	15.26%	72	11.80%	610
1945	5,024,174	26.23%	159	27.13%	586
1946(6月)	5,145,325	25.98%	153	26.11%	586
1946(11月)	5,430,593	28.26%	182	29.03%	627
1951	4,939,380	26.27%	103	16.48%	625
1956	5,514,403	23.56%	150	25.21%	595
1958	3,882,204	18.90%	10	1.83%	546
1962	4,003,553	20.84%	41	8.82%	465
1967	5,039,032	22.51%	73	14.99%	487
1968	4,434,832	20.02%	34	6.98%	487
1973	5,085,108	21.39%	73	14.96%	488
1978	5,870,402	20.55%	86	17.62%	488
1981	4,065,540	16.17%	44	8.96%	491
1986	2,739,225	9.78%	35	6.65%	573
1988	2,765,761	11.32%	27	4.70%	575
1993	2,331,339	9.30%	24	4.16%	577
1997	2,523,405	9.92%	35	6.07%	577
2002	1,216,178	4.82%	21	3.64%	577
2007	1,115,663	4.29%	15	2.60%	577
2012	1,792,923	6.91%	10	1.73%	577

社团决策的基石：个体决策

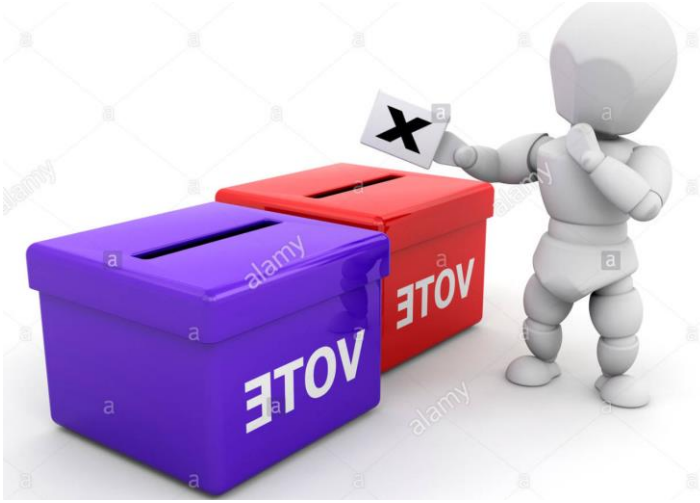
- 为实现社团的群体决策，个体首先要基于其偏好进行投票/打分/排序等
- 个体的偏好表达应该具有何种性质？
  - 首先，个体偏好应具有完备性
  - 即：对于任意一对备选项X和Y，应有明确的偏好
    - $X > Y$，或$X < Y$
  - 同时偏好两者，或两者都不喜欢，是不被考虑的
    - 选择困难症患者的噩梦

• 社团决策的基石：个体决策

- 为实现社团的群体决策，个体首先要基于其偏好进行投票/打分/排序等
- 个体的偏好表达应该具有何种性质？
  - 其次，个体偏好应具有传递性
  - 即：如果有X > Y，Y > Z，则必须有X > Z
    - 如果出现违反传递性的情况，则个体无法做出决策
    - 因此，必须首先保证个体有清晰明确的表态
  - 如果完备性和传递性均有保证，则存在“全序”
    - 即，所有备选项可以用一个顺序列表来表示

- **个体决策到群体决策**

- 假定每位表决者已经形成了自己偏好的“完序”列表，如何形成群体决策
  - 注意：形成的决策应具有“合理性”，即体现群体的意见
  - 何为群体的意见？**少数服从多数**
    - 如果多数人认为 X > Y，则群体意见中应优先选择 X



- **个体决策到群体决策**

- 假定每位表决者已经形成了自己偏好的“完序”列表，如何形成群体决策
  - 如何保证最后的群体决策体现“群体性”？
    - 首先，投票者一般为奇数——避免无法判断“多数派”的情况
    - 其次，如果只有两个候选项，则选择更具优势的一方

$$X \succ_1 Y$$
$$X \succ_2 Y$$
$$Y \succ_3 X$$

$$X \succ Y$$

- **个体决策到群体决策**

- 假定每位表决者已经形成了自己偏好的“完序”列表，如何形成群体决策
  - 然而，实际情况中往往不止两个候选项
    - 例如：晚餐去本科生食堂、研究生食堂，还是东门外?
    - 此时，我们可能会面临一个麻烦的情况……

成员1：X > Y > Z
成员2：X > Z > Y
成员3：Y > X > Z

X > Z 得到 3 票，全票通过
X > Y 得到 2 票，多数票
Y > Z 得到 2 票，多数票

X > Y > Z，最终结果

- **个体决策到群体决策**

- 假定每位表决者已经形成了自己偏好的“完序”列表，如何形成群体决策
  - 然而，实际情况中往往不止两个候选项
    - 例如：晚餐去本科生食堂、研究生食堂，还是东门外?
    - 此时，我们可能会面临一个麻烦的情况……

成员1：X > Y > Z
成员2：Z > X > Y
成员3：Y > Z > X

Z > X 得到 2 票，多数票
X > Y 得到 2 票，多数票
Y > Z 得到 2 票，多数票

？传递性被破坏了!

- **个体决策到群体决策**

- 上述例子表明一个道理：即使每个个体的偏好关系都具有“完序”（即完备且传递），但其整体结果仍然可能不具有传递性！
- 孔多塞（Condorcet）悖论：
  - 即使每个个体都满足完备性和传递性，其整体情况仍然可能自相矛盾
  - 合理的个体行为 + 合理的聚合方式 → 不合理的群体结论！

- **个体决策到群体决策**

- 上述例子表明一个道理：即使每个个体的偏好关系都具有“完序”（即完备且传递），但其整体结果仍然可能不具有传递性！
- 事实上，孔多塞悖论在日常生活中也经常可见
  - 曾虑多情损梵行，入山又恐别倾城，世间安得双全法，不负如来不负卿

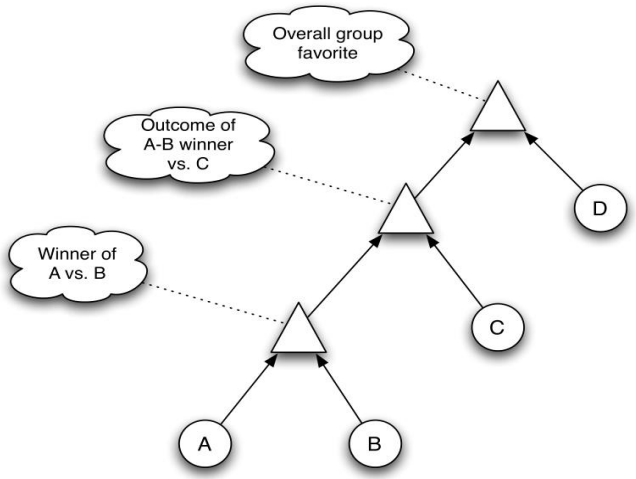
大学	全国排名	班平均规模	奖学金
X	4	40	$3000
Y	8	18	$1000
Z	12	24	$8000

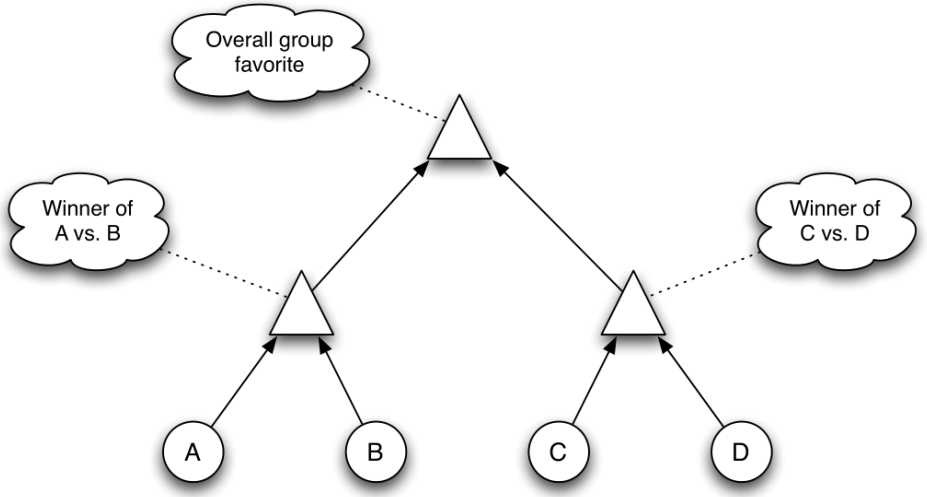
- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 一种启发式思路：给投票者 / 评判标准添加权重
    - 所有人一律平等，但有些人比其他人更平等
    - 可以根据权威性、性格、历史投票等多方面因素确定权重
      - 有关如何添加权重的问题稍后讨论

- **个体决策到群体决策**

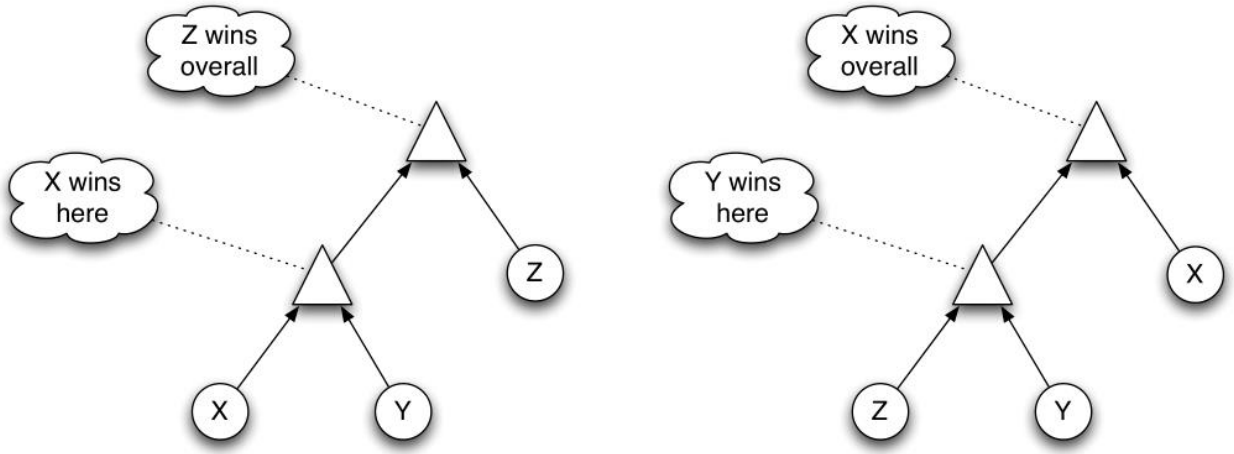
- 如何解决孔多塞悖论?
  - 另一种启发式方法：对于备选项施加一个排序
    - 按照排序依次进行表决，类似淘汰赛的形式
    - 当前轮的两个备选项比较后，胜者在与下一个备选项进行比较，直至获得最终选择





- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 另一种启发式方法：对于备选项施加一个排序
  - 需要注意的是：此时排序至关重要！（策略议程设置问题）



(a) *An agenda in which Z wins.*

(b) *An agenda in which X wins.*

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 第三种启发式方法：将排序转化为积分，统计总分达成决策
  - 波达计数法（Borda Count，1770）
    - 将N个候选项转化为对应的得分，第一名赋值为 N-1，以此类推，最后一名得0分
    - 根据每个成员的积分进行累加，最后得到的赋值和进行排序，形成群体决策
      - 注：此时需要对于平分情况进行额外约定

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 波达计数法实例

	第一偏好	第二偏好	第三偏好	第四偏好
个体1	A	B	C	D
个体2	B	C	A	D
得分	A=3+1	B=2+3	C=1+2	D=0+0
群体偏好	B	A	C	D

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 积分制看起来非常公平合理，然而有没有做手脚的空间呢?

影评家	公民凯恩	教父
1	1	0
2	1	0
3	1	0
4	0	1
5	0	1
得分	3	2

假定有五个成员

对两部电影进行打分

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 积分制看起来非常公平合理，然而有没有做手脚的空间呢?

影评家	公民凯恩	教父	低俗小说
1	2	1	0
2	2	1	0
3	2	1	0
4	0	2	1
5	0	2	1
得分	6	7	2

如果引进了第三个选项
会有何种变化?

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 积分制看起来非常公平合理，然而有没有做手脚的空间呢?

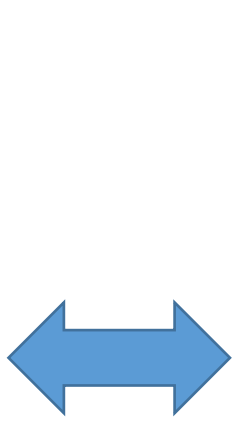
影评家	公民凯恩	教父	低俗小说
1	2	1	0
2	2	1	0
3	2	1	0
4	0	2	1
5	0	2	1
得分	6	7	2

然而此时，如果成员4、5希望《教父》胜出，
可以通过策略性投票，
使得最后的结果出现变化

- **个体决策到群体决策**

- 如何解决孔多塞悖论?
  - 上面的例子告诉我们：策略性偏好误报对于积分制会形成冲击
    - 虽然第三方实际上没有获胜的可能，但会干扰其他人对于排序的判断

投票率	55.7%[1][2] ▲ 0.8 %	
		
获提名人	**唐纳德·特朗普**	希拉里·克林顿
政党	共和党	民主党
家乡州	纽约州	纽约州
竞选搭档	**迈克·彭斯**	蒂姆·凯恩
选举人票	**304**[3][4][5]	227[6]
胜出州/省	**30 + 缅-2**	20 + DC
民选得票	62,984,828[7]	**65,853,514**[7]
得票率	46.1%	**48.2%**



伯尼·桑德斯

佛蒙特州联邦参议员 (2007 -)
竞选
失去提名：2016年7月26日 **13,167,848张初选票及1,846张代表票**
[59]

- **衡量群体决策的更多策略**

- 回到我们开头的问题：如何进行决策
  - 前面的例子都是基于排序结果，假如是二分投票（O/X）或满意度打分呢?
  - 一般而言，最基本的策略都是少数服从多数
    - 即根据赞同票的个数，赞同数最多者被选中作为集体决策
    - 然而，如果不是二分投票，而是进行打分，情况会更加复杂
      - 如何确定每个备选项的得分?

- **衡量群体决策的更多策略**

- 回到我们开头的问题：如何进行决策
  - 备选项得分判定的三种策略
    1. 平均满意度（Average Satisfaction，与总体满意度等价）

$$GR_i = average(r_{i,j}) = \frac{\sum_{j=1}^{n} r_{i,j}}{n}$$

    2. 最小不满意度（Minimum Misery），即基于最低得分评估不满意度

$$GR_i = min(r_{i,j})$$

    3. 最大满意度（Maximum Satisfaction），即基于最高得分评估满意度

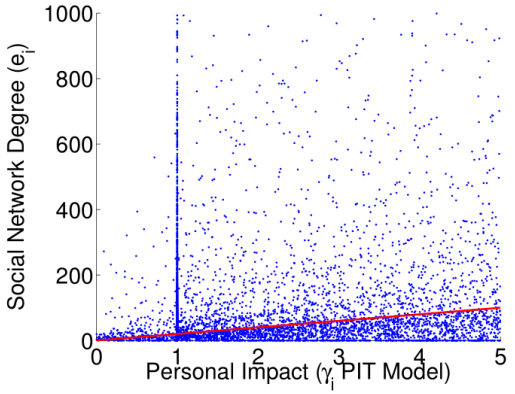
$$GR_i = max(r_{i,j})$$

成员投票的分量问题

- 还有一个重要问题没有解决：每个成员的投票是否等价？
  - 一般情况下，我们将所有人的投票视作等价的
    - 在正式的投票环节中，这个约定尤为重要且常见
  - 然而，在现实的群体决策中，权威人士的意见往往更为重要
    - 显然，不同成员的投票分量往往并不等价
    - 如何衡量一张票价值几何？

- **成员投票的分量问题**

- 常见的决定选票分量的策略
  - 一般而言，我们根据成员的权威性赋予不同权值
    - 例如，在专业问题上，权威人士的投票比一般人士更有价值
    - 例如，论文的评委问题：一般人甚至不具备投票资格
  - 如何衡量成员的权威性?
    - 社交权威性：如度、介数、 PageRank等
    - 主题权威性
  - 投票公平性问题?



Xingjie Liu, et al., Exploring Personal Impact for Group Recommendation, CIKM-2012

- **成员投票的分量问题**

- 常见的决定选票分量的策略
  - 与此同时，性格因素同样会影响投票分量
  - 例如，对于倾向于和他人合作的成员，他们的意见往往是可以被“牺牲”或“忽略”的
  - 相反，那些“难伺候” 的成员，他们的意见更容易被采纳
    - 按闹分配?

Juan A. Recio-Garcia, et al., Personality Aware Recommendations to Groups, RecSys-2009

- **成员投票的分量问题**

- 常见的决定选票分量的策略
  - 然而，老实人也不见得一直愿意受欺负
    - 如果存在多轮决策，意见总是不被满足的情况下，会积累不满情绪

$$dbrm(u,i) = \frac{1}{|\sum_{v \in G} t_{u,v}|} \sum_{v \in G \wedge v \neq u} t_{u,v}(pred(v,i) + p_v) + \alpha(\mathbf{1} - \mathbf{s_v})\mathbf{p_v}$$

    - 如果某人的意见总是得不到满足，那么他/她的权重将得到提升
      - S用于表示对上一轮推荐结果的满意程度（Satisfaction）
      - 偏差越大，失望越多，积累的权重就越高，下一轮就越可能获得满足

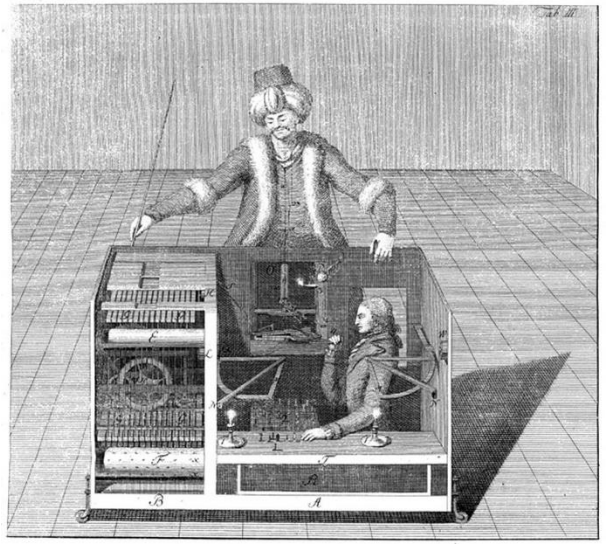
LARA QUIJANO-SANCHEZ, et al., Social Factors in Group Recommender Systems, ACM TIST-2013

本节目录

- 社团挖掘
  - 层次聚类技术
  - 划分聚类技术
- 社团决策
- **众包与群体智能概述**

- **社会智能的新时代**

- 人工智能盛行的今天，人的价值更进一步凸显
  - 一方面，人工智能依赖于海量数据进行学习
    - 高质量难以获取且代价高昂
    - 大量数据缺乏可靠标注信息
  - 另一方面，人天然地具有提炼、学习知识的能力
    - 有趣的是，社交网络正是人们获取知识的重要渠道
- 如能从这种社交行为中有效提炼“群体智能”，将对人工智能技术起到极大补充和助力

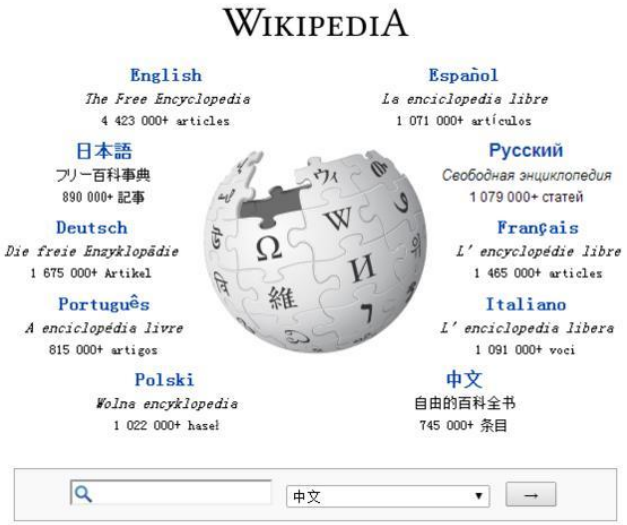


- **群体智能能帮我们做些什么**

- 如何运用群体智能？体现在不同的层面
  1. 将个体视作群体的传感器与信息传递渠道
     - 地震等自然灾害的信息快速收集
     - 重大公共事件的有效应对
  2. 将个体视作群体的信息分析与处理器
     - 经典案例：波士顿爆炸案的“人肉缉凶”
       - 小心背后潜藏的“人肉”伦理问题

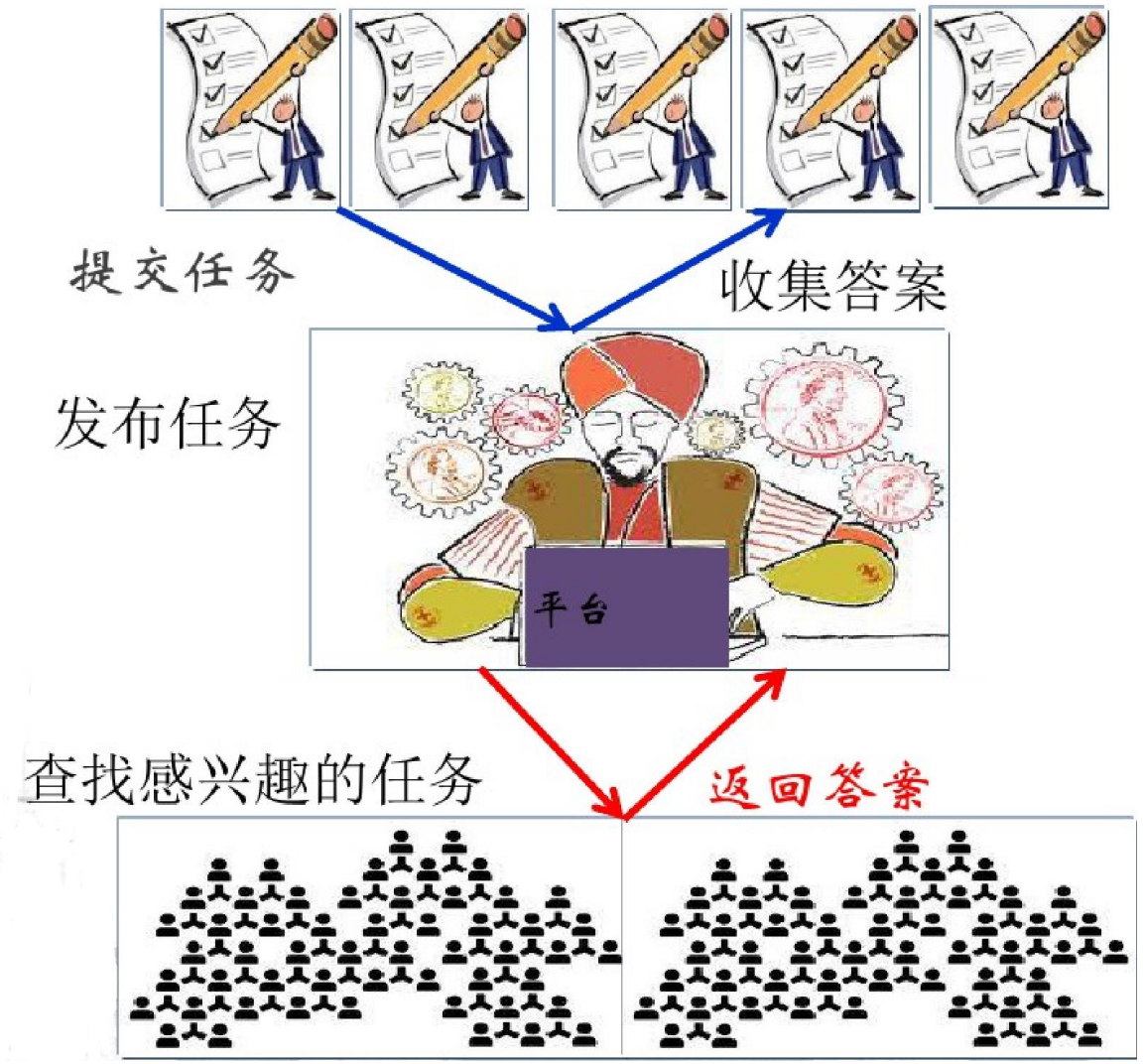
• 群体智能与众包

- 事实上，前面的例子可以归纳为“众包”（Crowdsourcing）思想的应用
  - 众包与外包的区别何在？后者面向已知、特定的雇员，而前者的人员不定，数量众多
  - 重要的是“开源”思想的体现！
    - 众包的精髓来自于“开放”、“共享”
    - 最成功的应用莫过于大众维护的百科类网站，如Wikipedia
    - 众包带来的隐患：质量问题
      - 有趣的案例：百度百科——中国十大名校



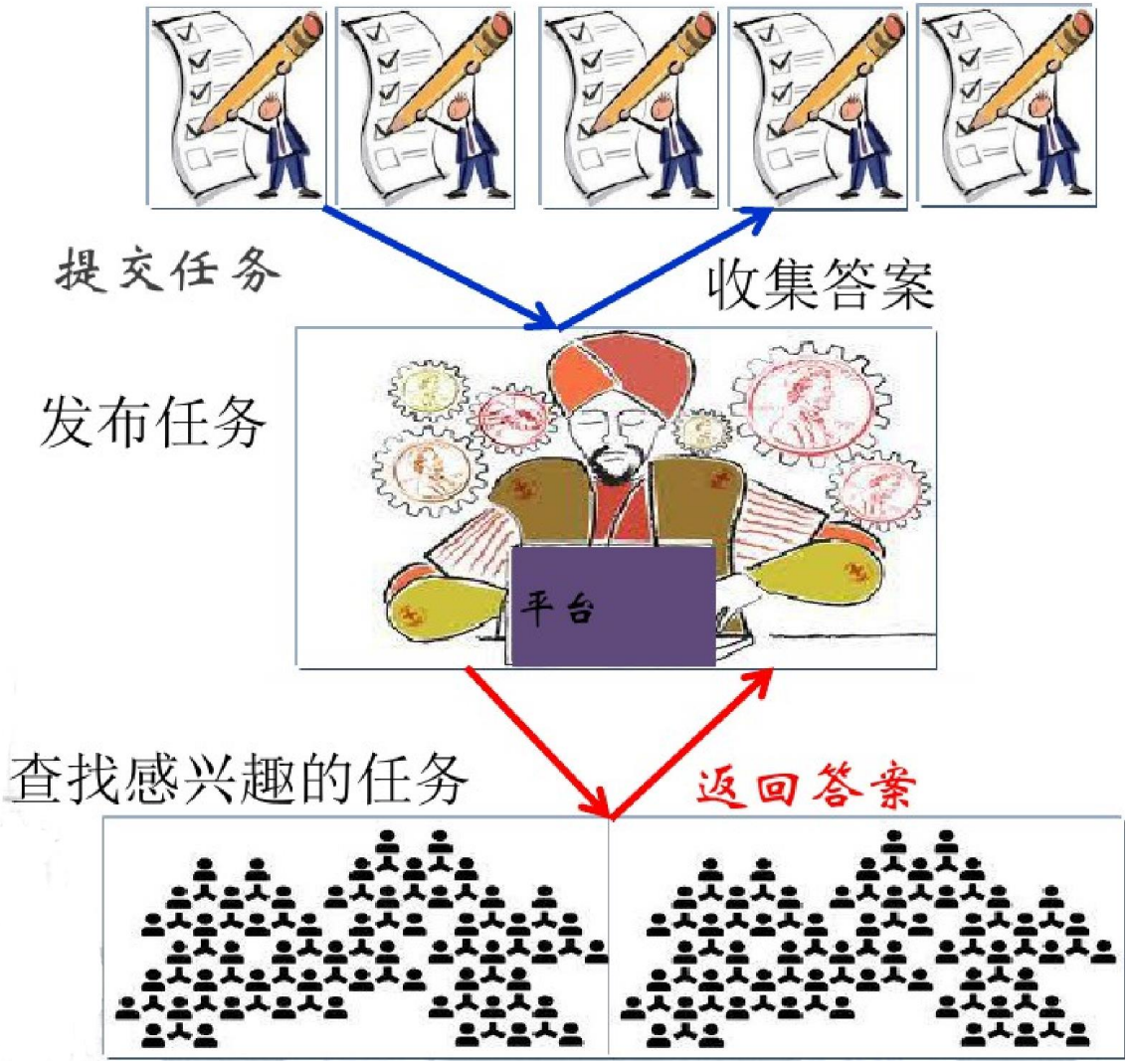
- **群体智能与众包**

- 一个众包任务应包含三方面基本元素
  - 请求者：
    - 提交任务、回收答案并评估结果
  - 工人
    - 针对任务开展工作并提交
    - 将获得与答案满意度匹配的收益
  - 任务平台
    - 任务发布与管理



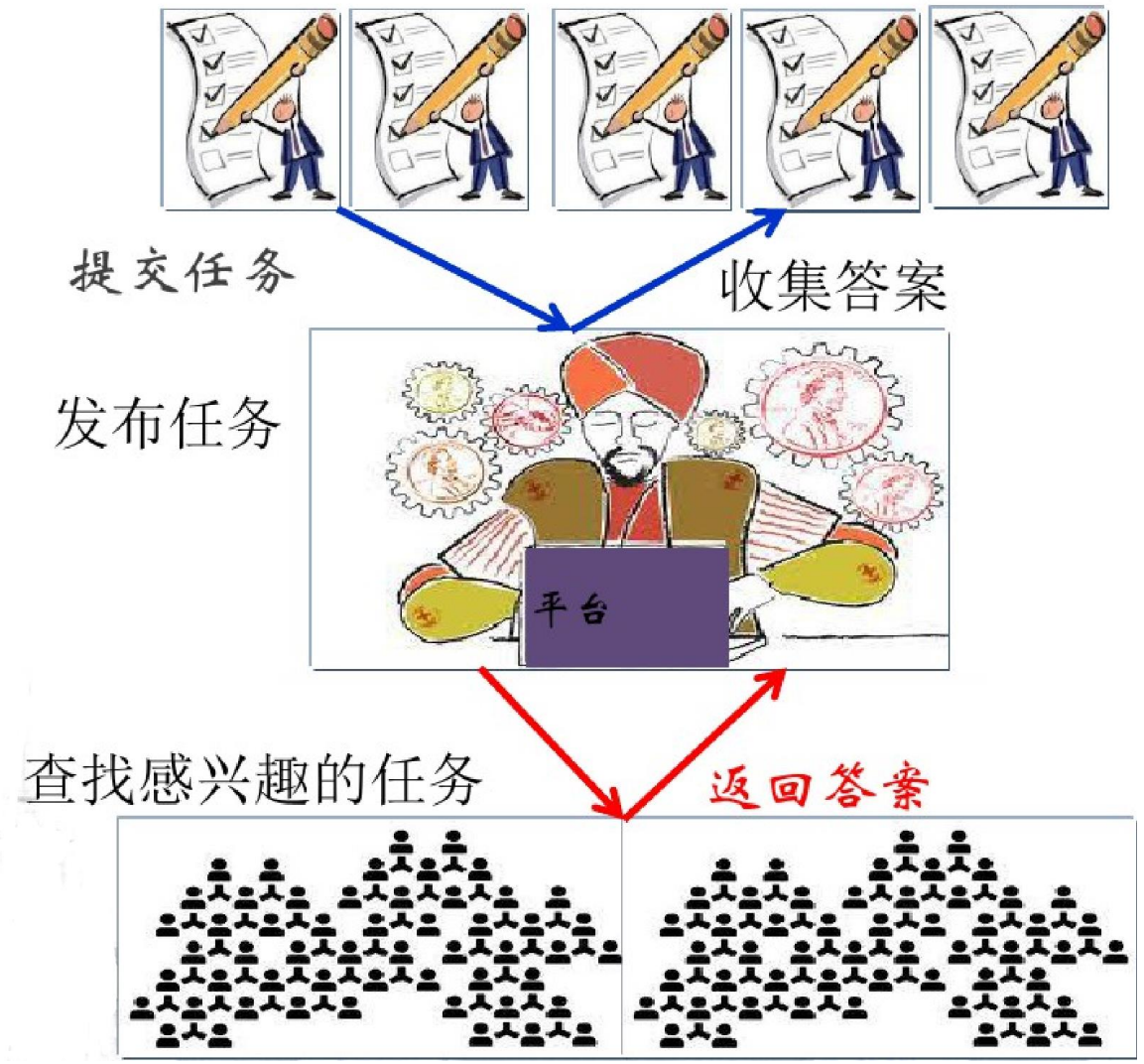
• 群体智能与众包

- 如何保证众包结果的可靠性？
  - 首先对任务需要的专业性进行评估
  - 如果任务需要较强的专业性，则需要对工人的资格进行评判
    - 例如，审稿任务中的专家筛选
    - 优点：确保答案质量
    - 缺点：需要额外成本，影响完成效率，且如何保障评判的公平性？



- **群体智能与众包**

- 如何保证众包结果的可靠性？
  - 其次，对于大众化任务，也需要质量控制
  - 基本的质量控制：结果的一致性
    - 即对于同样一条信息，如果两个人标注统一才可以得分
    - 相应的，大多数人的一致性标注将作为最终答案

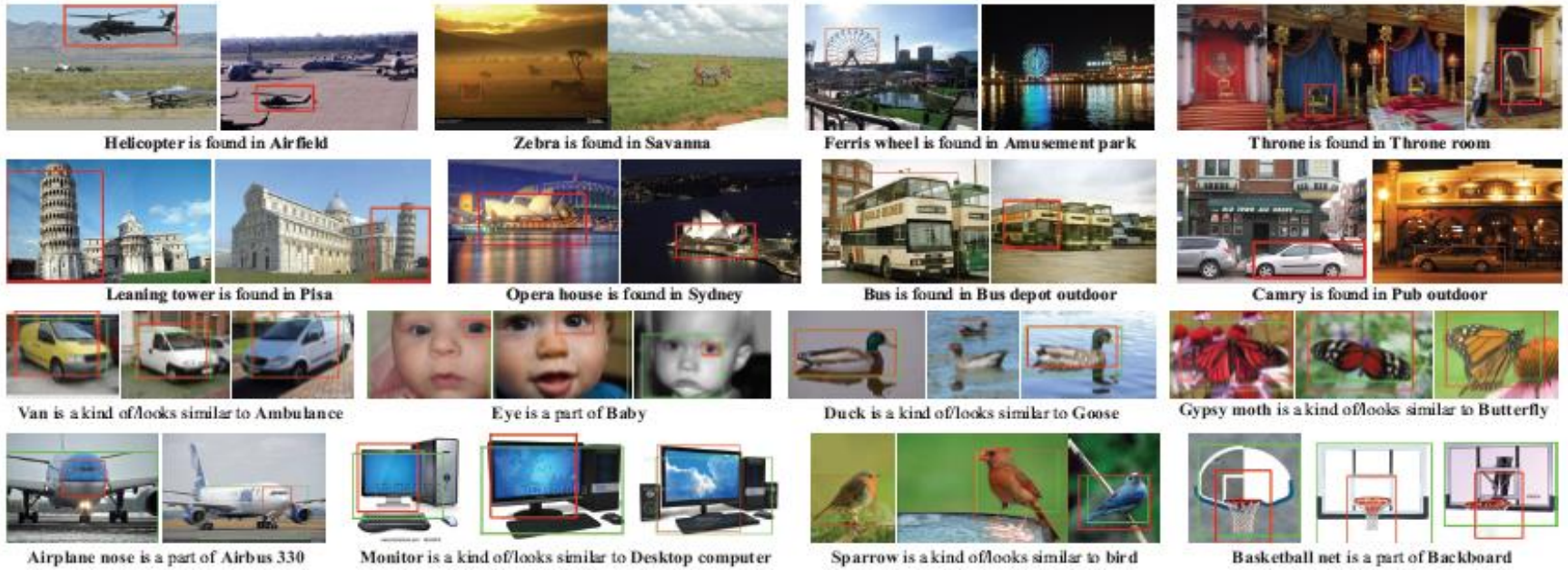


• 群体智能与众包

- 如何保证众包结果的可靠性？
  - 除了输出的一致性，输入一致性也需纳入考虑范畴
    - 一些主观性较强的标注（如多模态信息）发挥空间大，很难达成完全一致，此时需要反其道而行之
    - 例如，一度流行的“我画你猜”App
    - 又如，给出双方对于同一首歌打出的标签，让他们猜对方标注的是哪一首歌，如果一致则答案有效

• 群体智能与众包

- 如何保证众包结果的可靠性?
  - 更进一步，还可以借助机器学习技术辅助，半自动地获得标准答案
    - 通过迭代优化，逐步完善模型学习的精度，并相应地获得高质量标注



- **群体智能与众包**

- 如何保证众包结果的可靠性?
  - 最后，当然不要忘记，通过奖惩机制和声望系统激励工人
    - 根据工人的历史提交情况进行回报，质量越高收益越高
    - 反之，则降低奖励甚至进行惩罚
      - 例如，剥夺工人未来的工作机会

本章小结

社团挖掘

- 基于聚类的社团挖掘方法
  - 层次聚类技术
  - 划分聚类技术
  - 确定社团数的模块度原则
- 社团决策与表决问题
- 众包与群体智慧概述

tongxu@ustc.edu.cn

写在最后

- 在一万五千年前，大腿骨折是致命的，只能等在原地，被野兽吃掉。但，这根股骨愈合了，它是人类文明诞生的标志。
- 意味着它受伤以后，有人为他处理伤口，有人为他提供水和食物；有人保护他，不受野兽的攻击。
- 团结，延续着文明的火种。

——《流浪地球2》

tongxu@ustc.edu.cn

写在最后

- 在一万五千年前，大腿骨折是致命的，只能等在原地，被野兽吃掉。但，这根股骨愈合了，它是人类文明诞生的标志。
- 意味着受伤后，有人为你提供水和食物；有人保护他，不受野兽的攻击。
- 团结，延续着文明的火种。

社会网络，是人类团结协作、守望相助的纽带

也是铸就文明的基石

——《流浪地球2》

tongxu@ustc.edu.cn