# 结构平衡性

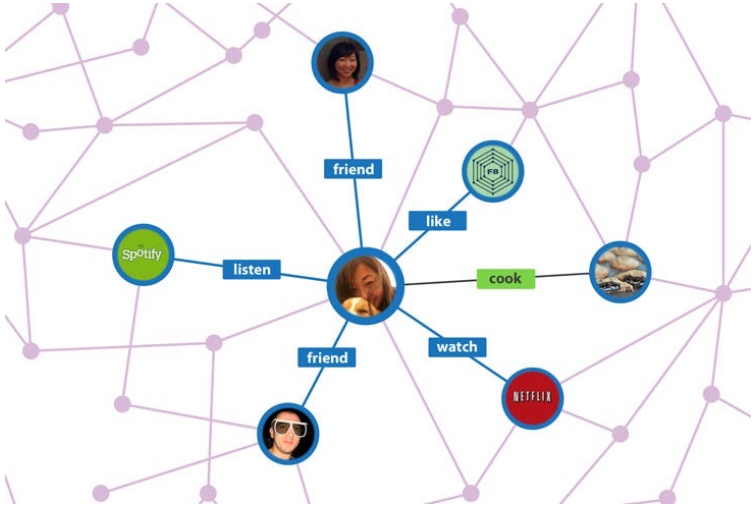
姜少峰

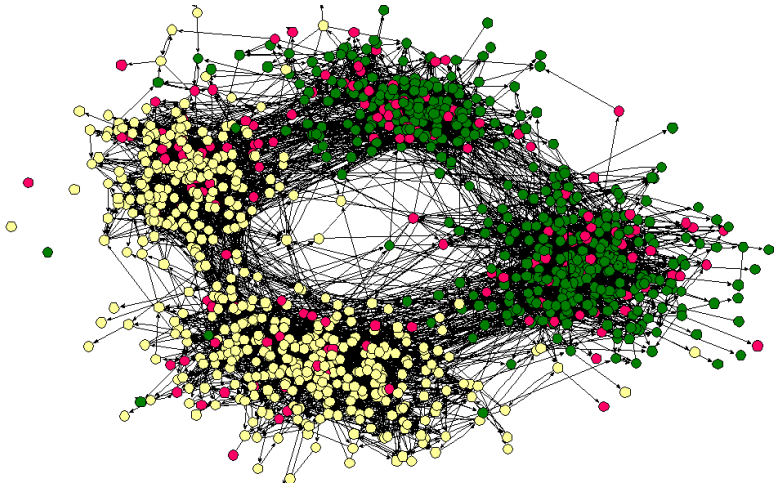

北京大学前沿计算研究中心

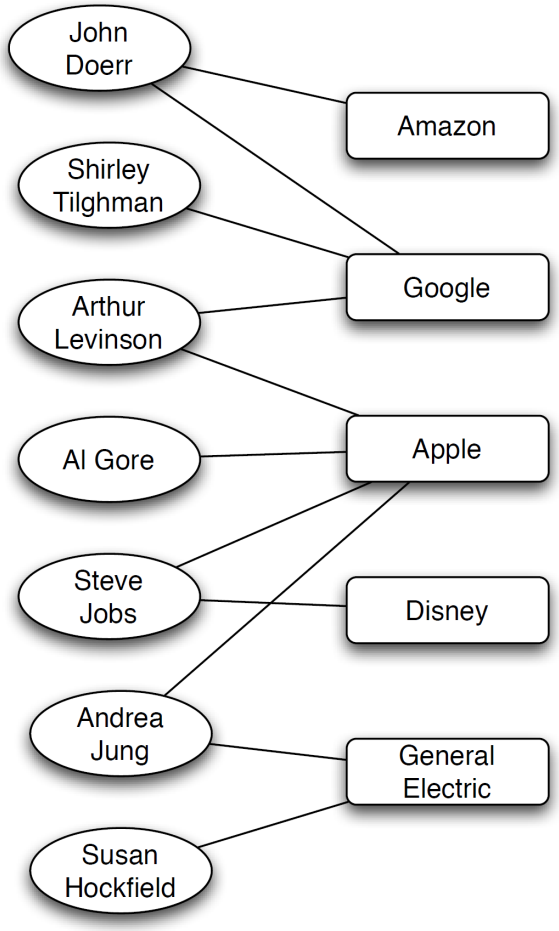
Center on Frontiers of Computing Studies, Peking University

# 正关系与负关系

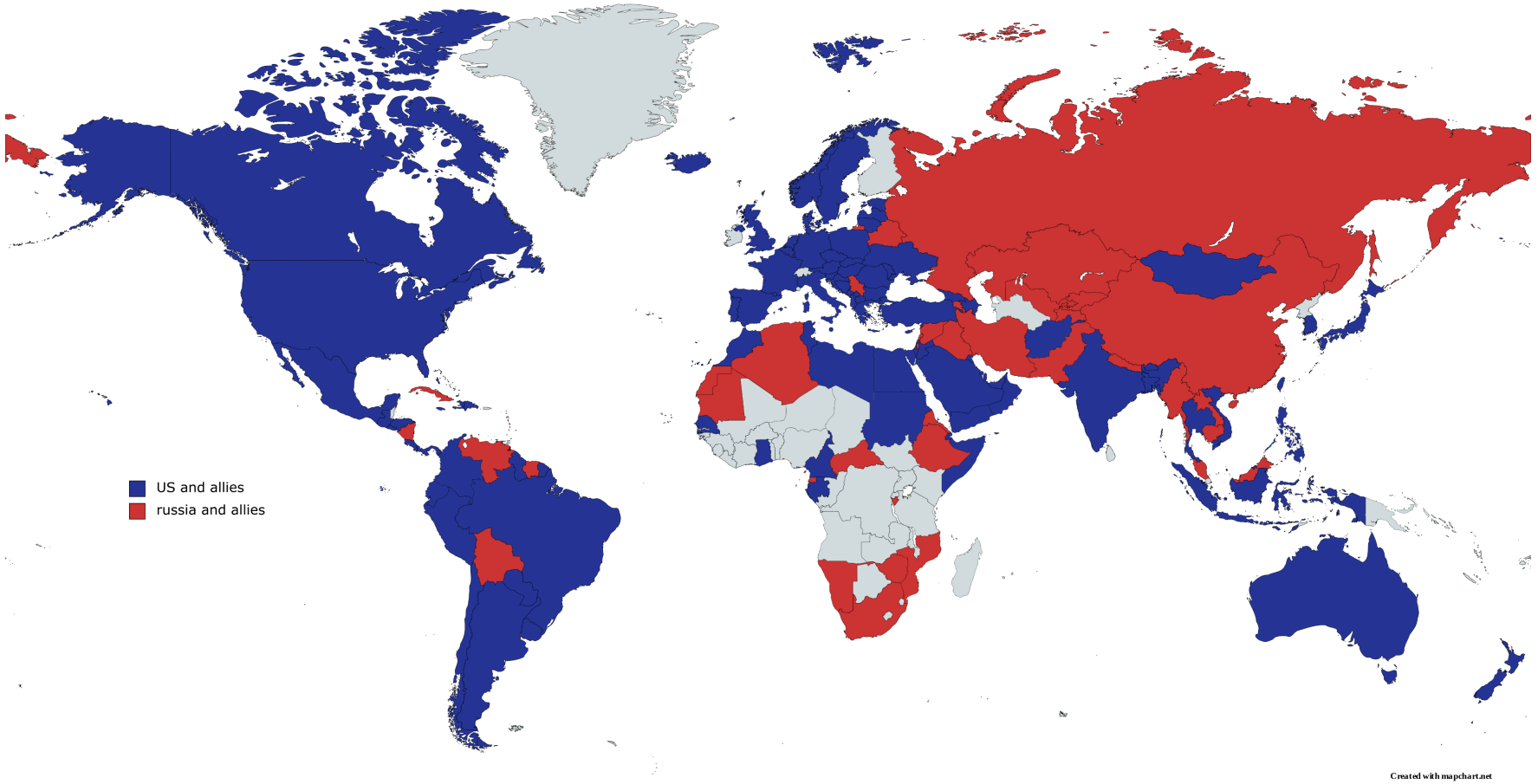
- 之前，我们考虑的网络的边都是“正面”含义的
  - 合作，友谊，信息共享，同属某个团体 etc.



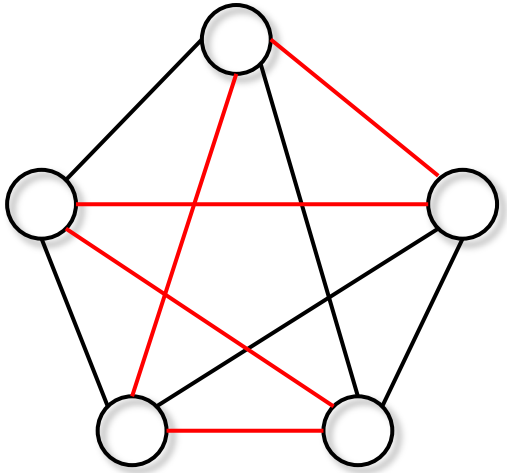


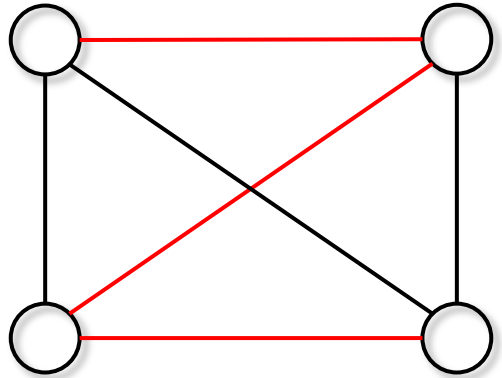
- “负面”关系也很典型；我们探究“负面”关系对网络结构的影响



# 模型

- 简化模型：任何两点都有连边，且有+或者-标识
  - +代表两个端点是朋友，-代表是敌人
  - 没有不认识/不持立场的情况
  - 称这种图为（完全）“标注图”

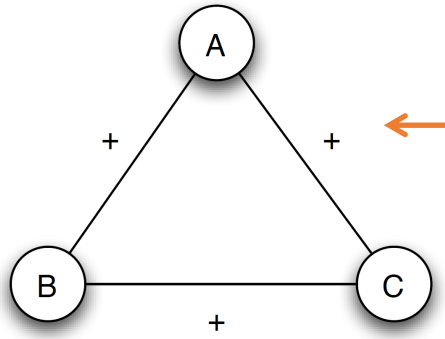


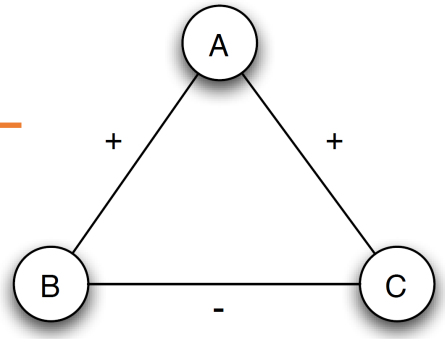


# 平衡关系 Heider 1940, Cartwright and Harary 1950s

- 考虑三个点所有的+/-组合情况，有的关系更加“合理”

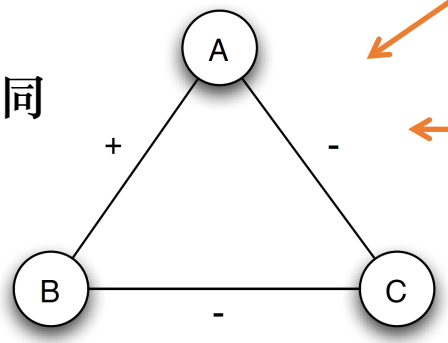
ABC互为朋友，关系平衡

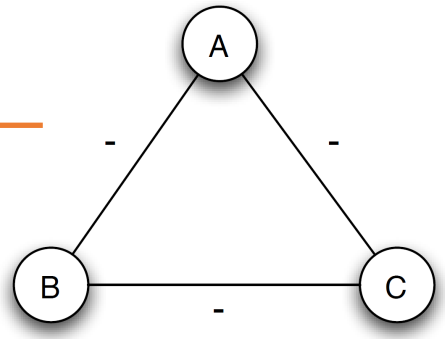



A和BC是朋友，但是BC敌对，关系不平衡：A可能撮合BC成为朋友；也可能A与BC中一个结盟对抗另一个

A和B是朋友，但是他们有共同的敌人C：关系平衡



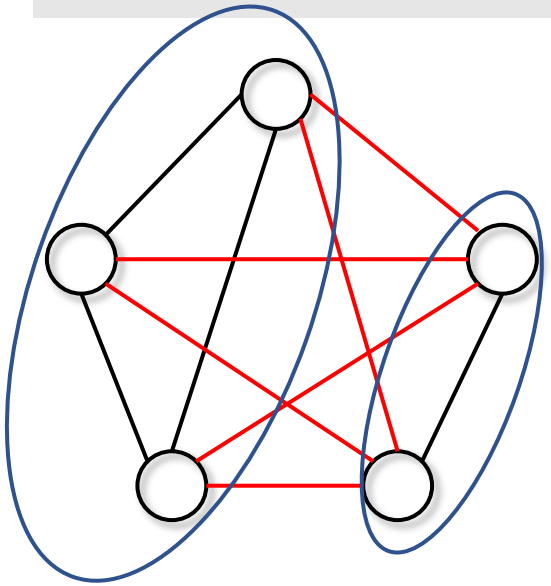

ABC互为敌人，关系不平衡：可能会导致二者结盟对抗第三者

结构平衡性质：对于任何三个节点，与他们相连的三条边要么全+要么恰有一条+

# 平衡定理

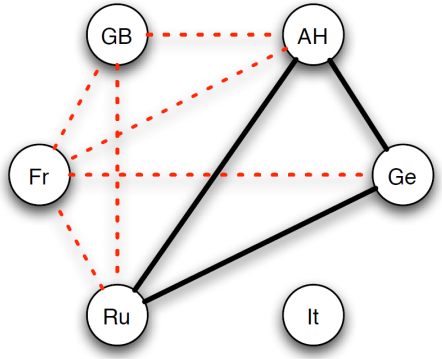
Frank Harary, 1953

如果一个（完全）标注图是平衡的，则要么它所有节点两两都是朋友，要么它的节点可以分为两组X和Y，其中X和Y组内的节点两两是朋友，而X中每个点都和Y是敌人

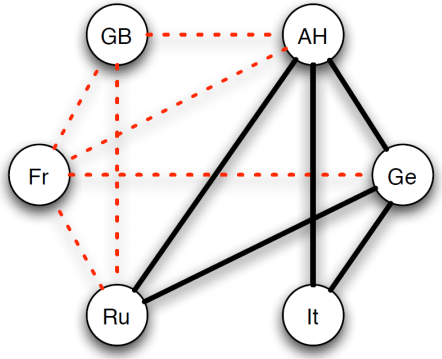


极化网络中，人们的分组经常很极端，难以改变立场

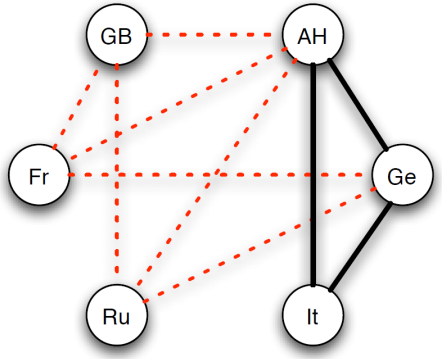
# 用结构平衡解释国际关系



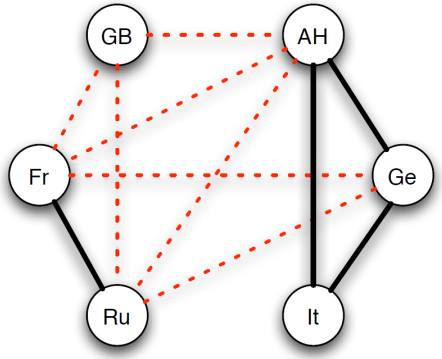

(a) *Three Emperors' League 1872–81*



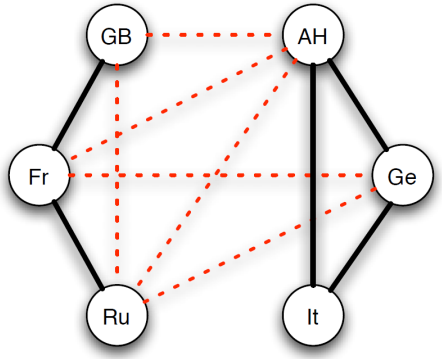

(b) *Triple Alliance 1882*



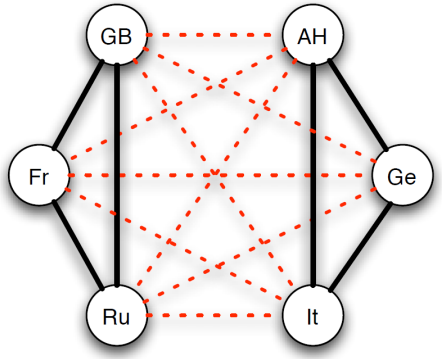

(c) *German-Russian Lapse 1890*




(d) *French-Russian Alliance 1891–94*



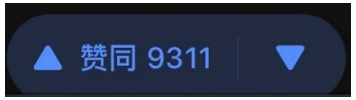

(e) *Entente Cordiale 1904*




(f) *British Russian Alliance 1907*

# 社交网络中的正负关系

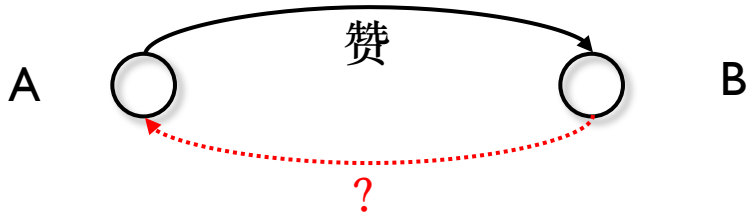
- 很多社交网站都有对于别的用户“赞同”“不赞同”的功能，这构成了一种用户间正负关系的描述




知乎


reddit

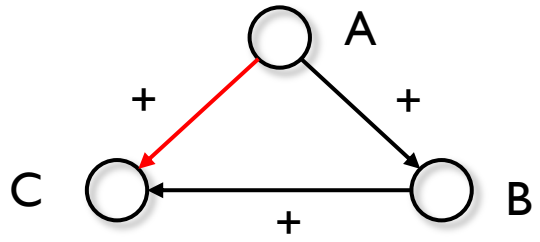
- 不同点：
  - 关系是有向的：A对B的+/-不代表B也有同样的+/-，甚至可能没有B->A联系



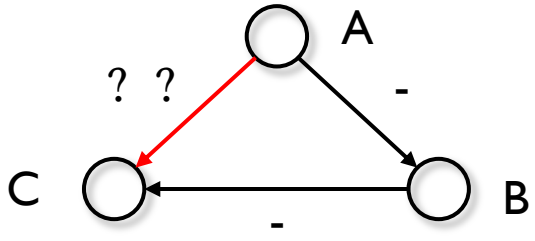

A赞了B，但是B甚至可能不认识/不在意A

# 有向图的平衡关系？

- 另一个不同点在于关系的平衡更加微妙复杂




A对B表达认可，B对C表达认可，那么A大概也会认可C




A对B表达不认可，B对C表达不认可，那么A会如何评论C?

- A对C表达认可：与--+类型的无向图平衡原因类似
  - 例如，假设ABC每人属于2个阵营的一个，那么这说明AC大概是同一个阵营的（B是AC的共同敌人）
- A对C表达不认可：如果A不认可B是因为B的水平低，B不认可C也是同样的原因，那么A应该更加不认可C

# 平衡定理的证明

# 平衡定理

Frank Harary, 1953

如果一个（完全）标注是平衡的，则要么它所有节点两两都是朋友，要么它的节点可以分为两组X和Y，其中X和Y组内的节点两两是朋友，而X中每个点都和Y是敌人

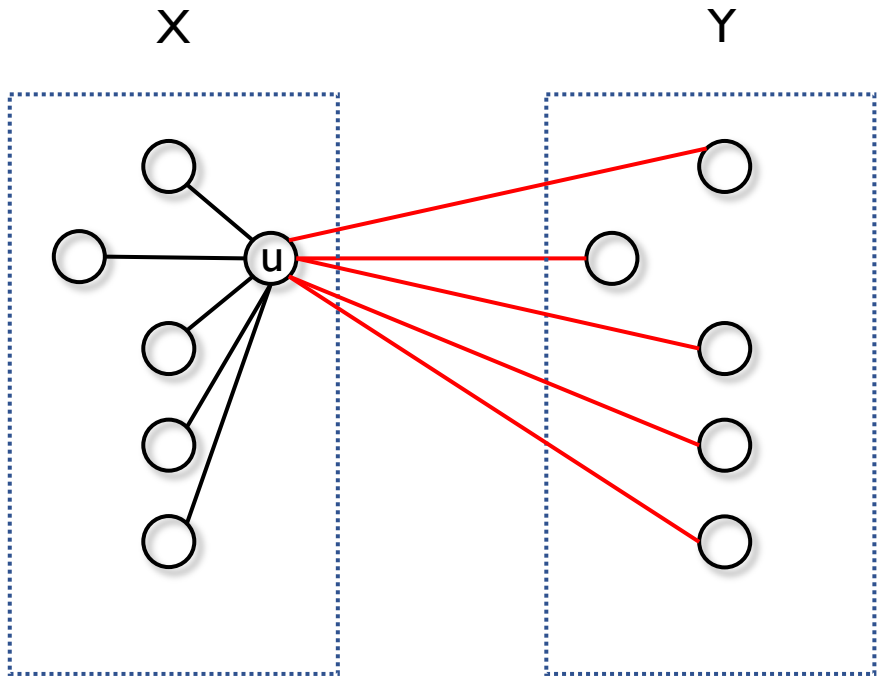
数学语言翻译：

对于满足结构平衡性质的标注图

- 要么图中边全是+
- 要么可以将节点分成X、Y，使X、Y内部边全+，跨越X、Y的边全-

# 定义划分

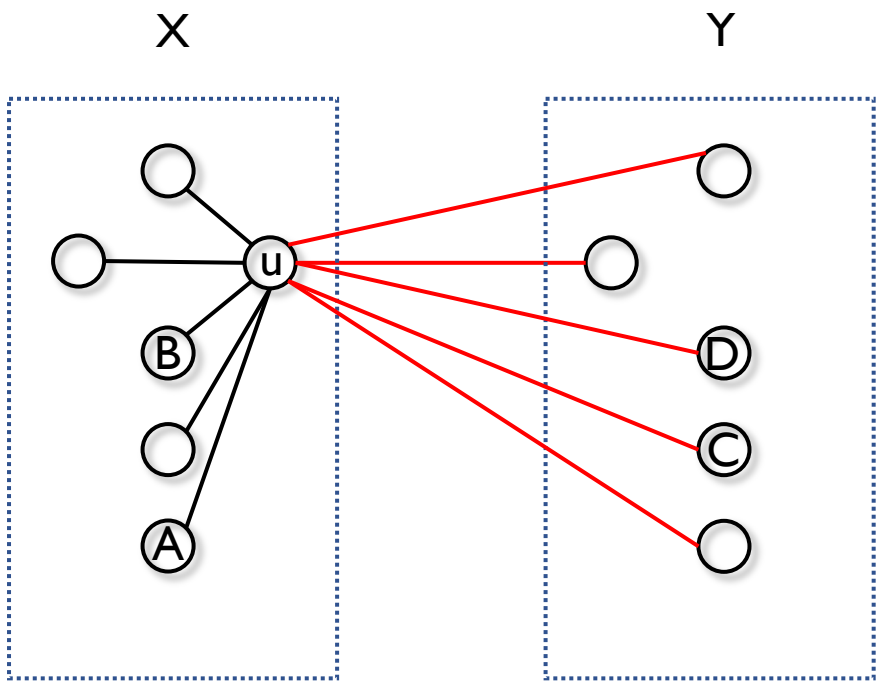
- 考虑图中存在至少一条-边的情况 （全+自动满足定理）
- 我们试图构造X和Y
  - 必然存在一条-边，任意找到一条-边u-v
  - 将u放入X
  - 将所有u的+边端点放入X，-边端点放入Y




结构平衡性质：对于任何三个节点，与他们相连的三条边要么+++要么-+-

# 验证定义

- 验证X、Y内部都是+：
  - AB = +
  - CD = +
- 验证X、Y之间都是-：
  - AC = -

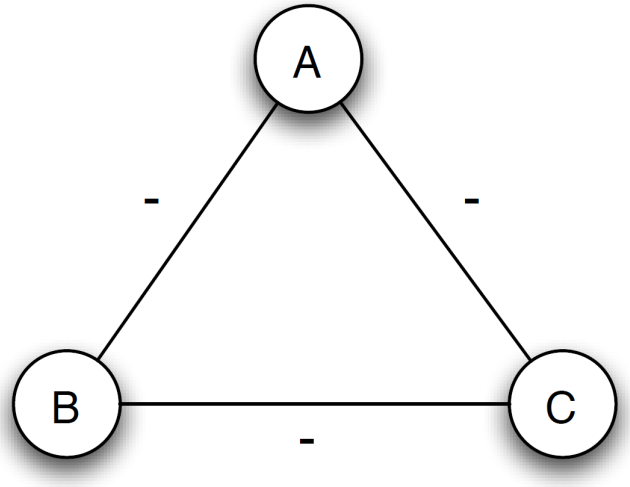



结构平衡性质：对于任何三个节点，与他们相连的三条边要么+++要么-+-

# 弱平衡性质

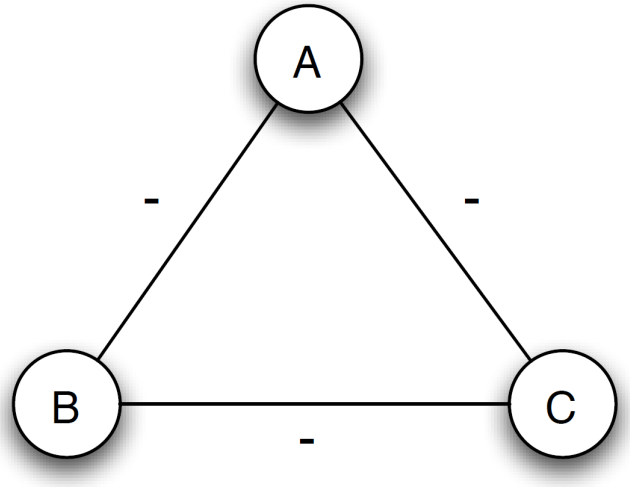
# 弱平衡性质

之前的定义：




ABC互为敌人，关系不平衡：
可能会导致二者结盟对抗第三者

James Davis 1967：




三个人完全敌对，缺乏动力/途径化敌为友，
因此事实上该关系也是较为稳定的

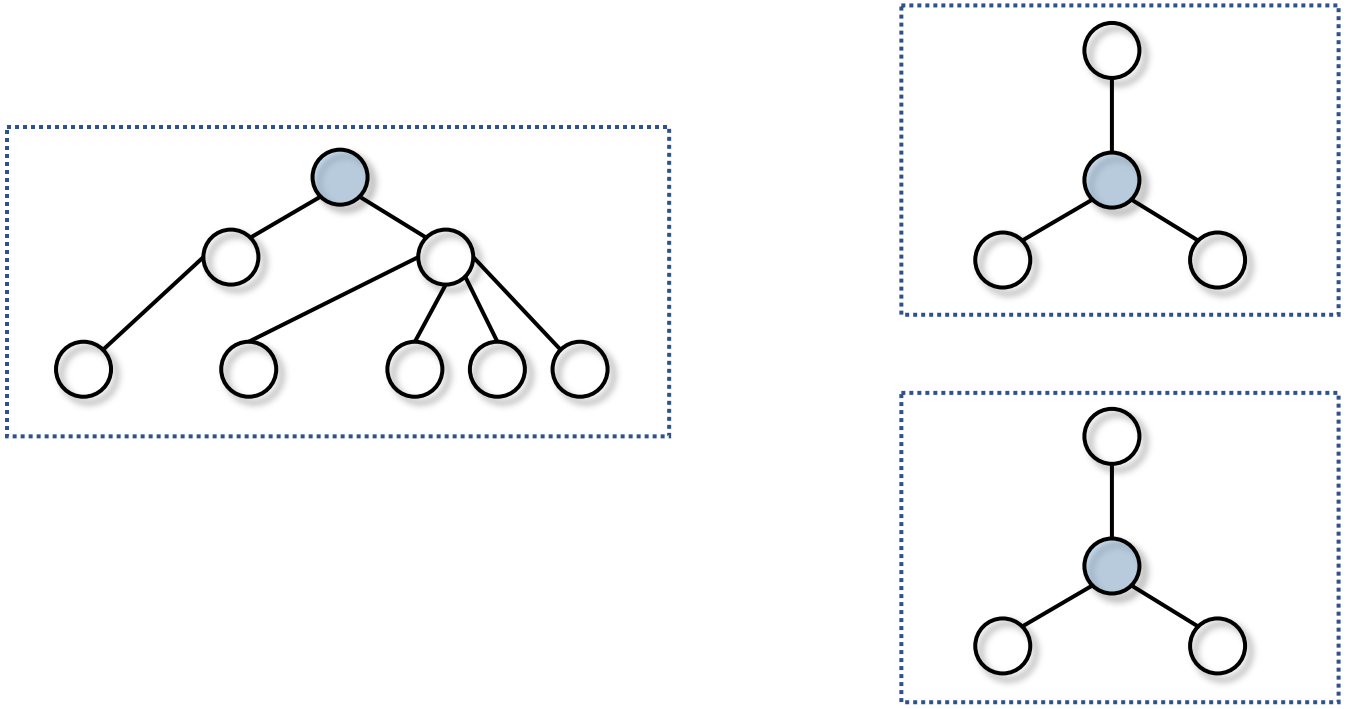
弱平衡性质：任何三个节点，都不存在两个正关系和一个负关系的连接方式

# 弱平衡网络结构定理

如果一个（完全）标注图是弱平衡的，则它的节点可以分为若干组，使得组内节点两两是朋友，而跨组的节点两两都是敌人

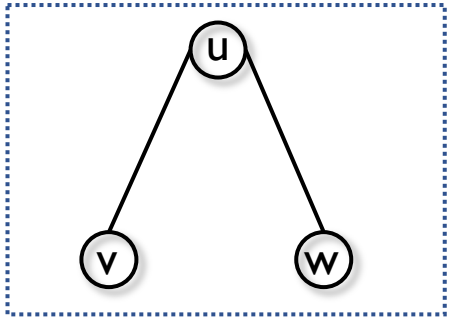
# 证明：定义分组

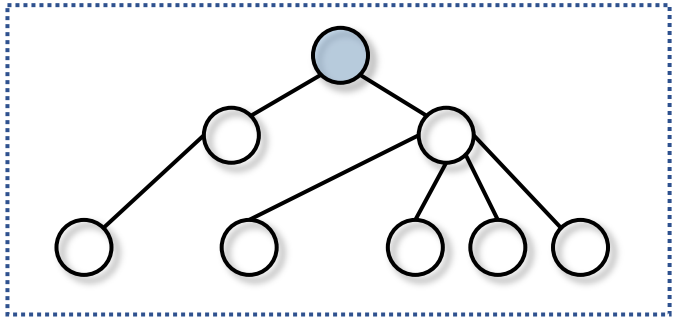
- 定义分组：
  - 任取一个节点，从该点开始在全部+边上做BFS，将能到达的节点划为一组
  - 任取一个尚未归组的点重复上述过程，直到所有点都归组



# 证明：组内

- 组内都是+边
  - 重要性质：对于节点uvw，已知u-v和u-w为+，那么v-w必为+
  - 在用于定义组的BFS树上反复使用该性质

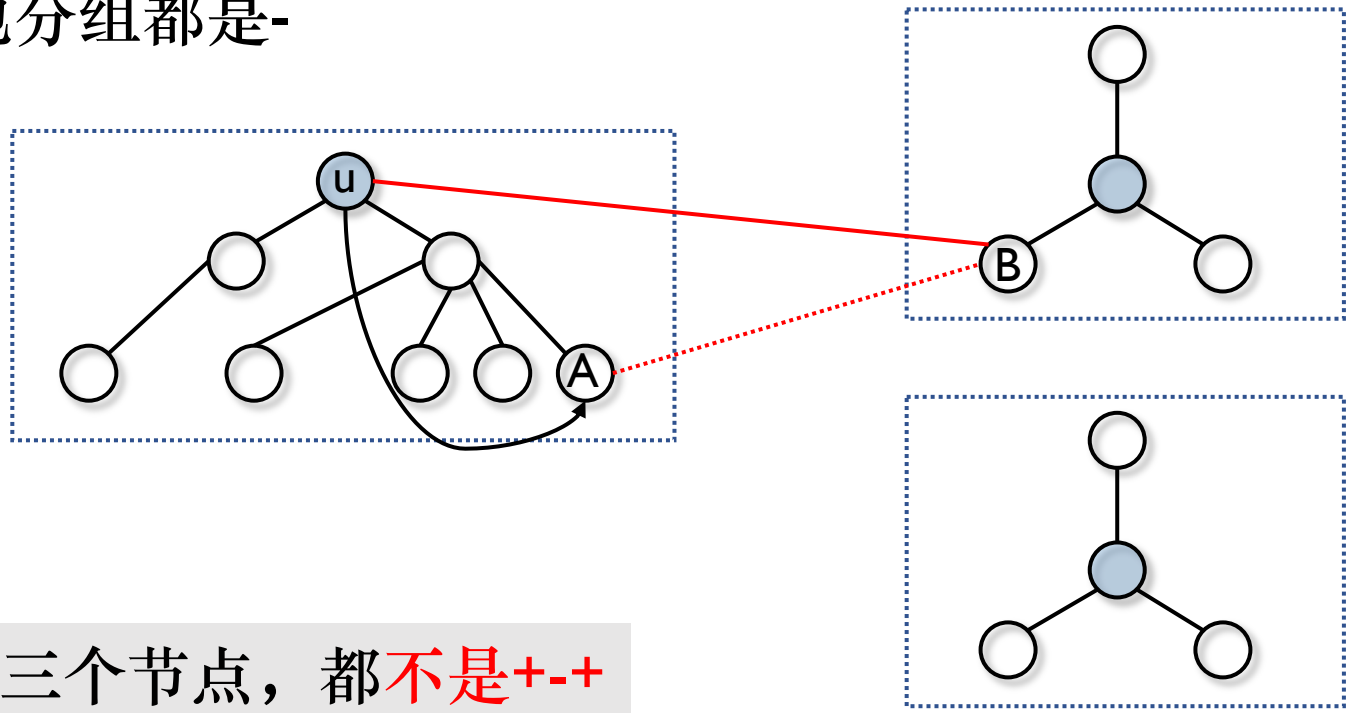


弱平衡性质：任何三个节点，都不是+-+

# 证明：组间

- 根据该分组规则，组间无+边
  - 考虑A和B之间的边
  - 观察：根节点到其他分组都是-




弱平衡性质：任何三个节点，都不是+-+

# 更一般的模型：非完全图 (Harary 1953)

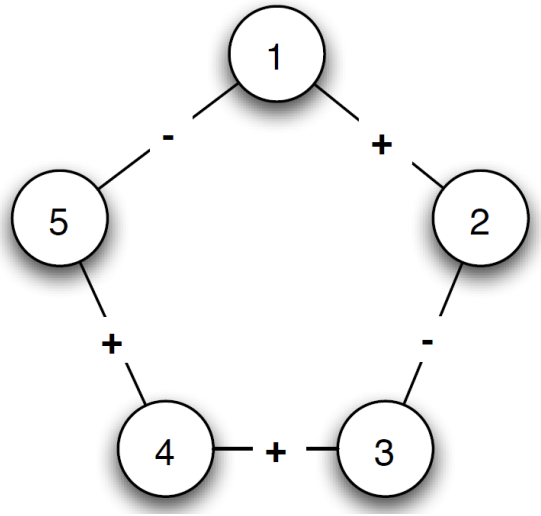
- 之前（强）平衡关系的定义需要完全图，现在考虑一个自然扩展
- 两个定义方法
  1. 称一个标注图是平衡的，如果可以将其补全成平衡的完全图
  2. 称一个标注图是平衡的，如果可以将其分成两个对立的集合（-只在集合间）
- 由于平衡定理，这两个定义事实上是等价的

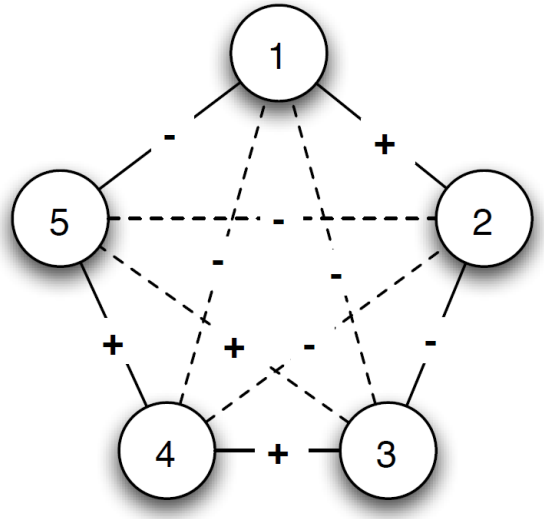
**平衡定理（完全图）**：如果一个（完全）标注图是平衡的，则要么它所有节点两两都是朋友，要么它的节点可以分为两组X和Y，其中X和Y组内的节点两两是朋友，而X中每个点都和Y是敌人

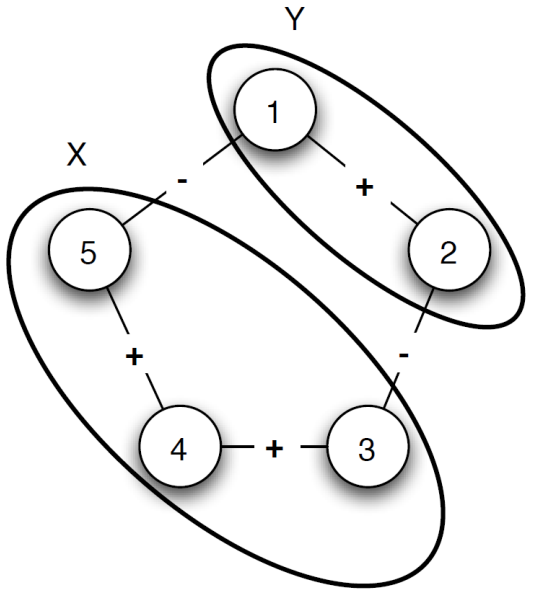
结构平衡性质：对于任何三个节点，与他们相连的三条边要么+++要么-+-

- 1->2？ 2->1？

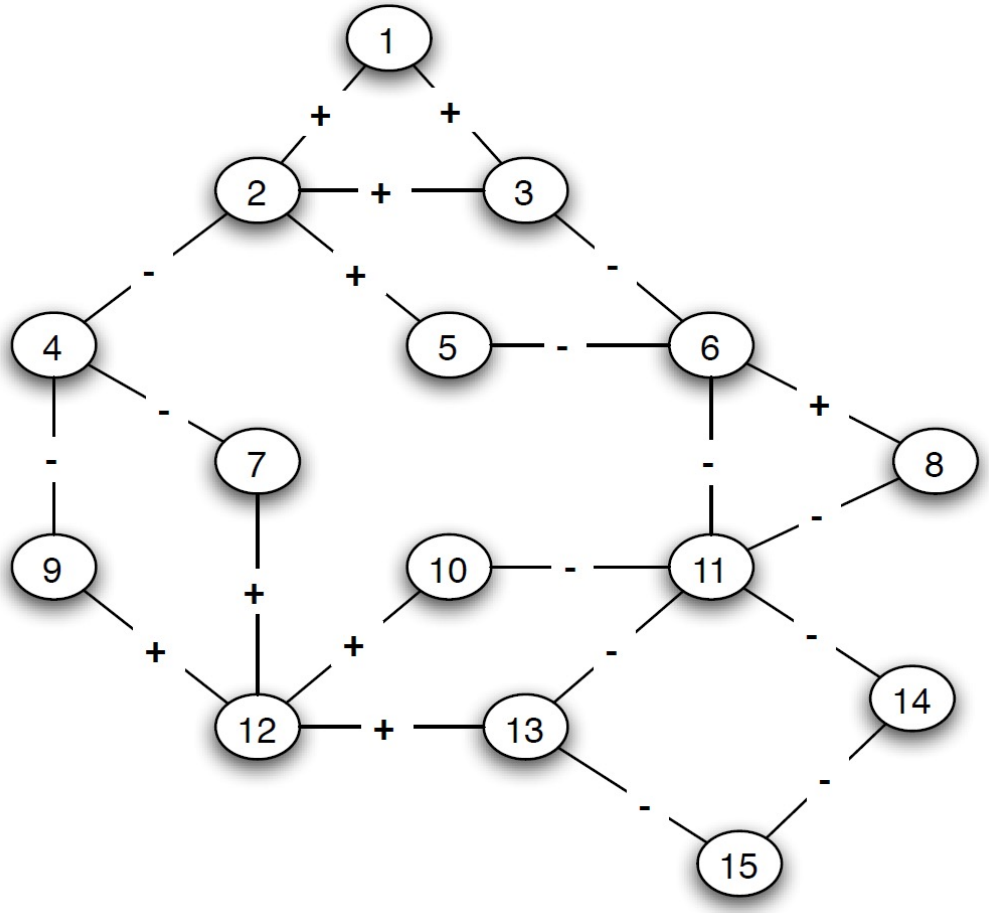
# 非完全图的平衡：示例



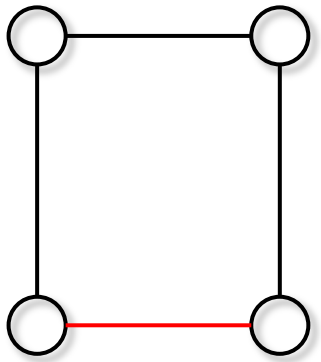
# 如何判断?



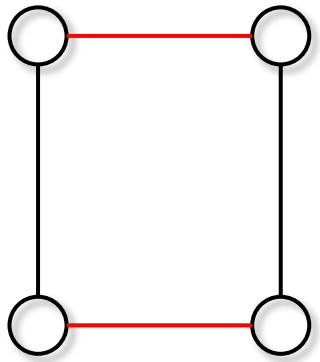
# 平衡定理（非完全图）

一个无向图是平衡的，当且仅当不存在含有奇数个-边的简单回路

不满足



满足

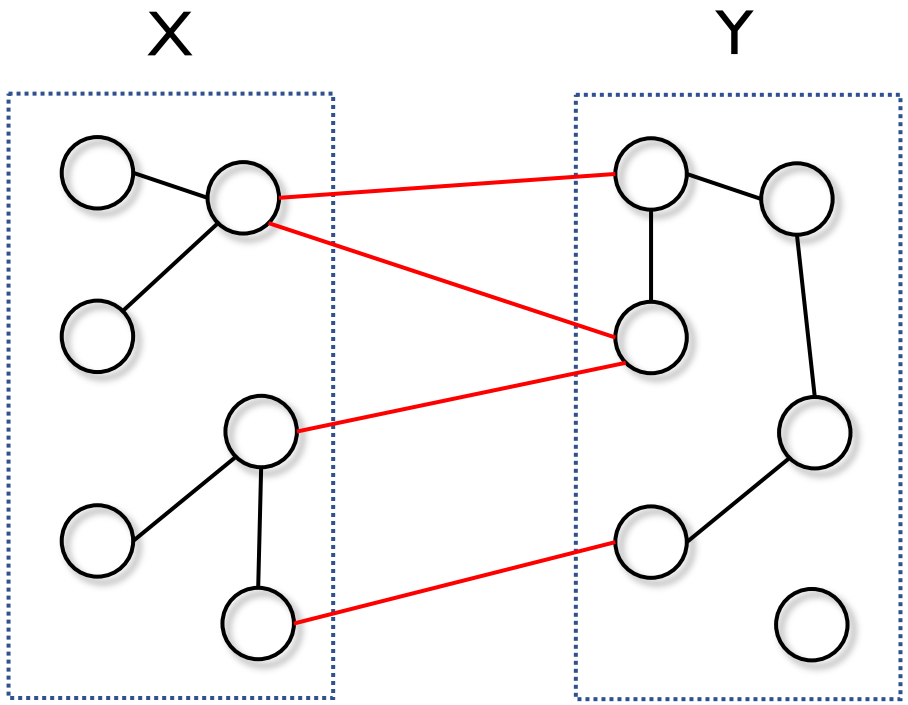


称一个标注图是平衡的，如果可以将其分成两个对立的集合（-只在集合间）

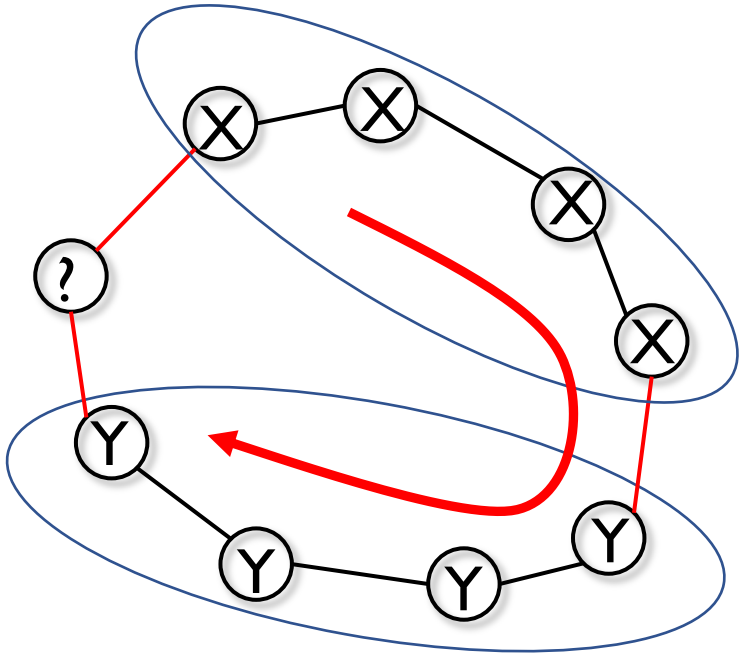
称一个标注图是平衡的，如果可以将其补全成平衡的完全图

# 平衡 -> 不存在奇数-边回路

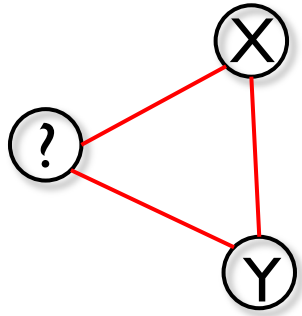
称一个标注图是平衡的，如果可以将其分成两个对立的集合（-只在集合间）




划分示例



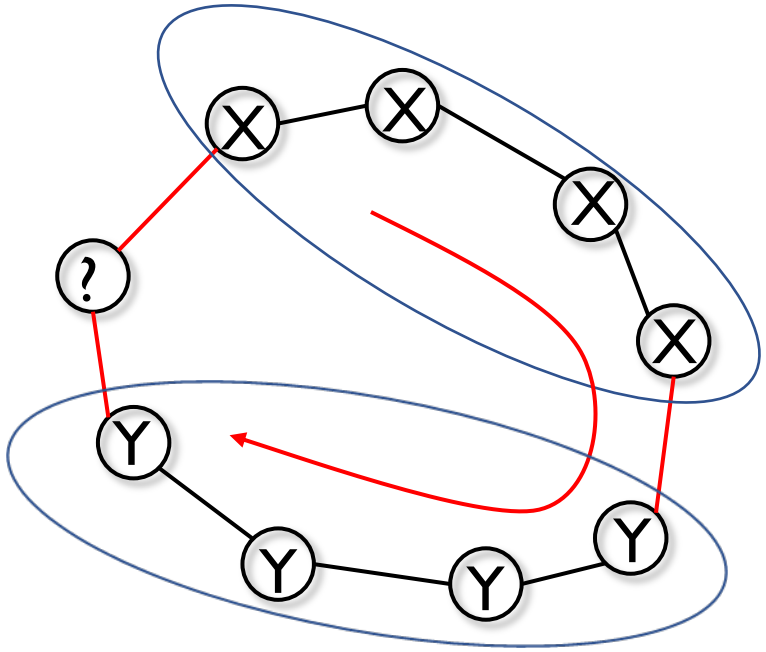

假设有一条奇数-路



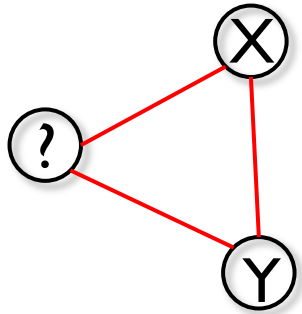

缩点：合并全+子路径

# 平衡 -> 不存在奇数-边回路

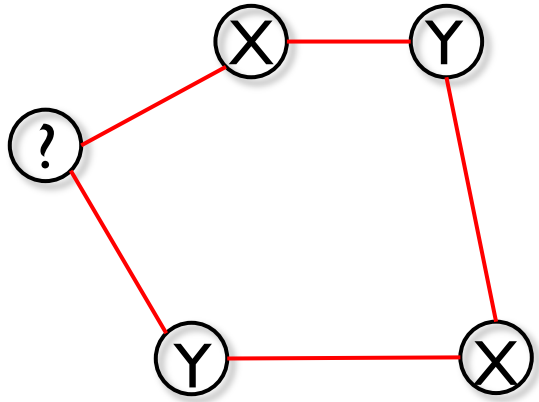
- 假设存在一条奇数-路
  - 缩点：把（最大）全+子路径缩成一个点
  - 得到：纯-回路




假设有一条奇数-路




缩点：全+子路径




每条边的两端点都分别属于XY；
如果是奇数圈，最后会导致矛盾！

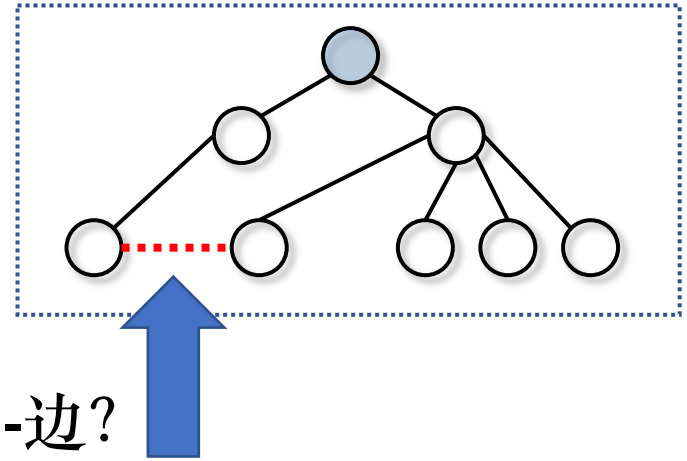
# 不存在奇数-边回路 -> 平衡

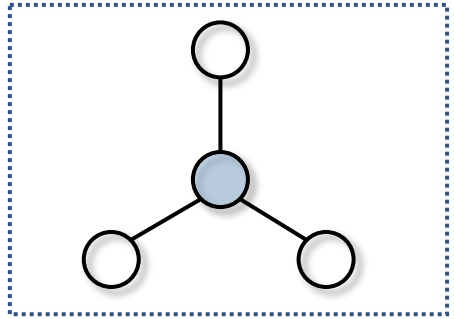
称一个标注图是平衡的，如果可以将其分成两个对立的集合（-只在集合间）

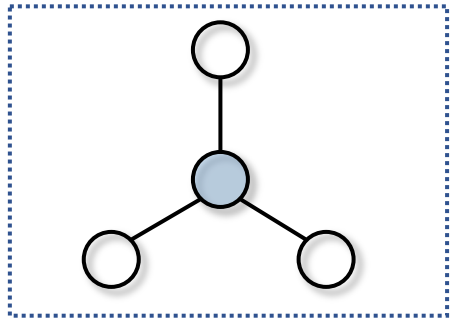
- 为证明平衡，只需要验证定义，即给出一种划分
- 总体策略：缩点，将缩点后的点划分成对立的两部分

# 缩点

- 缩点过程
  - 任取一个节点，从该点开始在全部+边上做BFS，将能到达的节点划为一组
  - 任取一个尚未归组的点重复上述过程，直到所有点都归组





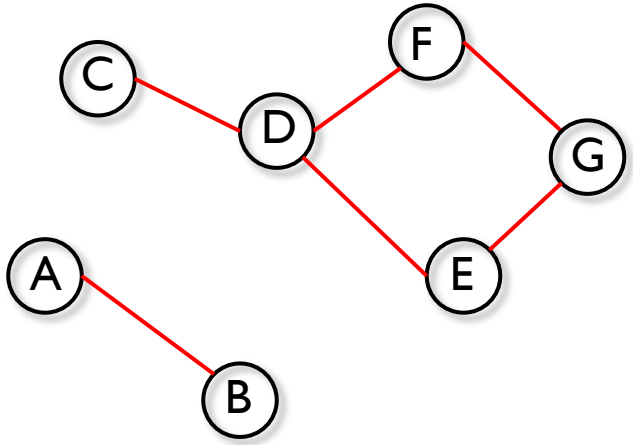


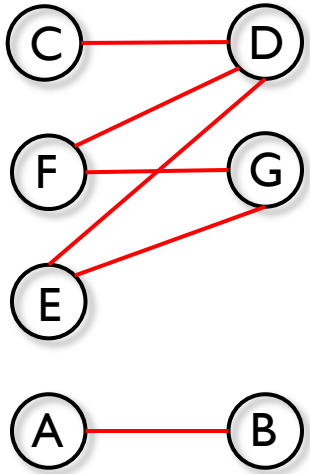
- 这个归组平衡吗?
  - 需要验证：组内是不是无-边?
  - 如果有，那么能找到一个含有1条-边的回路，矛盾

# 缩点后的划分

称一个标注图是平衡的，如果可以将其分成两个对立的集合（-只在集合间）

- -边此时只出现在组间
  - 注意：-边构成的图未必是路径，可以是任意图
  - 我们要的划分等价于将此图划为二分图



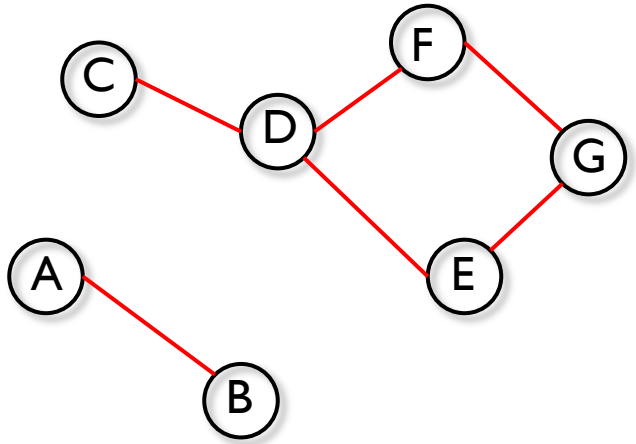



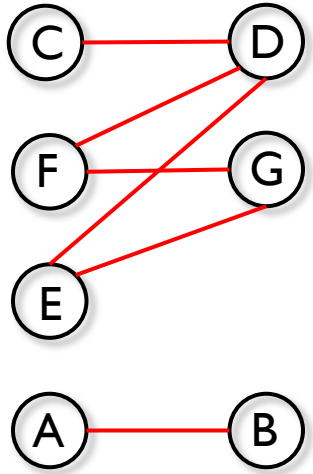
# 如何判断一个图是否是二分图?

- 定理：一个图是二分图，当且仅当没有奇数长度的回路
  - 我们仅需要“无奇数长度回路 -> 二分图”的方向　（另一个方向？）
  - 因此不存在奇数-边回路 -> 平衡

# 证明梗概

- 首先，对每个连通分量可以独立处理

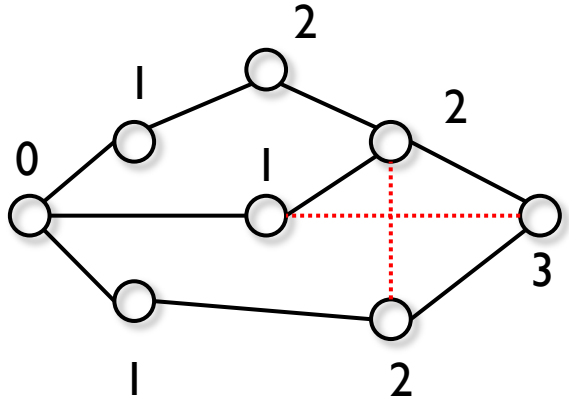






- 假定图连通
  - 任取一点u，从u做BFS
  - 奇数层一边，偶数层一边
  - 如果无奇数长度圈，则层数同奇偶的点没有边



# 另一个方向的推广：近似平衡性

- 依然考虑完全标注图，但放松对平衡性的要求
- 比如：若一个完全标注图中至少99.9%的三角形是平衡的，那么
  - 存在一个至少包含90%节点的集合，这些节点组对中至少90%互为朋友，或
  - 可将节点划为两个集合X和Y，满足
    - 集合X、Y内部至少有90%节点对互为朋友
    - 横跨X、Y的节点对至少有90%互为敌人

# 近似平衡定理

- 一般地，若一个完全标注图中至少$1-t^3$比例的三角形是平衡的，则
  - 存在一个至少包含(1-t)n节点的集合，这些节点对中至少1-t比例互为朋友，或
  - 可将节点划为两个集合X和Y，满足
    - 集合X、Y内部至少有1-t比例节点对互为朋友
    - 横跨X、Y的节点对至少有1-t比例互为敌人

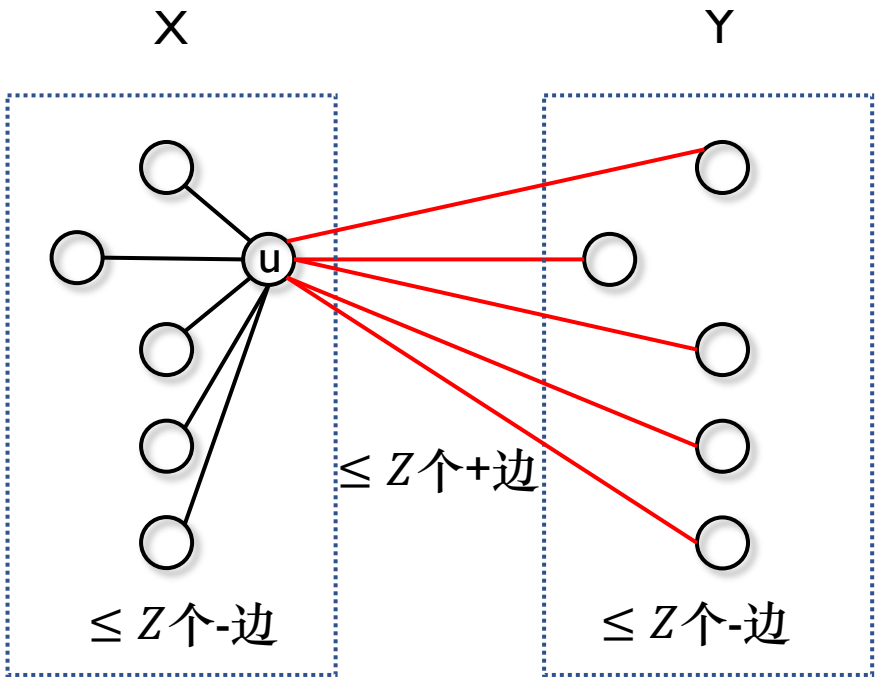
三角形是平衡的：三条边要么+++要么-+-

# 思路：找到一个“关键点”

- 根据假设，有至多$t^3$比例的非平衡三角形
- 对于每个点u，设u的权重w(u)是以它为某个顶点的非平衡三角形数目
- 全图总权重至多：$t^3 \cdot C_n^3$；平均每个点权重至多：$t^3 \cdot \frac{n^2}{2}$
- Averaging argument：必然存在一个点u权重w(u)至多：$t^3 \cdot \frac{n^2}{2}$
- 选择该u作为关键点

# 思路：剖分成X和Y

- 设$Z := t^3 \cdot \frac{n^2}{2}$
- 将与u连接的+边端点放入X，其余放入Y
- 观察：
  - 在X中，每条-边对应一个含u的非平衡三角形
  - 因u只属于$Z$个非平衡三角形，故X有$\leq Z$个-边
  - 类似地，在Y中也有$\leq Z$个-边
  - 同样，考虑跨越X、Y的边，也只有$\leq Z$个+边



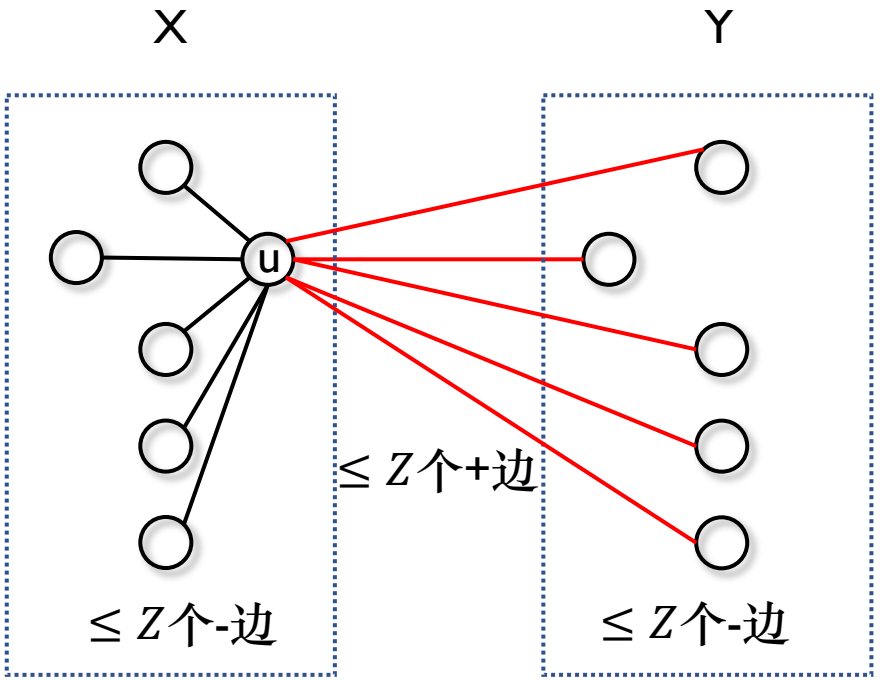

三角形是平衡的：三条边要么+++要么-+-

# 思路：按照|X|和|Y|分情况讨论

- $Z := t^3 \cdot \frac{n^2}{2}$
- 设|X| = x， |Y| = y
- 情况一：x > (1 – t) n，或者y > (1 – t) n

求证：存在一个至少包含(1-t)n节点的集合，这些节点组对中至少1-t比例互为朋友

  - 验证：$C_x^2 - Z \geq (1-t) \cdot C_x^2$
  - 等价于： $t \cdot C_x^2 \geq Z$
  - 亦即$t \cdot \frac{x^2}{4} \geq \frac{t^3 n^2}{2}$
  - 只需证$t \cdot \frac{(1-t)^2 n^2}{4} \geq \frac{t^3 n^2}{2}$

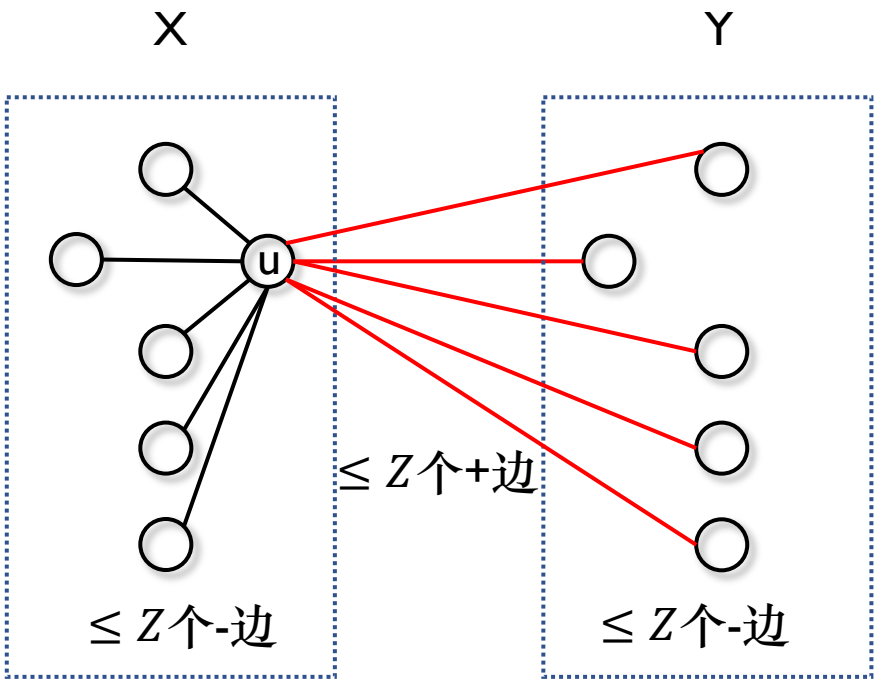


# 情况二

- $Z := t^3 \cdot \frac{n^2}{2}$
- 设|X| = x， |Y| = y
- 情况二：x和y都至多是(l – t) n

求证：可将节点划为两个集合X和Y，满足
集合X、Y内部至少有l-t比例节点对互为朋友
横跨X、Y的节点对至少有l-t比例互为敌人

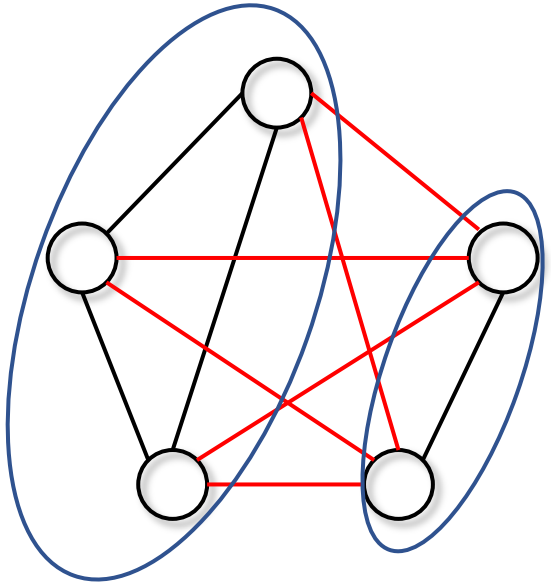
  - 说明x和y都> t n
  - 验证：
    - 集合X内部：$C_x^2 - Z \geq (1-t) \cdot C_x^2$
    - 跨越XY：$xy - Z \geq (1-t)xy$



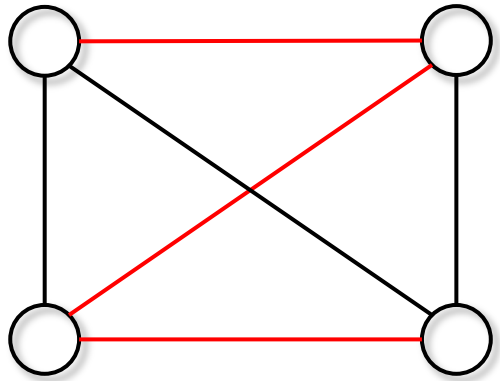
# 计算问题：Correlation Clustering

# Correlation Clustering

- 输入完全标注图$G(V, E^+, E^-)$
- 对于一个V的划分$\{S_i\}$，称所有在$S_i$里的-边和跨越$S_i$的+边为分歧边
- 目标是将整个图划分为若干个子集，最小化分歧边的个数



另外一个相关问题：最大化一致边数



- 可以理解为没有平衡关系的图的最小分歧划分

# Correlation Clustering的其他应用

- 与k-means等比较：不需要选k，不需要嵌入$\mathbb{R}^d$

- +/-可以理解成某个分类器f给出的二元相似度判定
- Correlation clustering即根据分类器给出的信息尽可能的恢复类别信息

# Document de-duplication

amazon.com

**Thanks for your order, Martin Flimberton!**

**Want to manage your order online?**
If you need to check the status of your order or make changes, please visit our home page at Amazon.com and click on

**Purchasing Information:**

**E-mail Address:** lostiddude@yahoo.com

**Billing Address:**
Martin Flimberton
1434 Main Street Road
Glenbert Iows, Illinois 60121
United States

**Order Grand Total: $53.99**

Get the Amazon.com Rewards Visa Card and get **$30 instantly** as a

**Order Summary:**

**Shipping Details : 8thdayconsulting**

| | |
|---|---|
| **Order #:** | 104-3041649-8513858 |
| **Shipping Method:** | Standard Shipping |
| Items: | $50.00 |
| Shipping & Handling: | $3.99 |
| | ------ |
| Total Before Tax: | $53.99 |
| Estimated Tax To Be Collected:* | $0.00 |
| | ------ |
| **Order Total:** | **$53.99** |

**Delivery estimate:** Oct. 24, 2012 - Nov. 8, 2012

1 **"Microsoft Office 2010: Essential (Shelly Cashman Series)"**
Shelly, Gary B.; Paperback; $50.00
In Stock
Sold by: 8thdayconsulting

Preston
amazon.com

**Thanks for your order, Preston Presterton!**

**Want to manage your order online?**
If you need to check the status of your order or make changes, please visit our home page at Amazon.com and click on Y

**Purchasing Information:**

**E-mail Address:** mepartydj@yahoo.com

**Billing Address:**
Preston Presterton
259 Greenpoint DR
DALLAS, TX 75231-9126
United States

**Order Grand Total: $97.41**

Get the Amazon.com Rewards Visa Card and get **$30 instantly** as an Amazon.com Gift Card.

**Order Summary:**

**Shipping Details : buybackselyria**

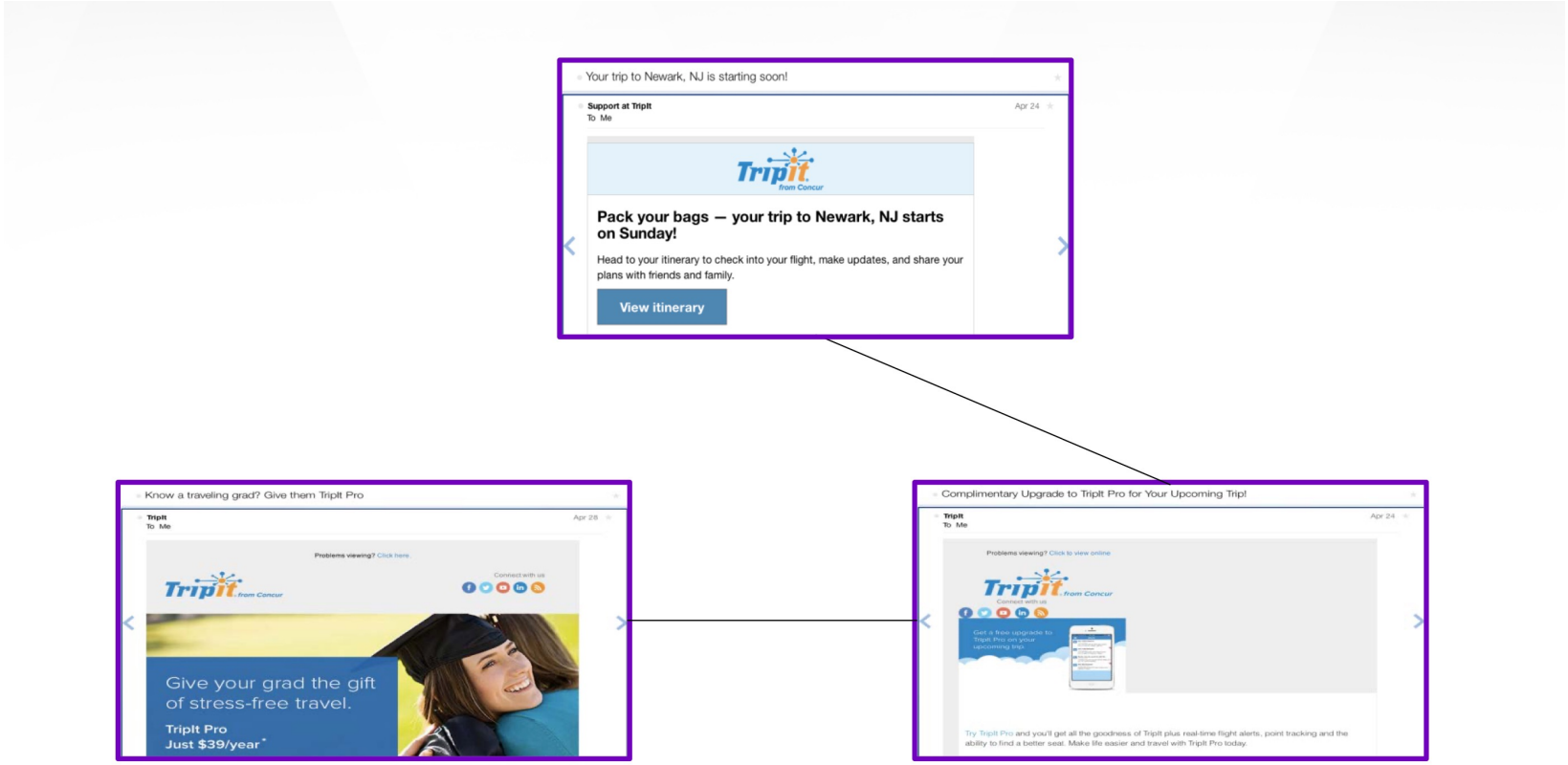
| | |
|---|---|
| **Order #:** | 002-1903988-3076225 |
| **Shipping Method:** | Standard Shipping |
| Items: | $30.68 |
| Shipping & Handling: | $2.98 |
| | ------ |
| Total Before Tax: | $33.66 |
| Estimated Tax To Be Collected:* | $0.00 |
| | ------ |
| **Order Total:** | **$33.66** |

**Delivery estimate:** Oct. 16, 2012 - Oct. 31, 2012

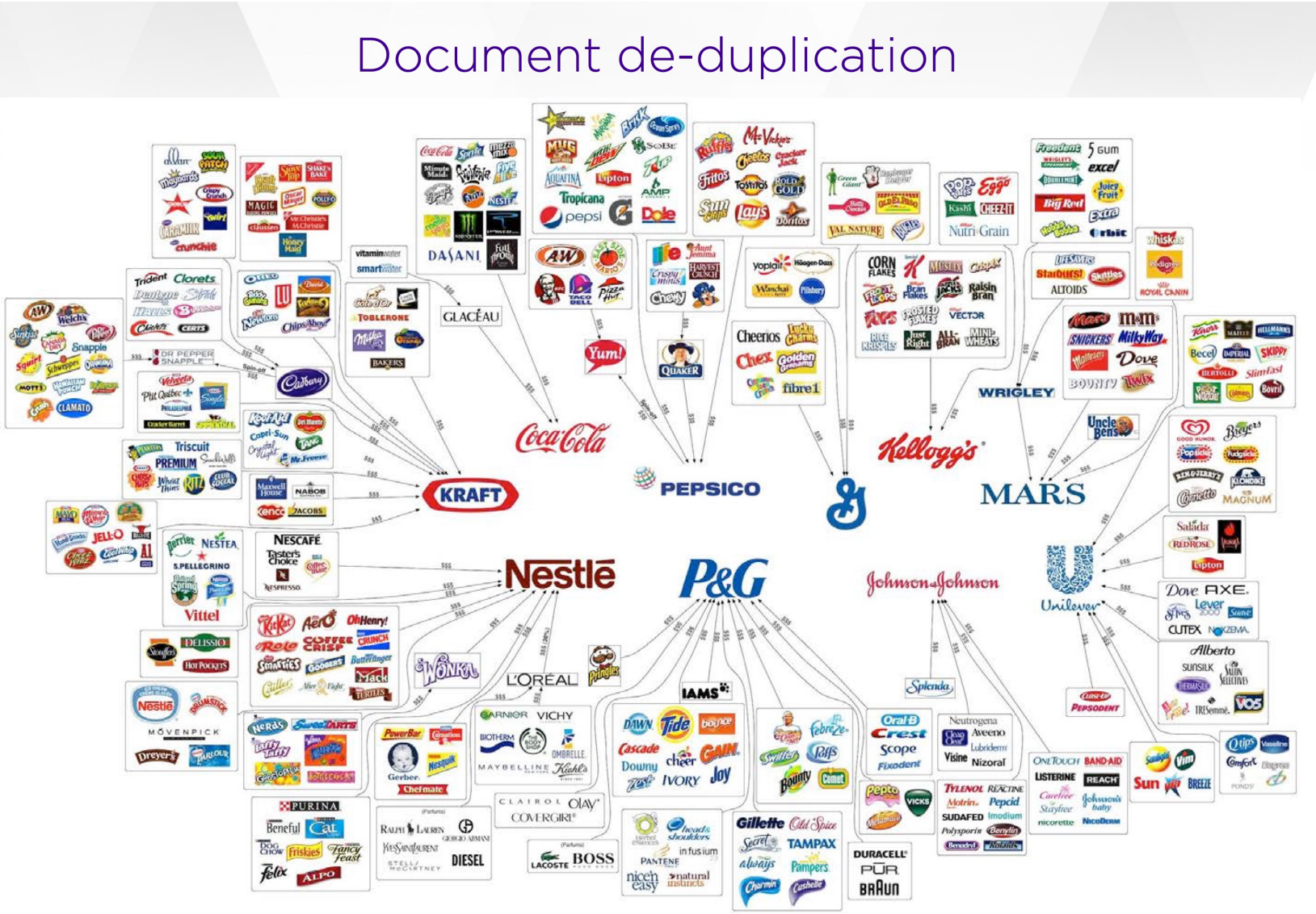
1 **"Fawlty Towers: The Complete Collection Remastered"**
Cleese, John; DVD; $30.68
In Stock
Sold by: buybackselyria

They are not identical

- 判断“相同”是很难的；谁和谁应该放一块?



# Document de-duplication



- Correlation clustering仍是研究热点
  - 大数据算法
  - 其他变种
  - Etc etc

# 一个3-近似的算法

- Ailon-Charikar-Newman, JACM 08

KwikCluster$(G = (V, E^+, E^-))$

If $V = \emptyset$ then return $\emptyset$
Pick random pivot $i \in V$.
Set $C = \{i\}, V' = \emptyset$.

For all $j \in V, j \neq i$:
    If $(i,j) \in E^+$ then
        Add $j$ to $C$
    Else (If $(i,j) \in E^-$)
        Add $j$ to $V'$

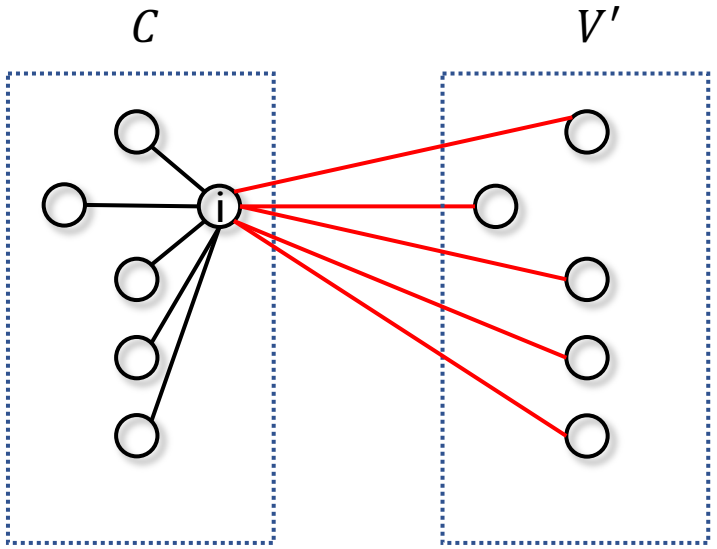
Let $G'$ be the subgraph induced by $V'$.

Return $C \cup$ KwikCluster$(G')$ .

在该算法中，边$(j,k)$ 何时会成为分歧边?

- 观察：pivot直接连接的都不是分歧边

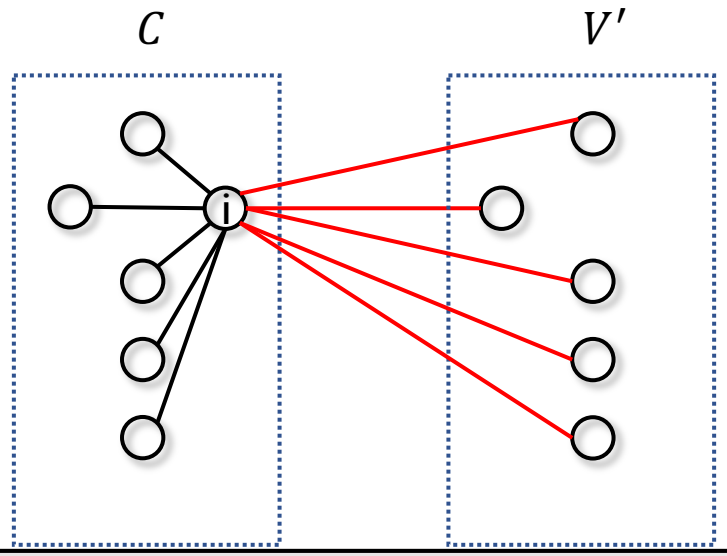
结论：在算法中，$(j,k)$成为分歧边，仅当存在i，使得i, j, k同在一个递归调用中，i是pivot，且(i, j, k)是++-型三角形




如果图是关系平衡的，则该算法一定找到满足关系平衡定理的划分!

# ALG的上界

- 称++-型三角形(i, j, k)为“坏”三角形，并设T为坏三角形的集合
- 对坏三角形$t \in T$，设事件$A_t$为t中三点有一点（曾）被选为pivot
- Claim：$\mathbb{E}[\mathrm{ALG}] \leq \sum_{t \in T} \Pr[A_t]$
- 证明要点：
  - 根据上页讨论，只有属于某个坏三角形的边才可能成为分歧边（试解释---？）
  - 在算法执行中，任何坏三角形$t \in T$中，仅当t中有pivot点才会有分歧边，并且分歧边个数至多是1



# OPT的下界

- 通常很难（上界容易，因为任何可行解都是上界）
- 一种通用方法：利用线性规划

LP:
min. $c^T x$
s.t. $\mathrm{A}x \geq b, x \geq 0$

Dual-LP (DLP)
max. $b^T y$
s.t. $\mathrm{A}^{\mathrm{T}} y \leq c, y \geq 0$

- 定理（strong duality of LP）：设x，y分别是上述LP和DLP可行解。则 $\mathrm{LP}(x) \geq \mathrm{DLP}(y)$，且等号取在最优解$x^*, y^*$上
- 思路：我们给出一个LP和一个DLP的解y，使得

$$\mathrm{OPT} \geq \mathrm{LP}(x^*) \geq \mathrm{DLP}(y) = \frac{1}{3}\sum_{t \in T} \Pr[A_t]$$

再结合$\mathbb{E}[\mathrm{ALG}] \leq \sum_{t \in T} \Pr[A_t]$可以得到3-近似

# 写出LP松弛

- 对每个边/点对e $= (j, k)$ 设置一个变量$x_e \in [0, 1]$
  - 当$x_e = 1$ 时对应分歧边，目标函数则是minimize分歧边数

LP:

min. $\sum_e x_e$

s.t. $x_{e_1} + x_{e_2} + x_{e_3} \geq 1 \quad \forall t = (e_1, e_2, e_3) \in T$

- 观察：OPT $\geq$ LP$^*$，因为LP是原问题的松弛（relaxation）
  - 松弛：任何原问题解都对应一个LP解，并且该解目标函数不大于原问题的
  - $x_{e_1} + x_{e_2} + x_{e_3} \geq 1$是因为++-三边不可能同时满足

# 推导DLP

- 根据公式写出DLP

LP:
min. $c^T x$
s.t. $\mathrm{A}x \geq b, x \geq 0$

Dual-LP (DLP)
max. $b^T y$
s.t. $\mathrm{A}^{\mathrm{T}} y \leq c, y \geq 0$

LP:
min. $\sum_e x_e$
s.t. $x_{e_1} + x_{e_2} + x_{e_3} \geq 1 \quad \forall t = (e_1, e_2, e_3) \in T$

DLP:
max. $\sum_{t \in T} y_t$
s.t. $\sum_{t \in T: e \in t} y_t \leq 1 \quad \forall e$

- $y_t$可以理解成对于每个坏三角形的一个权重/概率

# 构造DLP的解y

DLP:

max. $\sum_{t\in T} y_t$

s.t. $\sum_{t\in T:e\in t} y_t \le 1 \qquad \forall e$

- 回忆：对坏三角形$t \in T$，事件$A_t$为t中三点有一点（曾）被选为pivot
- 定义事件$B_e$为：边e被算法选为分歧边，则

$$\Pr[B_e \wedge A_t] = \Pr[B_e \mid A_t]\Pr[A_t] = \frac{1}{3}\Pr[A_t]$$

- 又观察到若$t, t' \in T$ 有e作为公共边，则$A_t, A_{t'}$ 是互斥事件
- 因此$\forall e$，有 $1 \ge \sum_{t\in T:e\in t} \Pr[B_e \wedge A_t] = \sum_{t\in T:e\in t} \frac{1}{3}\Pr[A_t]$
- 因此设$y_t := \Pr[A_t]$，就可以得到一个DLP的可行解！
- 并且 $\mathrm{OPT} \ge \mathrm{DLP}(y) = \sum_{t\in T} \frac{1}{3}\Pr[A_t]$

再结合$\mathbb{E}[\mathrm{ALG}] \le \sum_{t\in T} \Pr[A_t]$可以得到3-近似

# 线性规划方法

- 线性规划方法是设计近似算法的核心方法
  - 先解LP，后rounding
  - Primal-dual
  - LP-hierarchy
  - …
- 一般地，除线性规划外，还可以考虑凸优化
  - SDP + rounding得到max-cut的“最优”近似比
- 《The design of approximation algorithms》 Williamson and Shmoys.