社会计算

第四节

节点权力与节点地位

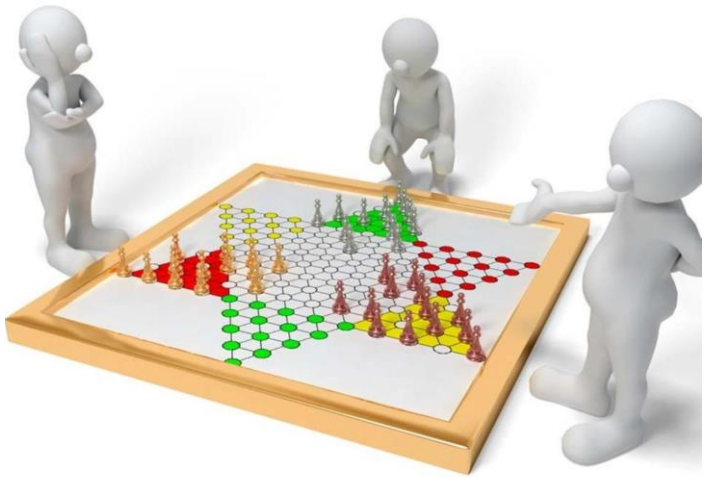
徐童

2024.3.18

上节回顾

博弈中的行为逻辑总结

- 基于上述概念总结所得的简单博弈行为推理原则
  - 如果两个参与人都有严格占优策略，则可以预计他们均会采取严格占优策略；
  - 如果只有一个参与人有严格占优策略，则这个参与人会采取严格占优策略，而另一方会采取此策略的最佳应对（一定会有！）
  - 如果双方都不存在严格，则寻找纳什均衡
    - 如果只有一个纳什均衡，则该均衡对应决策结果
    - 如果存在多个纳什均衡，则需要额外信息辅助推断
      - 例如协调博弈（性别战），鹰-鸽博弈等
      - 再次强调：纳什均衡有助于缩小范围，但不能保证有结果



本节主旨

- **博弈中参与者的地位问题**

- 上节课所介绍的大多数博弈中，参与者的地位都是对等的，但也有例外
  - 对等博弈往往对应着对称的收益矩阵
  - 部分博弈中，一方依赖于另一方获得收益
    - 依赖性带来地位和权力的不对等！

不对等↓

		Suspect 2	
		NC	C
Suspect 1	NC	$-1,-1$	$-10,0$
	C	$0,-10$	$-4,-4$

Figure 6.2: Prisoner's Dilemma

		Firm 2		
		A	B	C
Firm 1	A	$4,4$	$0,2$	$0,2$
	B	$0,0$	$1,1$	$0,2$
	C	$0,0$	$0,2$	$1,1$

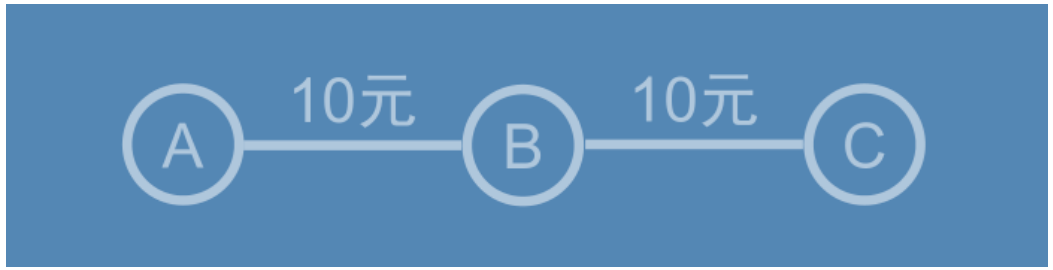
Figure 6.6: Three-Client Game

本节主旨

- **上古回顾：弱关系带来“中介”的议价能力**

- 弱关系形成的过程往往依赖于中介的“引荐”
  - 相应的，这种“掮客”行为也成为了中介的“社会资本”
    - 例如，网络交换实验中，中介可能获得更大收益

往往以平分收场



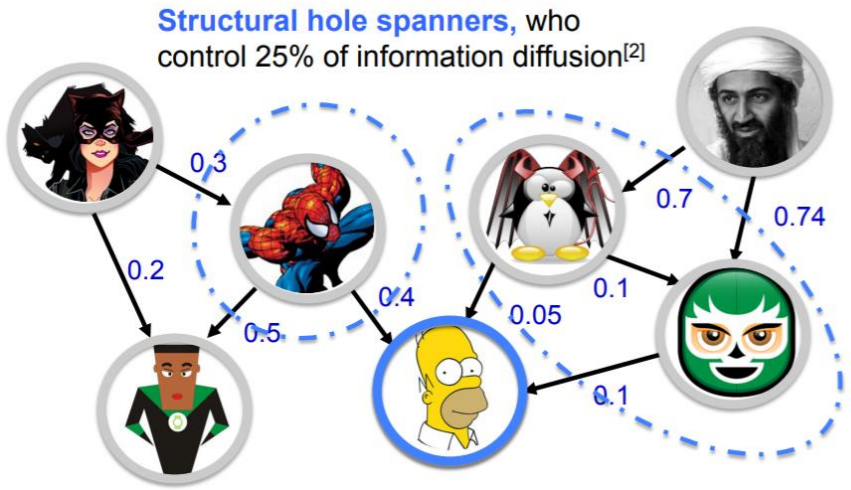
如果仅可与邻居交涉，则B收益更大

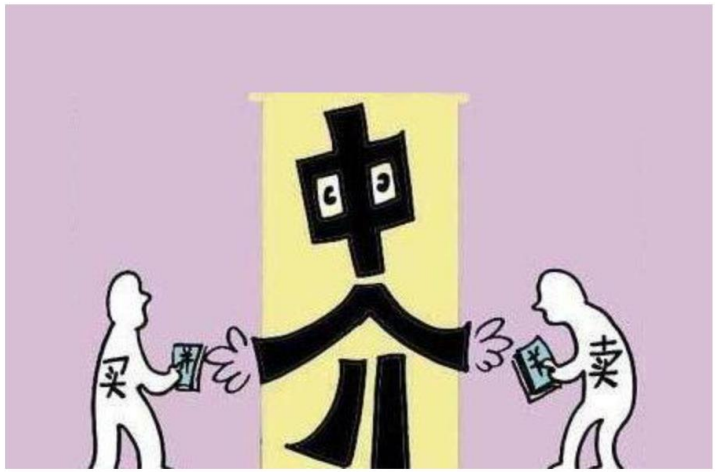
本节目录

- **网络中的议价与权力**
  - 网络中的权力交换
  - 纳什议价模型
  - 网络交换的稳定性问题
- 网络节点的地位
  - 基于中心性的启发式衡量
  - PageRank及其衍生模型
  - HITS算法：地位与类别的双重区分

- **网络中的节点存在着不同**

- 我们曾经说过：网络中节点 / 关系的分工不同
  - 不同的分工带来不同的职责，也带来不同的权力
  - 节点在网络中的支配性地位，将赋予该节点更多的社会性机会 / 利益





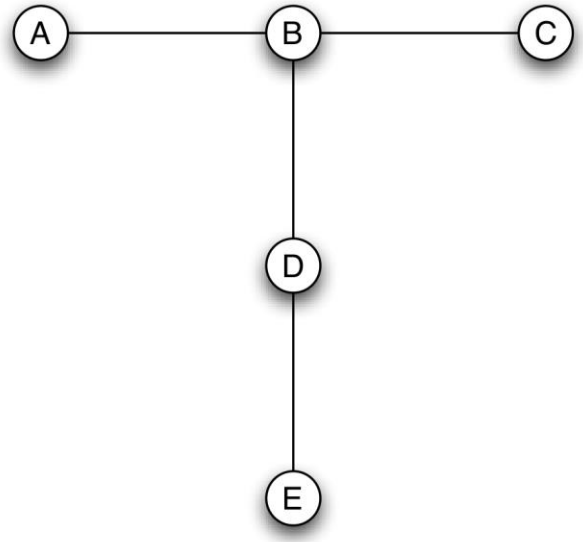
- **课堂小实验：网络交换**

- 征集 5 名志愿者同学，按照如下流程进行实验（约10分钟）

① 每位同学扮演右图中A-E中的一个节点（抽签决定）

② 游戏将在 5 轮内结束，在每一轮中，每位同学可以通过纸条与自己的网络邻居进行交流，并遵循以下规则：

- 纸条内容为“自己的编号：对方的编号，自己的占比：对方的占比”，例如“A：B，3：7”
- 每一轮可以同时向多个邻居发出纸条，也可以只发出一张空白纸条
- 如果同意对方的分配条件，可以在纸条上注明OK并退还纸条
- 纸条由工作人员统一收发，如果某节点没有收到纸条，则工作人员将补发一张空白纸条



- **课堂小实验：网络交换**

- 征集 5 名志愿者同学，按照如下流程进行实验

① 每位同学扮演右图中A-E中的一个节点（抽签决定）

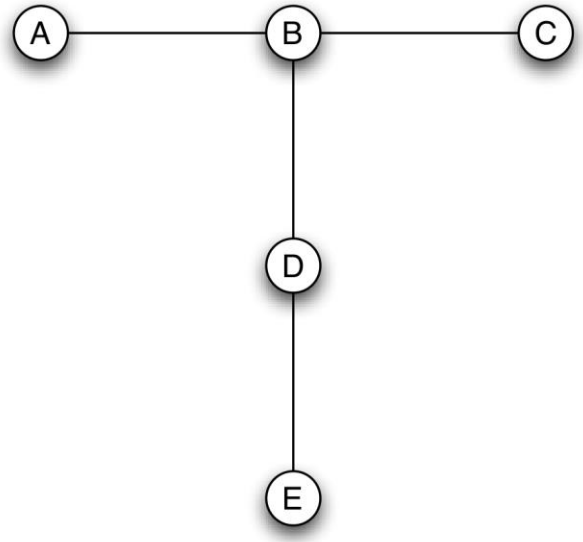
② 游戏将在 5 轮内结束，在每一轮中，每位同学可以通过纸条与自己的网络邻居进行交流

③ 如果收到了对方表示确认（OK）的纸条，同时自己仍然认可这一协议，请举手向工作人员示意协议达成，否则协议作废

④ 5 轮结束后，达成协议的按照协议进行分配，没达成协议的收益为 **0**

- 每个节点只能和**最多一个邻居**达成协议（排他性）
- **目标是尽可能最大化自己的收益！**

⑤ 游戏结束后，请 5 名志愿者同学简要介绍自己的交流过程及动机



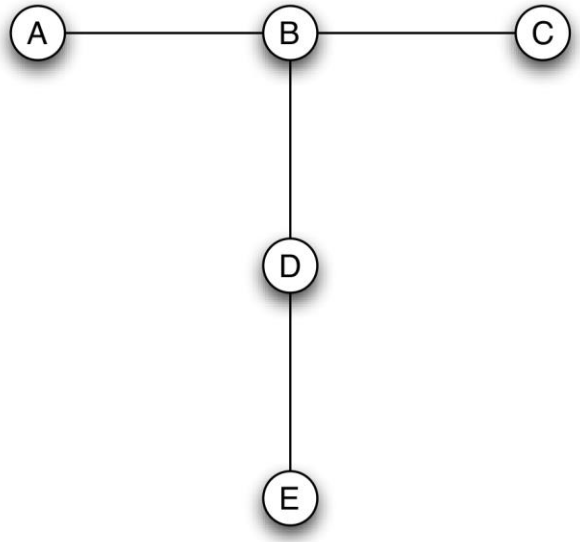
- **网络中的议价与交换**

- 这个实验教会了我们什么？审视关系上的价值

  - 当面临利益冲突时，谁占据着优势的地位？
  - 权力的均衡性从何而来？
    - 权力来自于稳定的关系，即双方认可，而不是单方面强加的
    - “均衡”≠“相等”，地位的不平等是普遍存在的

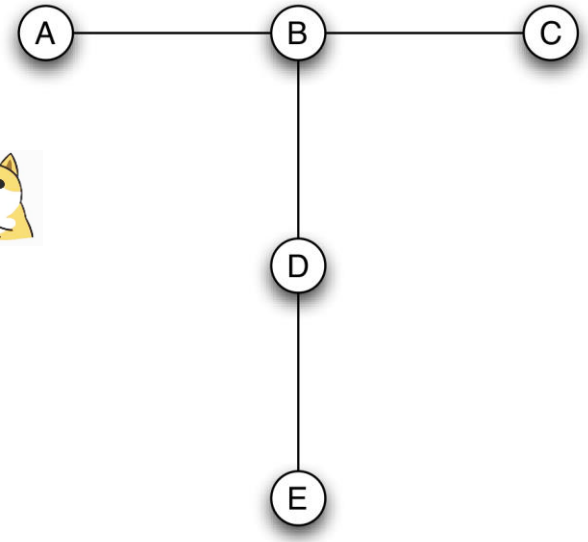
- **网络中的议价与交换**

- 一些影响权力的基本论述：（1）依赖性
  - 社会关系带来价值，但价值的来源不同，部分节点的价值来源于其他节点
    - 例如，社团的介绍人
      - 回顾第二讲中“星形网络”的形成过程
    - 在右图中，节点A、C的资源来源为B；相应的，B的地位/价值来自于A、C



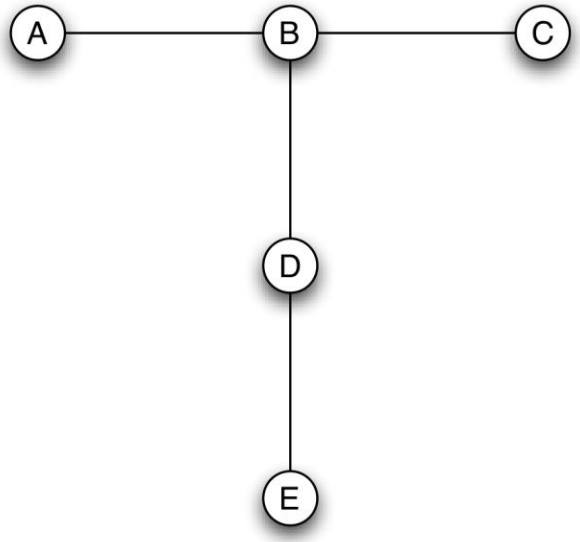
• 网络中的议价与交换

- 一些影响权力的基本论述：（2）排他性
  - 由于网络结构的不同，部分节点具有一项“特权”，即有能力排除其他节点
    - 某种意义上，这种排他性意味着“选择的自由”（备胎自由） 
      - 例如，在右图中，当节点B面临选择“最要好的朋友”时，他/她可以从A、C中单方面挑选一个
      - 但对于A、C来说，他们除了节点B别无选择
    - 面临“被-选择”而不具备“选选择的自由”可能带来不安全感
      - 例如，子女与父母的关系与“二胎问题”



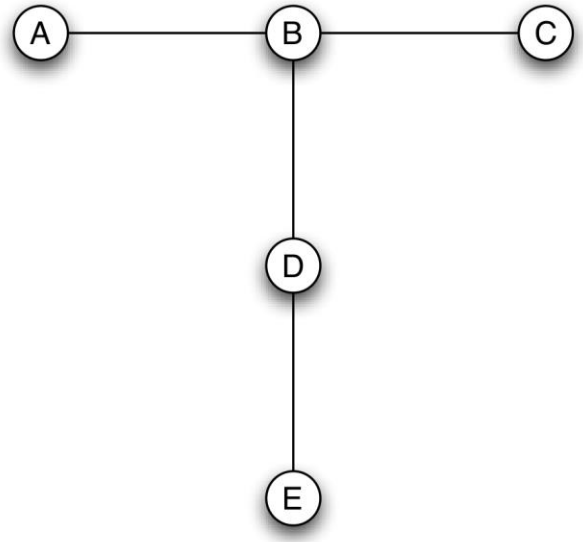
- **网络中的议价与交换**

- 一些影响权力的基本论述：（3）饱和性
  - 基本概念：某种可能带来回报的事物，随着其数量的增加，回报将逐渐减少
    - 前面两种论述都指向一个结论：B通过网络地位获得了价值
    - 而一旦这种价值达到某种饱和之后，B维持这些社会关系的兴趣会降低
      - 注意：不是想放弃关系，而是不满足于对等关系！
      - 只倾向于维护那些令他/她收益更多的关系



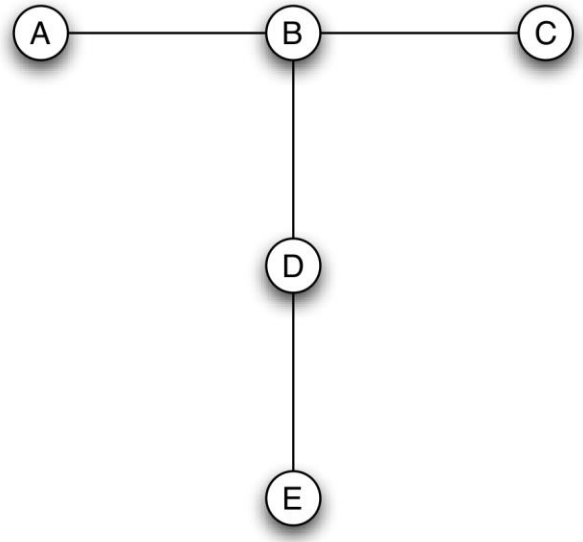
- **网络中的议价与交换**

- 一些影响权力的基本论述：（4）中心性
  - 直观来看，节点在网络中的中心性可以作为量化节点地位的线索
    - 例如，高介数的节点往往具有更高的权力
      - 有关介数将在本节课后半部分详细介绍
    - 在关心信息流动的场合中，介数往往是不错的提示指标
      - 但是，简单考虑中心性可能会被误导
        - 例如，后面所要提到的5-节点路径



网络中的议价与交换

- 回到我们先前的实验中来，我们简单总结一下实验结果
  - 从实验结果中，我们大体上可以看到以下规律：
    - B获得了最高的收益
    - D大体上可以获得尚可的收益，但远不如B
    - A或C中的一方将毫无收益
    - E很可能将获得收益
  - 造成这一现象的原因是什么？



- **网络中的议价与交换**

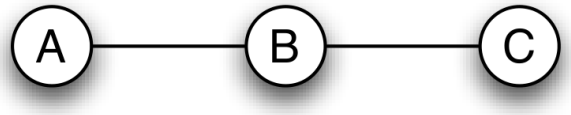
- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第一种情况：2-节点路径

(a) *2-Node Path*

  - 在基本的2-节点路径中，理论推导与实验验证的结果大体倾向于两者平分
    - 前提是：两者的利益相当，忽略其他复杂的影响因素

- **网络中的议价与交换**

- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第二种情况：3-节点路径

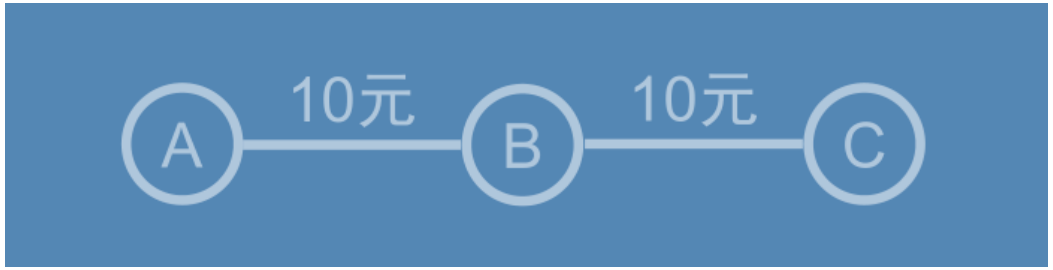


(b) *3-Node Path*

  - 在3-节点路径中，基于依赖性和排他性，可知具有“选择自由”的B更具有权力
    - 实验显示B的收益大致为5/6（不平等条约！）
      - 其中经历了一个“迫使”A、C多次降低要求的过程

- **网络中的议价与交换**

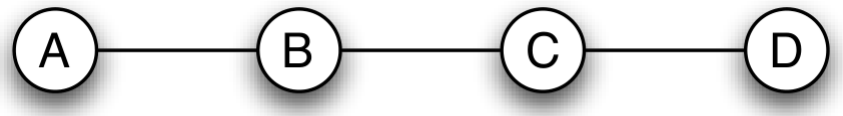
- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第二种情况：3-节点路径



  - 一个有趣的现象：如果忽略排他性，允许B同时与多个邻居达成协议
  - 此时，B会倾向于和A、C都对半平分（类似于我们“上古时期”所举过的例子）
    - 潜在原因：B对A、C的需求并不亚于A、C对B的需求（你们都是我的翅膀）

- **网络中的议价与交换**

- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第三种情况：4-节点路径

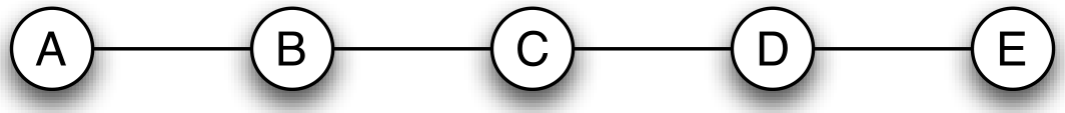


(c) *4-Node Path*

  - 此时，B虽然仍有优势，但相比于3-节点路径并不明显
    - 对于B来说，C虽然是备选，但这个备选存在风险：C同样也有备选（互为备胎） 
    - 因此，B往往会做出一定的让步（约在7/12至2/3之间）：弱权力!

- **网络中的议价与交换**

- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第四种情况：5-节点路径

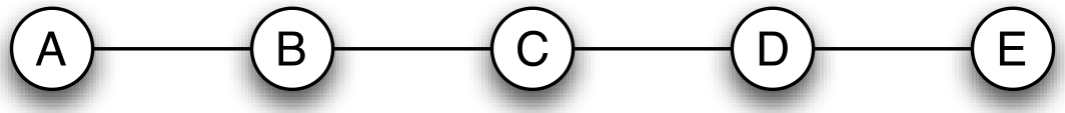


(d) *5-Node Path*

  - 最为尴尬的节点：C节点
    - 理想很丰满：左右逢源，多重备胎
    - 现实很骨感：B、D都有备胎，C往往只能“被”备胎，其实往往处于弱势地位

- **网络中的议价与交换**

- 我们先从一些更小的单元开始，讨论这一规律的形成原因
  - 第四种情况：5-节点路径

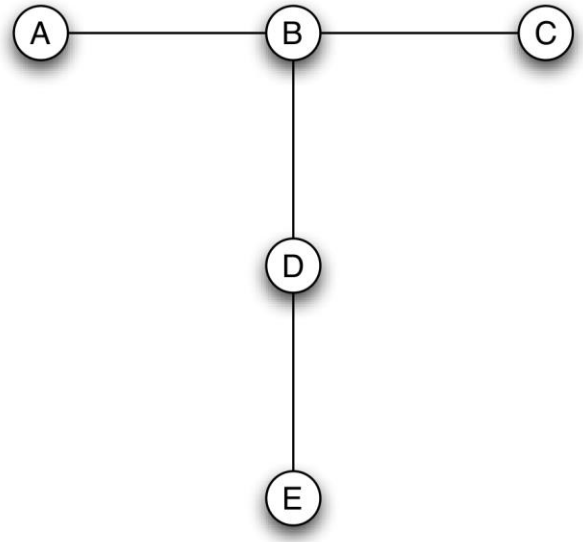


(d) *5-Node Path*

  - C节点拯救计划：采用与3-节点路径类似的思路，去除排他性（只去除B、D的排他性）
    - 此时，为争取更多交换机会，B、D将争取C的加盟
    - 相应的，C就获得了排他性与“选择的自由”

• 网络中的议价与交换

- 现在，我们来复盘先前实验的大致经过
  - 首先，由于B有能力排除A和C，因此B在于A、C的交涉中占优
  - 与之相应的是，D虽然名义上具有B作为备选，但实际上很难从与B的交涉中获取较大利益，而实际上处在和A、C类似的地位
    - 更类似于5-节点中的C，因为D很难情愿给出高于A、C的报价
  - 因此，D与E的交涉大体上处于均势
    - D可能稍微占优（如果成功实现对E的恐吓）



- **网络中的议价与交换**

- 假如，我们对网络结构进行一些调整
  - 某种意义上，相当于前一张图中处于弱势的节点实现了私下串联，共同反抗强权 B
  - 此时，B的地位大致相当于4-节点路径中的B，但更具优势（对方没有强有力的备选）
  - 最后的结果很可能是A-B成对，C-D成对，B所获得的收益相较于4-节点路径会略高

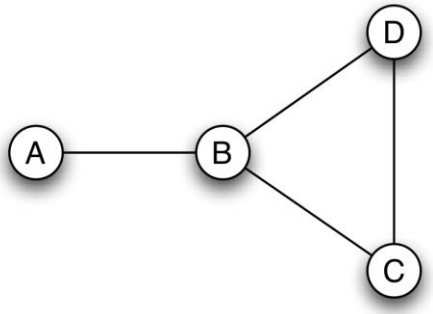


Figure 12.3: An exchange network with a weak power advantage for node B.

网络中的议价与交换

- 最后，我们讨论一种不稳定的交换情况
  - 在考虑排他性的情况下，任何一种可能达成的交换一定会被第三方破坏
    - 破坏的方式也很简单：向弱势一方让渡一点利益即可
    - 要命的是：这个过程可以无限循环，永远无法达成共识
  - 吐槽：谁说三角形具有稳定性？三角恋明明是最致命的

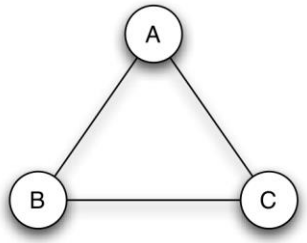


Figure 12.4: An exchange network in which negotiations never stabilize.

• 网络中的议价与交换

- 最后，我们讨论一种不稳定的交换情况
  - 但是需要注意的是：三角形的破坏性往往限于“独立性”明显的局部网络
  - 在大型网络中，三角形可能不会带来什么问题
    - 例如，先前提到的“柄状”网络中的三角形

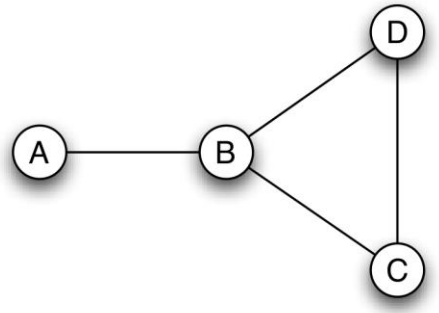
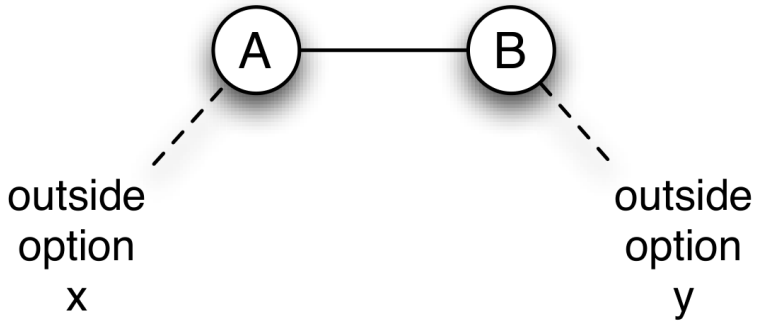


Figure 12.3: An exchange network with a weak power advantage for node B.

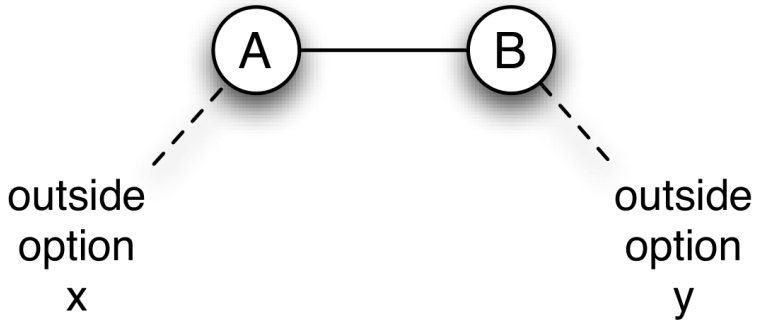
• 纳什议价模型

- 现在，我们了解了一些基本的交换实验。如何从博弈角度看待这些实验？
  - 回顾基本的二人博弈问题，我们约定：参与人所关心的事务只包含在收益矩阵中
  - 相比于博弈问题，在网络交换实验中，节点考虑的内容多出了什么？
    - 除2-节点问题之外的情况，一个节点往往面临多个报价（外部信息存在）
    - 显然，节点不会接受比当前收益更低的报价



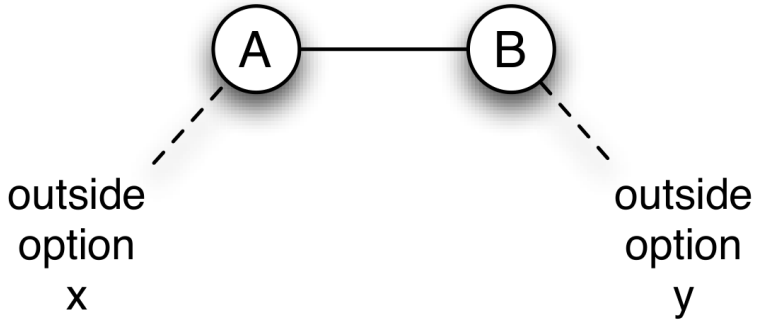
• 纳什议价模型

- 现在，我们了解了一些基本的交换实验。如何从博弈角度看待这些实验？
  - 我们通过数学形式，定义交换实验的纳什议价解问题
  - 假设博弈双方A、B，分别有外部选项 x 和 y，其中有 $x + y \leq 1$ （为什么？）
    - 如果出现 $x+y > 1$，则A、B不可能达成协议（因为至少有一方利益受损）
    - 由此，谈判的目的实际上在于分配剩余的 $s = 1-x-y$



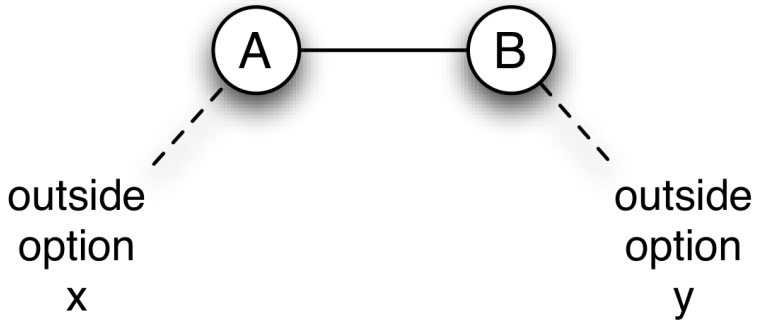
• 纳什议价模型

- 现在，我们了解了一些基本的交换实验。如何从博弈角度看待这些实验？
  - 我们通过数学形式，定义交换实验的纳什议价解问题
  - 从纳什的角度考虑，由于两者有着对等的谈判权力，那么他们将倾向于均分
    - 因此，A将得到 $x + s/2$，而 B 将得到 $y + s/2$
    - 这一结果也与人们的直觉相一致，因此可以预计实验结果也是类似的



• 纳什议价模型

- 现在，我们了解了一些基本的交换实验。如何从博弈角度看待这些实验？
  - 有趣的现象：纳什议价解的理论与实验结果不一致！
  - 外部选项收益更高的参与方，往往获得的收益增量要高于 s/2
    - “高状态”时倾向于夸大自己的外部选项，而“低状态”时倾向于贬低自己
    - 虚张声势的必要性！



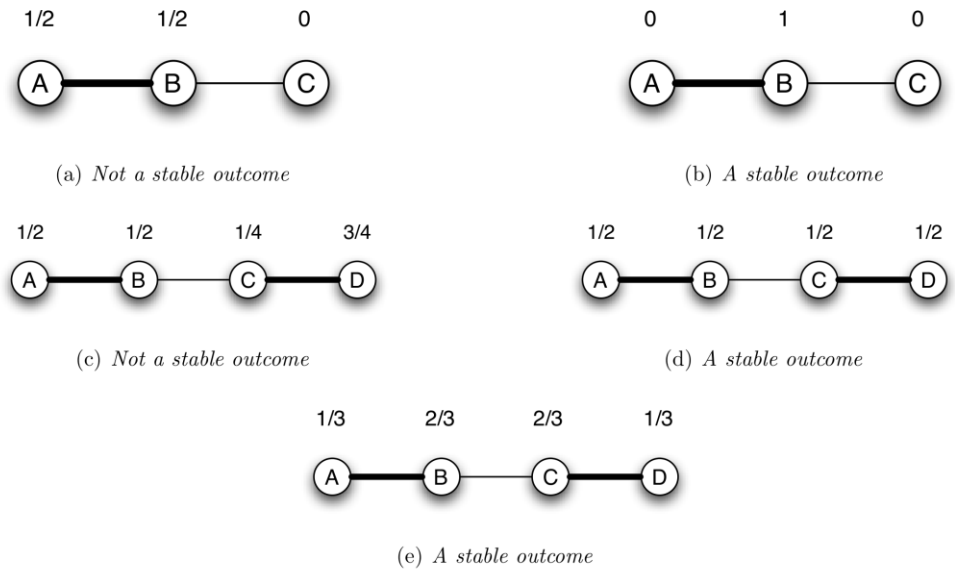
- **小拓展：最后通牒博弈**

- 一类有趣的博弈互动：终极博弈（最后通牒）
  - 参与双方仅进行一次互动（一锤子买卖）
  - A提出一个分配的方案
  - B选择接受或者不接受
    - 如果接受，则A、B按照方案进行分配
    - 如果不接受，则双方什么都无法得到
  - 猜猜看，双方的议价将会是怎样的结果？
    - 将作为实验一的一部分进行讨论

- **结果的稳定性**

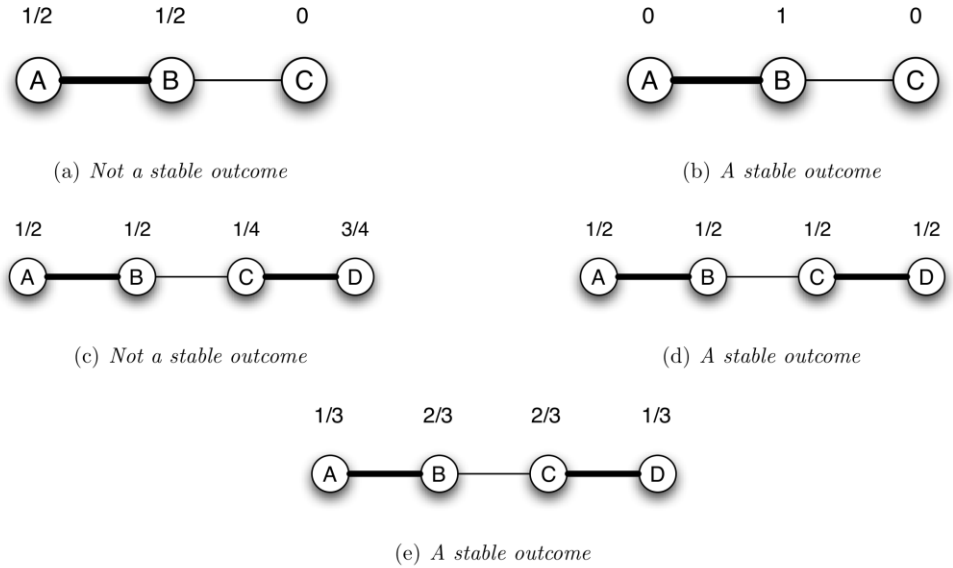
- 谈判，总要有个“结果”
  - 什么是结果：给定一个图，“结果”是其中一种量化的分配方案
    - 其中，每个节点有一个[0,1]区间内的赋值，非0代表参与了一次分配

右图这些结果，都能稳定存在么？→



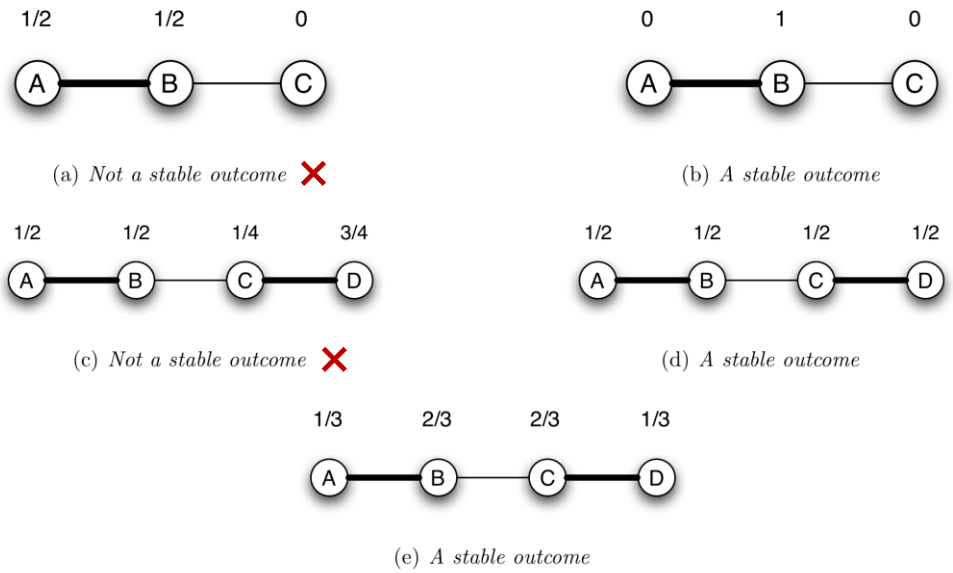
• 结果的稳定性

- 谈判，总要有个“结果”
  - 回顾纳什议价解中的基本要求： $x + y \leq 1$ 时，才有讨论议价解的必要性
  - 可以归纳出“不稳定性”的条件：结果中存在某一条边，两个端点X、Y价值和小于1
    - 本质上，这是一种“不平等条约”，在这种情况下，这条边的双方不如抛开现有的分配，直接自己成对去分配利益



• 结果的稳定性

- 谈判，总要有个“结果”
  - 与之相应的，是有关“稳定性”的说明：当前图中不包含任何不稳定性
    - 现在再来看看，下图中哪些分配是不稳定的？



(a) *Not a stable outcome* ✗

(b) *A stable outcome*

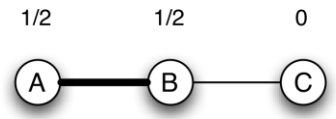
(c) *Not a stable outcome* ✗

(d) *A stable outcome*

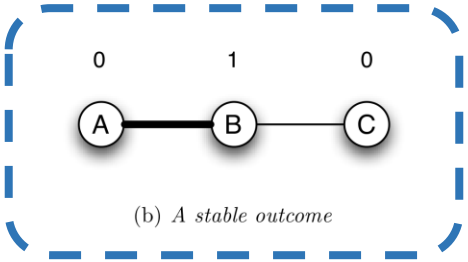
(e) *A stable outcome*

• 结果的稳定性

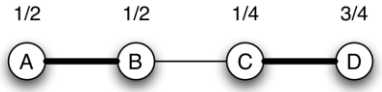
- 谈判，总要有个“结果”
  - 需要注意的是，“稳定”并不一定意味着“很可能出现”或者“温和的结果”
    - 例如下图中的(b)，分配结果对于A和C非常不利，但结果是稳定的



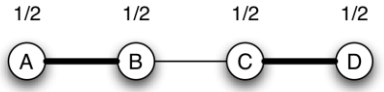
(a) *Not a stable outcome*



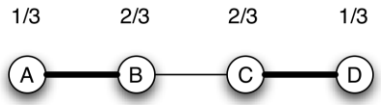
(b) *A stable outcome*



(c) *Not a stable outcome*



(d) *A stable outcome*

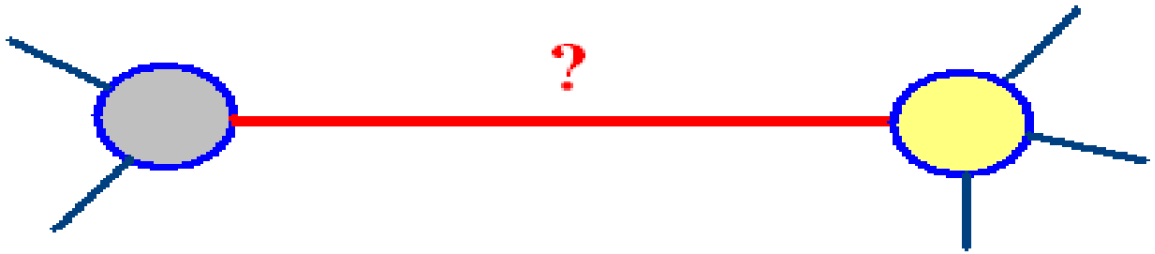


(e) *A stable outcome*

怎么样的分配是合理的？

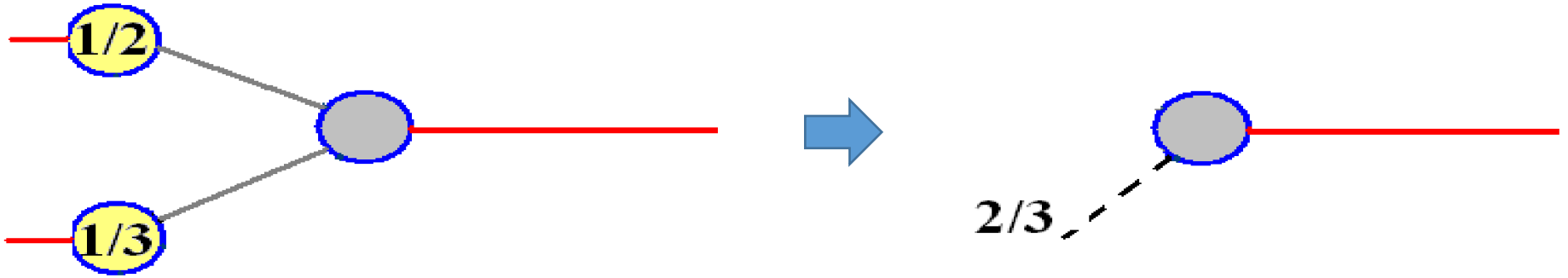
- **结果的平衡性**

- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 基于纳什议价解，如果价值分配是满足纳什议价解的，那么双方应该是满意的/平衡的
    - 问题在于，如何衡量纳什议价解的外部选项？
    - 更直接地说，就是纳什议价解中 x 和 y 的取值从何而来？



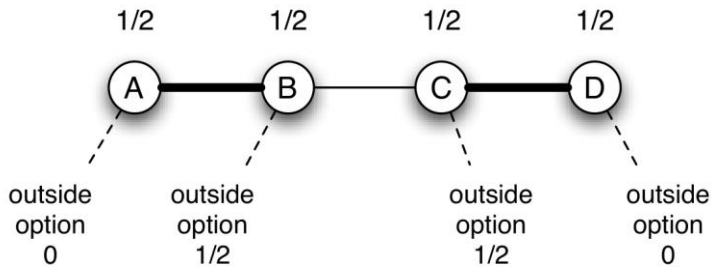
- **结果的平衡性**

- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 外部选项实际上体现了“网络其他部分的影响”
  - 一种直观的思路是，外部选项 = 放弃当前分配后所能获得的最大好处
    - 换言之即所谓“退路”（备胎）的价值



- **结果的平衡性**

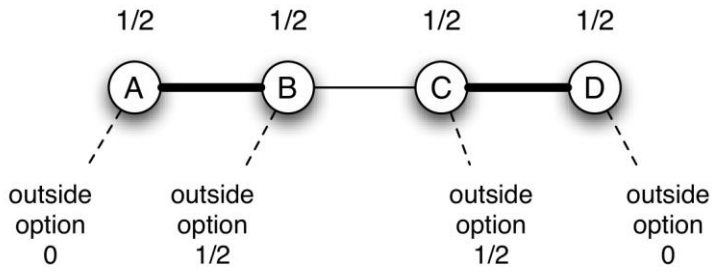
- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 我们以先前的一个“稳定结果”为例，来看看其是否具有平衡性



  - 图中形成了A-B，C-D两组分配，其中A、D没有备选项，外部选项为 0
  - B、C有外部选项，其价值来自于对方现有的收益，即1- ½ = ½
    - 注意，这里的外部选项是根据对方收益来的，即需要让出这部分收益来吸引对方

- **结果的平衡性**

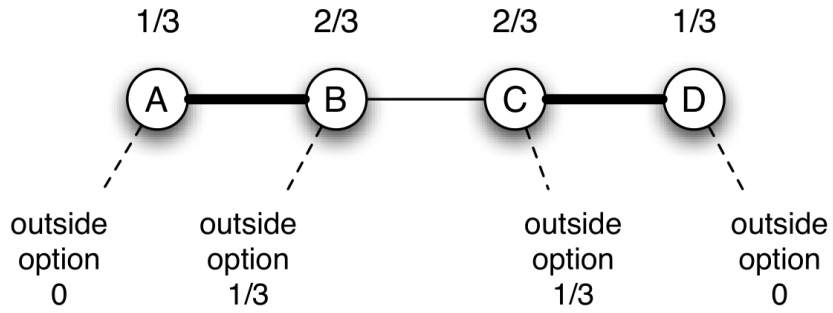
- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 我们以先前的一个“稳定结果”为例，来看看其是否具有平衡性



  - 因此，考虑A-B这对分配，其中 $x = 0$，$y = ½$，显然 $s = 1-x-y = ½$
  - 那么，在纳什议价解中，B的期望收益应该是 $y + s/2 = ¾$，低于现在的收益
    - 这种情况下，A-B的分配是不平衡的，B应该获得更高的收益

- **结果的平衡性**

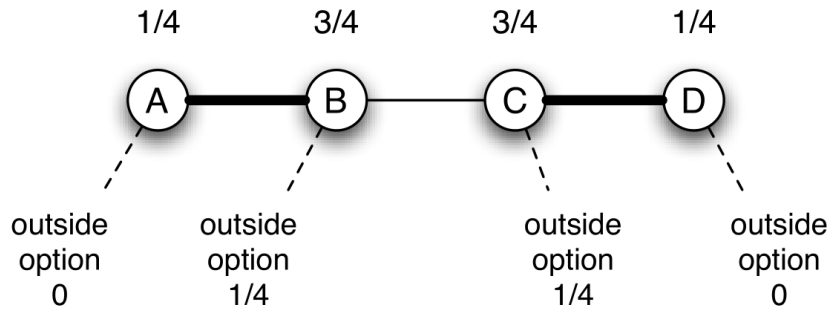
- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 我们再来看一个“平衡”的例子



  - 同样考虑A-B这对分配，其中 $x = 0$，$y = 1/3$，显然 $s = 1-x-y = 2/3$
  - 那么，在纳什议价解中，B的期望收益应该是 $y + s/2 = 2/3$，恰好等于现在的收益
    - 这种情况下，A-B的分配是平衡的

- **结果的平衡性**

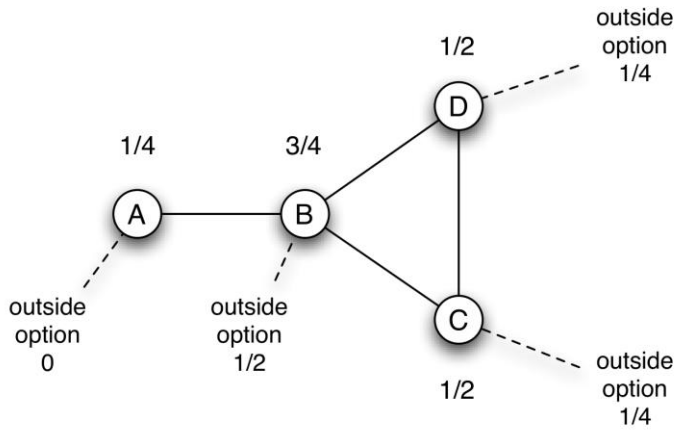
- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 另一种方式的“不平衡”



  - 同样考虑A-B这对分配，其中 x = 0，y = 1/4，显然 s = 1-x-y = 3/4
  - 那么，在纳什议价解中，A的期望收益应该是 x + s/2 = 3/8，低于现在的收益
    - 这种情况下，A-B的分配是不平衡的，且不平衡的因素来自于A

- **结果的平衡性**

- 稳定意味着结果不可能被外人破坏，但这种分配是否合理呢？
  - 事实上，这种“平衡性”的分析往往更接近实验结果，也具有更好的可解释性



  - 例如，我们先前讨论过的“柄图”，由于C-D的均势，B可以获得一个1/2的外部选项
    - 此时，B的期望收益可达½ + ½*½ = ¾，体现出了高于4-节点路径的优势（7/12 – 2/3）

本节目录

- 网络中的议价与权力
  - 网络中的权力交换
  - 纳什议价模型
  - 网络交换的稳定性问题
- **网络节点的地位**
  - 基于中心性的启发式衡量
  - PageRank及其衍生模型
  - HITS算法：地位与类别的双重区分

从节点权力到节点地位

- 先前，我们探讨了由于网络结构导致的节点议价权力不同
- 其中，四个形式化论述中的“中心性”引发了我们的思考
  - 在考虑信息流动的情况下，高介数的节点往往具有更高的权力
  - 其核心在于对于信息流通渠道的控制 / 垄断
  - 如何量化衡量这种垄断地位?

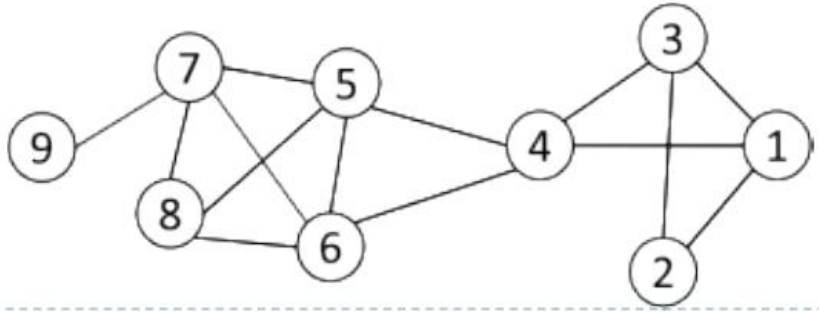
- **如何量化衡量节点的网络地位**

- 启发式方法（1）PageRank及其衍生模型
  - PageRank及其各种衍生算法如HITS都可以采用
    - 我们将在后半段，详细介绍PageRank及其各种算法的计算方式

- 启发式方法（2）中心性（Centrality）度量
  - 用于衡量网络中最重要的节点。常见中心性度量如度（Degree）、紧密度（Closeness）、介数（Betweenness）等

- **如何量化衡量节点的网络地位**

- 中心性度量（1） —— 度（Degree）
  - 基本思想：节点的邻居越多，对于信息传递的贡献越大，地位也就越高
  - 背景事实：许多现实世界网络具有无尺度性，只有少数节点度比较高
  - 计算公式： $C_D(v_i) = d_i = \sum_j A_{ij}$
    - 如果进行归一化，则公式为： $C'_D(v_i) = d_i/(n-1)$
  - 计算实例：
    - 节点 1 的度为 3
    - 归一化度为 3 / (9-1) = 3/8



- **如何量化衡量节点的网络地位**

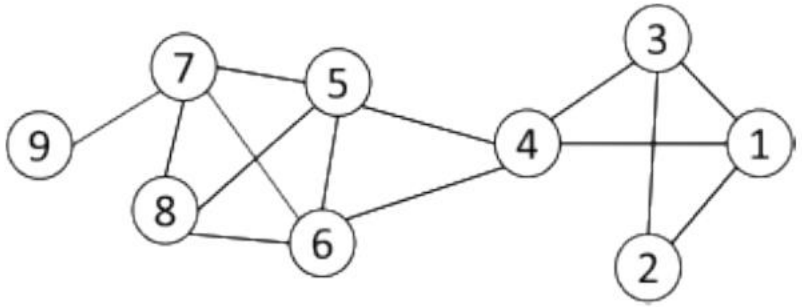
- 中心性度量（2） —— 紧密度（Closeness）
  - 基本思想：节点到其他节点越近，对信息传递的贡献越大，地位也就越高
  - 背景事实：距离越近，信息到达所需的中介越少，传递越快，成功率越高
  - 计算公式：

$$C_C(v_i) = \left[\frac{1}{n-1}\sum_{j\neq i}^{n} g(v_i, v_j)\right]^{-1} = \frac{n-1}{\sum_{j\neq i}^{n} g(v_i, v_j)}$$

    - 其中，$g(v_i, v_j)$ 定义为所谓测地距离（Geodesic Distance）
      - 即为两个节点之间的最短路径长度

- **如何量化衡量节点的网络地位**

- 中心性度量（2） —— 紧密度（Closeness）
  - 计算实例：



$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

- **如何量化衡量节点的网络地位**

- 中心性度量（3） —— 介数（Betweenness）
  - 基本思想：越多最短路径经过某节点，对信息传递的贡献越大，地位越高
  - 背景事实：与边介数思想类似，介数越高，信息传播中的中介地位越显著
  - 计算公式：

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

    - 其中，σ_{st} 表示节点 s 与 t 之间的最短路径数量（可能不止一条）
    - $\sigma_{st}(v_i)$ 表示s 至 t 的最短路径中经过节点 i 的路径数量

- **如何量化衡量节点的网络地位**

- 中心性度量（3） —— 介数（Betweenness）
  - 计算实例：

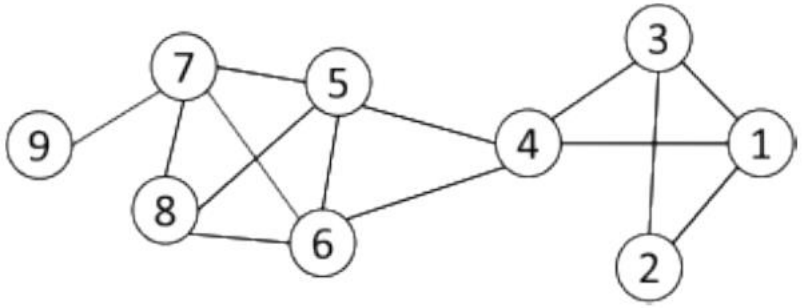


Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

	$s=1$	$s=2$	$s=3$
$t=5$	1/1	2/2	1/1
$t=6$	1/1	2/2	1/1
$t=7$	2/2	4/4	2/2
$t=8$	2/2	4/4	2/2
$t=9$	2/2	4/4	2/2

$$C_B(4) = 15$$

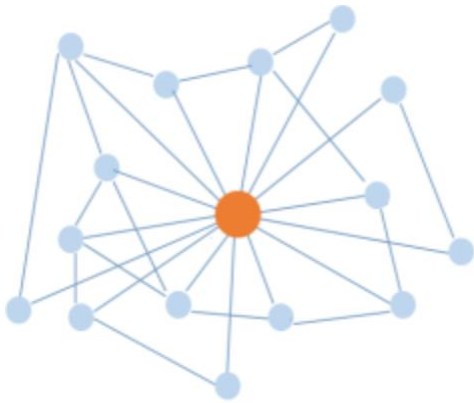
所有最短路径都经过节点 4

$$C_B(5) = 12 \times 0.5 = 6$$

但是，只有一半路径会经过节点 5（还可以走6）

- **中心性启发式方法的局限性**

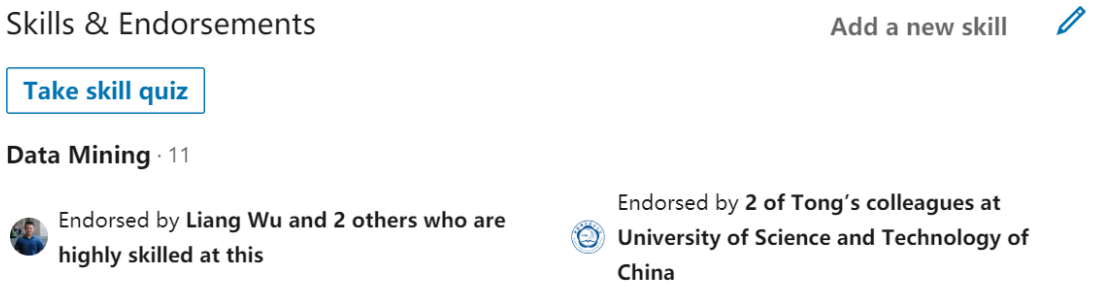
- 大致总结，可以将前述中心性度量方法分为局部 / 全局两类
  - 局部度量：如度
    - 仅考虑局部网络结构，局限性明显
  - 全局度量：如紧密度、介数
    - 考虑全局结构信息，但遍历最短路径开支巨大

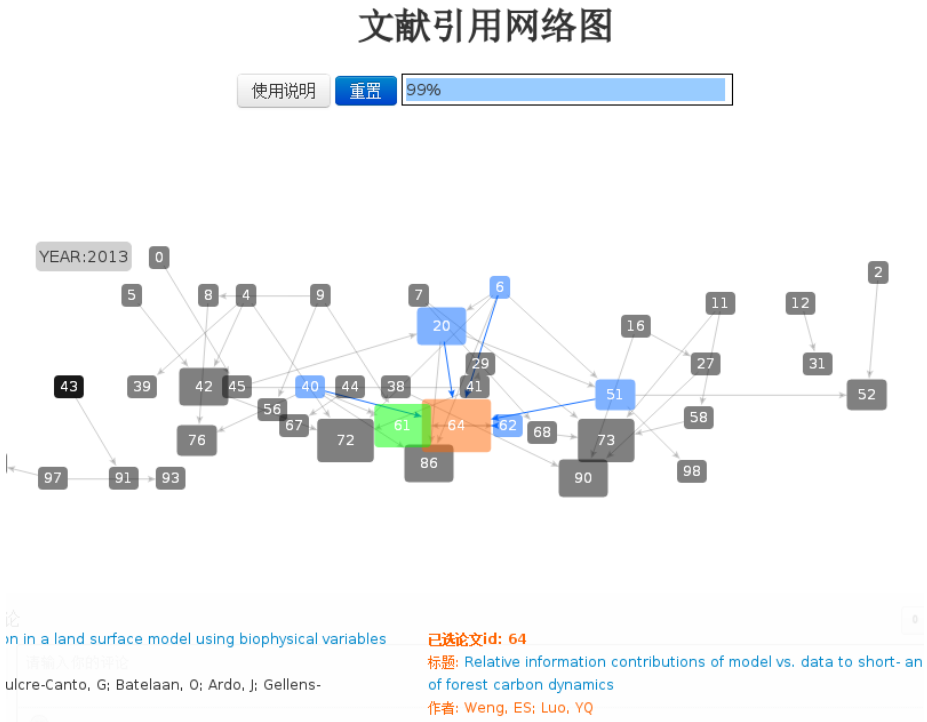


PageRank

- **如何有效度量全局性结构下的网络地位**

- 一个启发式想法：大家说好才是真的好：来自同行的endorsement
  - 区分“点赞”的人将获得更好区分效果



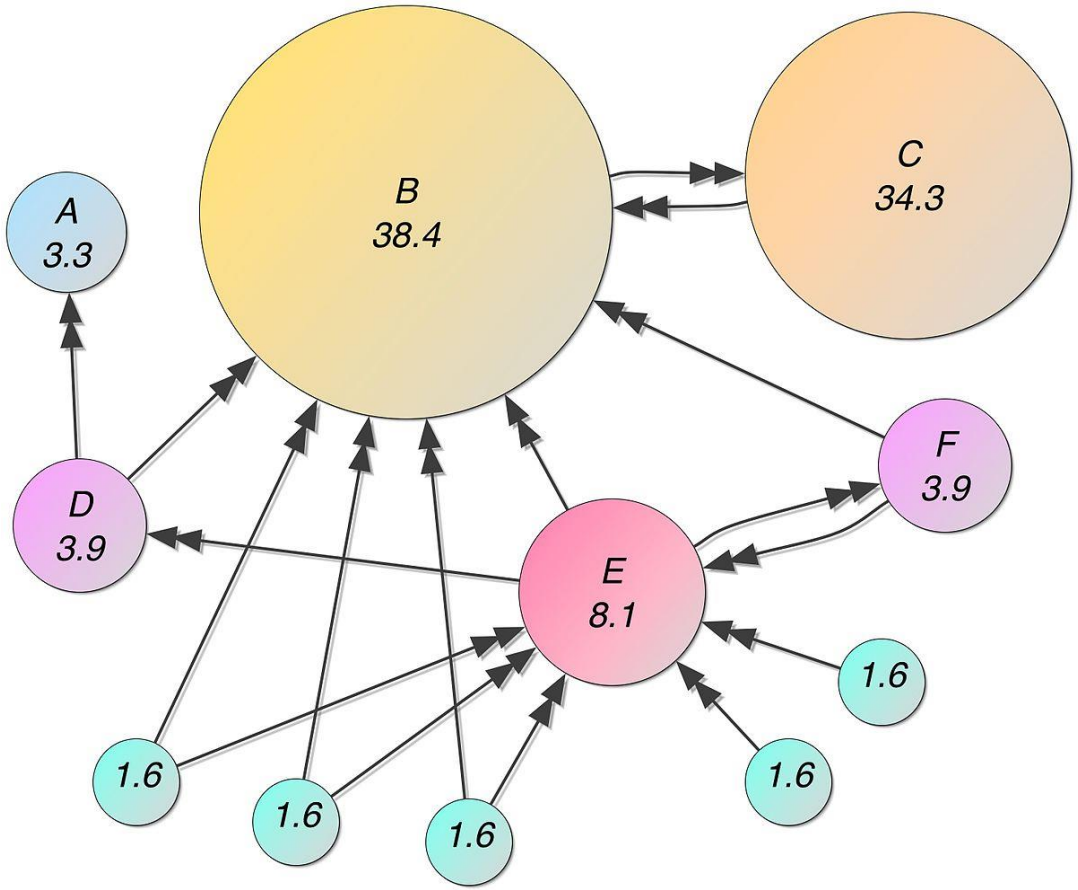


- **PageRank的历史**

- Sergey Brin 和Lawrence Page 在1998年提出了PageRank 算法
  - Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford InfoLab, 1999
    - 截止2024.3.18，引用18215次
    - http://www-db.stanford.edu/~backrub/pageranksub.ps
  - Sergey Brin, Lawrence Page, *The Anatomy of a Large-scale Hypertextual Web Search Engine*. WWW'98, Computer Networks 30(1-7): 107-117 (1998)
    - 截止2024.3.18，引用23783次
    - https://ai.google/research/pubs/pub334.pdf

- **PageRank的核心思想**

- 节点与节点之间的关系形成了一个巨大的有向图（注意方向性的引入）
  - 节点之间的有向连接表示了节点之间相互推荐的关系
  - 入度越多，则被推荐的次数越多，节点的地位就越高



- **PageRank的计算方法**

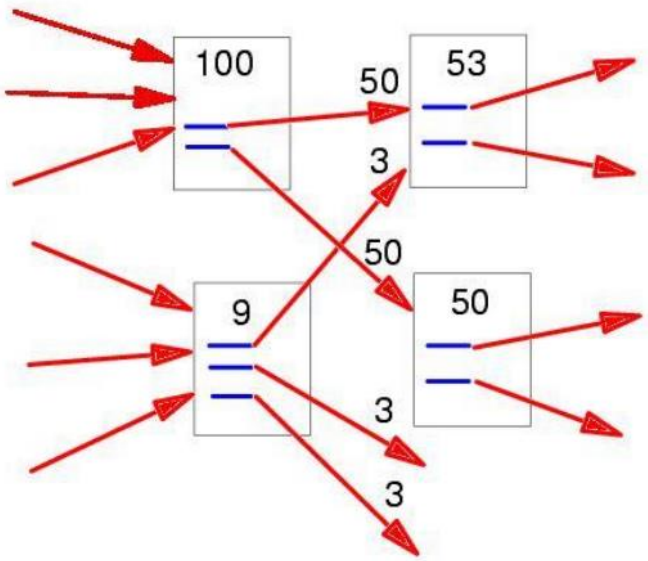
- PageRank的核心公式如下：

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

  - 其中：
    - PR(pi)为节点pi的PageRank值
    - PR(pj)为指向节点pi的某个节点pj的PageRank值
    - L(pj)为节点pj发出的链接数量
    - d为阻尼系数，取值在0-1之间
    - N为节点总数，M(pi)为链入pi的节点集合

• PageRank的设计思路

- PageRank背后的思想：
  - 总体思路：地位高的节点所推荐的节点，往往也是具有一定地位的
  - 三重衡量标准：
    - 链入链接数：单纯意义上的受推荐程度
    - 链入链接本身是否推荐度高：推荐本身是否具有权威性
    - 链入链接源页面的链接数：被选中的几率
      - 体现节点是否滥发推荐

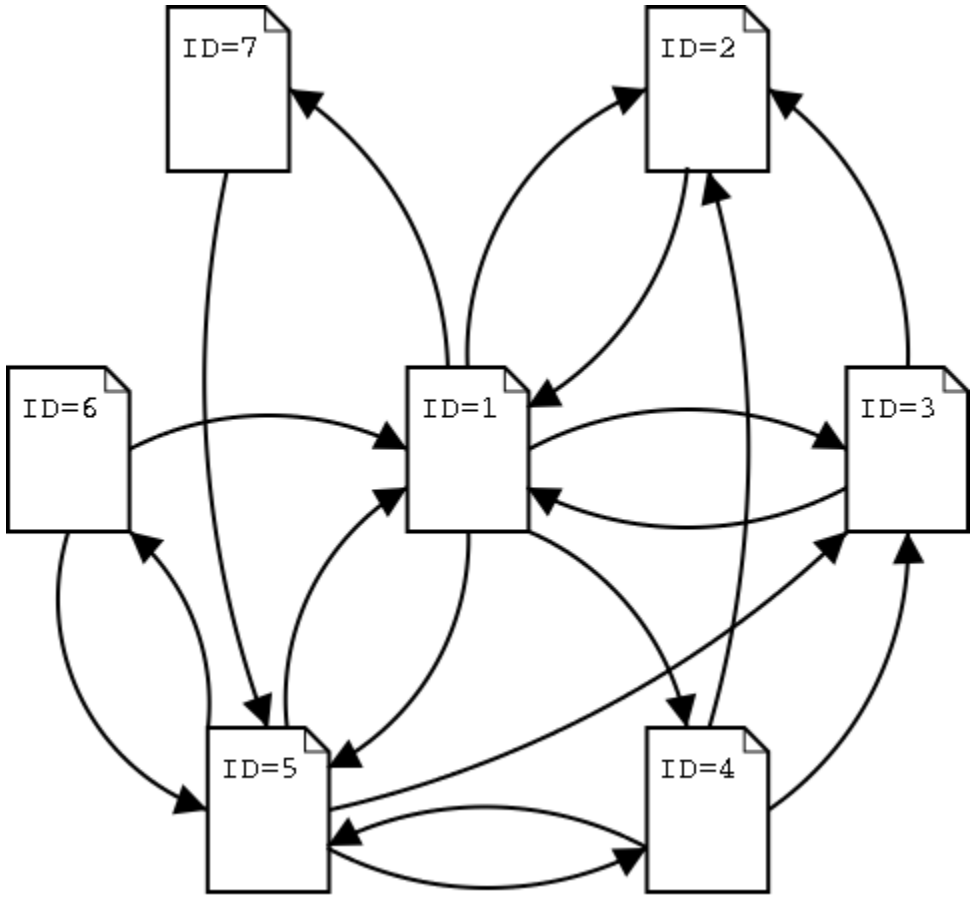


- **PageRank的实现过程**

- PageRank的迭代计算过程：
  - 采用近似迭代算法计算PageRank值
    - 首先给每个节点赋予一个初值，例如1/N
    - 然后，利用之前的公式进行迭代有限次计算，得到近似结果
- 先前提到的论文显示，实际进行大约100次迭代才能得到整个节点的PageRank。
- 中等规模的网络（如论文中提到的26M个网站），需要数小时完成这一过程

- **PageRank的计算实例**

链接源I D	链接目标 ID
1	2,3 ,4,5, 7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5



- **PageRank的计算实例**

- 从邻接矩阵到跳转矩阵
  - 倒置后，将各个数值除以非零元素数（即出边数量）

$$A=\begin{bmatrix} 0,1,1,1,1,0,1; \\ 1,0,0,0,0,0,0; \\ 1,1,0,0,0,0,0; \\ 0,1,1,0,1,0,0; \\ 1,0,1,1,0,1,0; \\ 1,0,0,0,1,0,0; \\ 0,0,0,0,1,0,0; \end{bmatrix}$$

$$R=\begin{bmatrix} 0, & 1, & 1/2, & 0, & 1/4, & 1/2, & 0; \\ 1/5, & 0, & 1/2, & 1/3, & 0, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 0, & 1/2, & 1; \\ 0, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 0, & 0, & 0; \end{bmatrix}$$

- **PageRank的计算实例**

- 开始计算，假设所有PageRank值初始均为1/N

$$R=\begin{bmatrix} 0, & 1, & 1/2, & 0, & 1/4, & 1/2, & 0; \\ 1/5, & 0, & 1/2, & 1/3, & 0, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 1/3, & 0, & 1/2, & 1; \\ 0, & 0, & 0, & 0, & 1/4, & 0, & 0; \\ 1/5, & 0, & 0, & 0, & 0, & 0, & 0; \end{bmatrix} + M=\begin{bmatrix} 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \\ 1/7 \end{bmatrix}$$

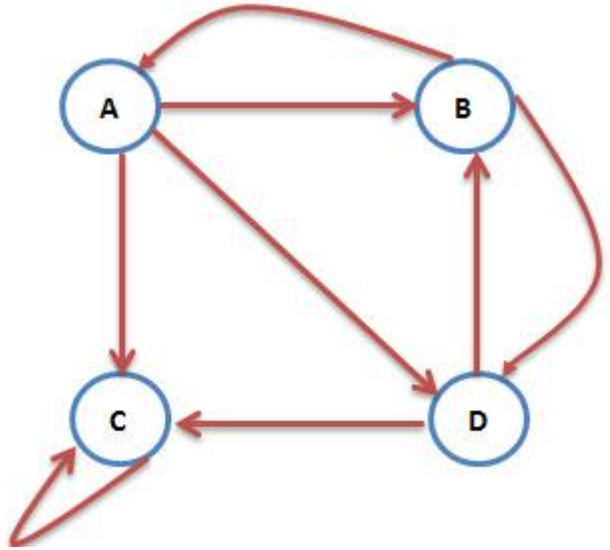
某轮迭代后，若M与M’对应维元素的差值高于阈值则继续迭代，直至收敛

当前M与M’差值过高，继续迭代

$$R \cdot M = \begin{bmatrix} 0.321 \\ 0.148 \\ 0.148 \\ 0.064 \\ 0.148 \\ 0.036 \\ 0.029 \end{bmatrix} = M'$$

- **PageRank中的两类特殊情况**

- PageRank的计算过程，实际上是一个马尔科夫过程
  - 如果计算中出现陷阱节点或终止节点，如何处理？
    - 陷阱节点：只有一条指向自己的边，没有其他出边
    - 终止节点：没有任何出边，如同黑洞
- 此外，孤立节点的存在也会产生一定影响
  - 没有任何入边，单纯使用马尔科夫链无法跳转
    - 仅有初始概率，不再更新，也不影响其他节点

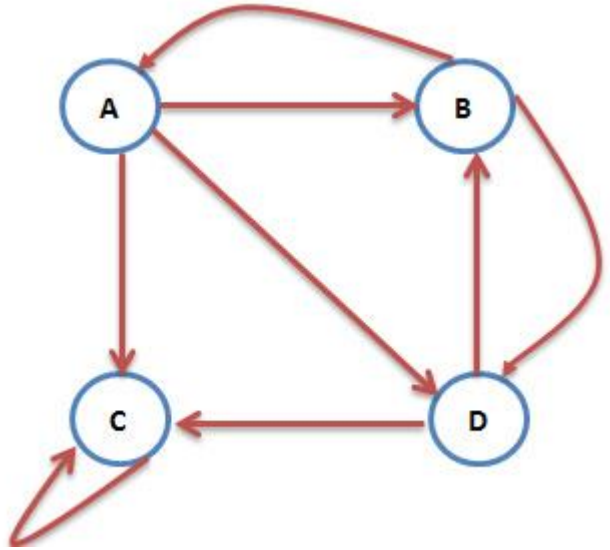


- **PageRank中的两类特殊情况**

- 几类特殊情况的解决：Restart机制
  - 回顾PageRank的公式

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

  - 其中的(1-d)/N的部分，相当于以一定概率重新选择起点
    - 此时，所有节点以一定等概率被选中
    - 由此，跳出了陷阱和黑洞的干扰
    - d一般选择为0.85左右



- **PageRank中的收敛性问题**

- 如前所述，PageRank的计算过程，实际上是一个马尔科夫过程。

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

  - 取$A = dM + [(1-d)/N]ee^T$，其中M为跳转矩阵，e为所有元素都为1的列向量
  - 则有$Pn+1 = APn$，形成马尔科夫过程

- 前述论文已从理论上证明，不论初始PageRank值如何选取，这种算法都保证了网页排名的估计值能收敛到他们的真实值。如何保证?

- **PageRank中的收敛性问题**

- 马尔科夫过程的三个收敛条件：
  - 转移矩阵A为马尔科夫矩阵
    - A矩阵所有元素都大于等于0，并且每一列的元素和都为1
  - 转移矩阵A为不可约的
    - 当图是强连通时，A为不可约，而Restart保障了这一条件
  - 转移矩阵A为非周期的
- 这三个条件，PageRank算法都满足，因此保障了其收敛性。

- **PageRank的拓展模型：个性化PageRank**

- PageRank在衡量节点的全局性地位方面具有显著的意义
- 然而，节点的个性化因素却没有在PageRank中得以体现
  - 每个节点有自己特定的偏好，这种偏好会影响其关系形成
  - 事实上，这种思想我们在第二章“链接预测”中曾经有所涉及

$$\mathrm{PR}(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{\mathrm{PR}(p_j)}{L(p_j)} \qquad \vec{q_x} = cP^T\vec{q_x} + (1-c)\vec{e_x},$$

  - 其中需要特别注意的是：起点限定为预设的某个节点

- **PageRank的拓展模型：主题敏感PageRank**

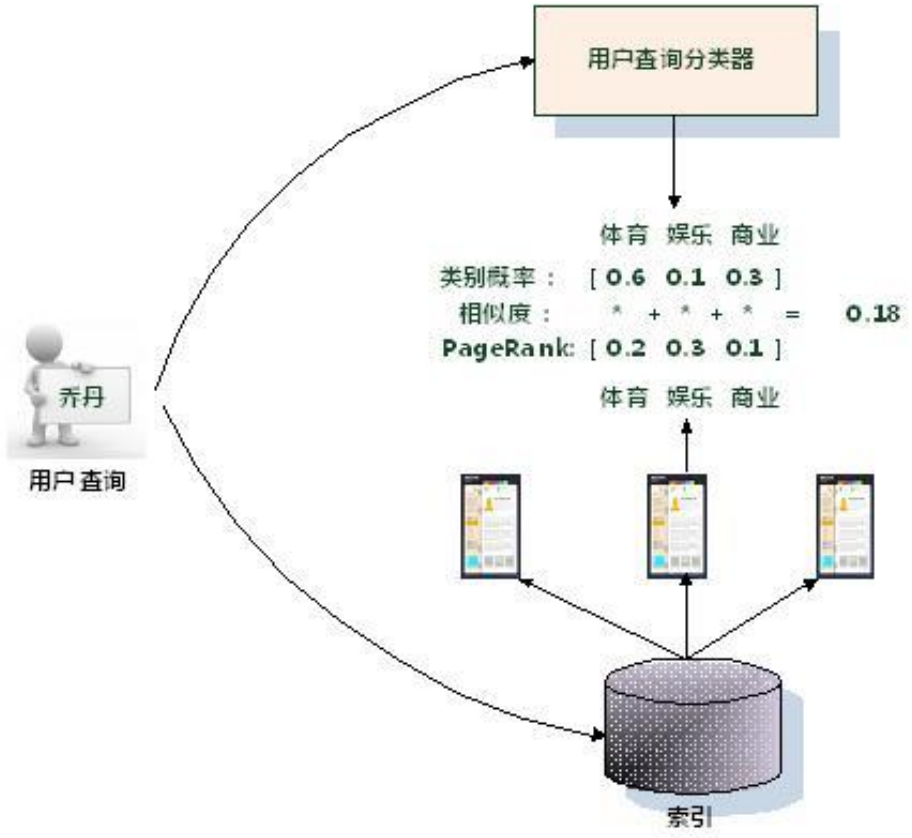
- 更进一步，考虑到针对每个节点进行个性化衡量代价巨大
- 如何既在一定程度上考虑偏好因素，又减轻计算资源的负担？
  - 一个折中的方案是，以主题为中介，为特定的主题计算相应的PageRank
    - TH Haveliwala, Topic-sensitive pagerank, WWW‘02，期刊版本引用3648次
    - 某种意义上，这一模型体现了“仅有同行的评价才有价值”的思想
      - 隶属不同主题的节点，其相互推荐的作用将被削弱

• PageRank的拓展模型：主题敏感PageRank

- Topic-sensitive PageRank与基本PageRank的区别
  - 首先，每个节点隶属于某个特定主题
    - 在论文中，作者参考Open Directory Project划分了16个类别
  - 其次，在计算部分，区别主要在于起始节点和Restart部分
    - 对于某个Topic来说，起点仅限于隶属于该Topic的节点
    - (1-d)e/N更改为了(1-d)s/|s|，s为一个N维向量
      - 如果某个节点隶属于该主题，则 s 中的该维为1，反之为0
      - |s| 表示 s 中为 1 的元素的个数。

• PageRank的拓展模型：主题敏感PageRank

- Topic-sensitive PageRank的输出
  - 最终，每个节点将得到一个Topic-sensitive的向量
    - 即使不属于某个Topic，节点也可以通过PageRank获得相应的数值
  - 在这种情况下，如果将主题的需求同样表示为向量，两个向量的内积可以反应节点在这种主题需求下的权威性



PageRank的总结

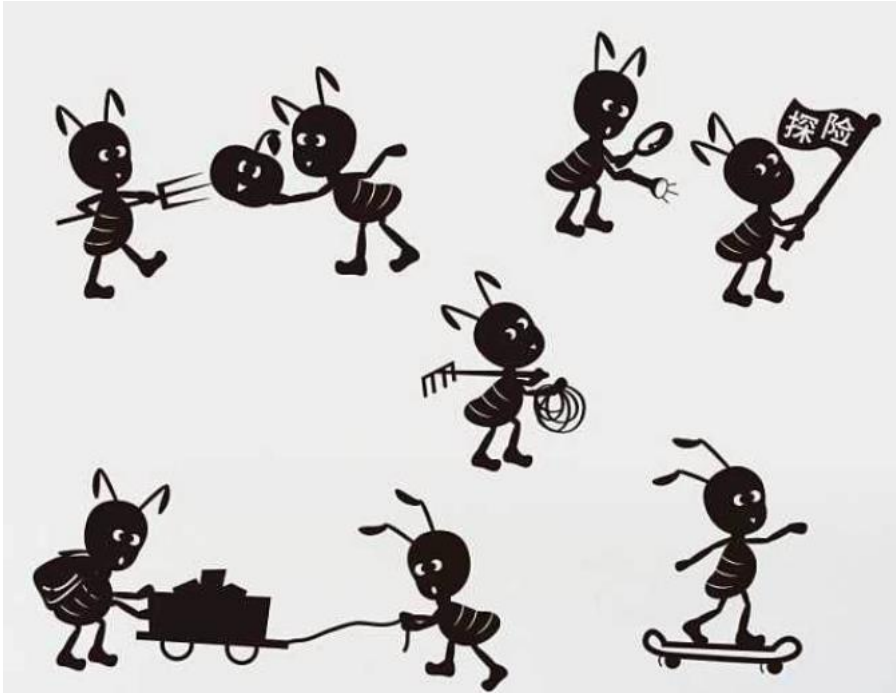
- PageRank是一种基于链接分析的全局节点地位量化算法
- 优点
  - 对于节点给出全局性的地位排序，并且可以离线完成以保障效率
  - 基础的PageRank算法本身独立于主题，可以实现结果的通用
- 缺点
  - 主题无关性，同时对于恶意链接、广告链接等缺乏区分
  - 旧节点得分更高，因为新节点往往少有入链（现象普遍存在，可能导致富者愈富）
  - 一般不能单独用于排序，需要与相关性排序方法相结合
  - 效率问题（很多图问题的通病）

- **从PageRank到HITS算法**

- Hilltop算法尝试了网页角色的区分，但仍有进一步的空间
  - 对于基础PageRank算法而言，不同链接的对外输出是平均的
  - 即使通过个性化或主题敏感进行了修正，但对节点的类型仍不区分
    - 以网站为例，专业信息站点与门户网站，在提供信息的种类上区别很大
    - 因此，将对外链接的输出和功能视作等价不符合实际情况

- **网络中的节点存在着不同**

- 另一方面，研究者也发现，在同一个社会网络中，不同节点往往也扮演不同角色
  - 例如，一个部门，往往有以下分工：
    - 部门经理：负责领导部门
    - 技术专家：负责提供技术指导
    - 项目经理：负责外部需求沟通
    - ……



- **从PageRank到HITS算法**

- 在PageRank提出的同一年（1988），康奈尔大学的Jon Kleinberg（回忆一下W-S-K模型？）提出了Hyperlink – Induced Topic Search（HITS）
  - Kleinberg博士认为，在衡量节点重要性时，也需要对于节点的两类不同功能进行区别对待

Kleinberg J M. Authoritative sources in a hyperlinked environment, In Proceedings of the ACM-SIAM symposium on discrete algorithms (SODA 1998). 1998. （引用3130次）

Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM (JACM), 1999, 46(5): 604-632. （引用10650次）

- **HITS算法的两个核心概念（节点基本功能）**

- 权威（Authority）节点与枢纽（Hub）节点的区分
- 权威节点：指某个领域或某个话题相关的高质量节点
  - 如科研领域的中科院之声，视频领域的优酷与爱奇艺等
- 中心节点：类似中介，指向了很多高质量的权威节点
  - 如“hao123”，各个浏览器自带的首页（手动滑稽）
- HITS的目的即找到并区分这些高质量“Authority”与“Hub”节点
  - 在搜索场景下“Authority”更为重要（因为更能满足用户的信息需求）

- **HITS算法的两个基本假设**

- 基本假设与核心概念相互对应
  - 基本假设1：好的Authority会被很多好的Hub指向

$$\forall p, a(p) = \sum_{i=1}^{n} h(i)$$

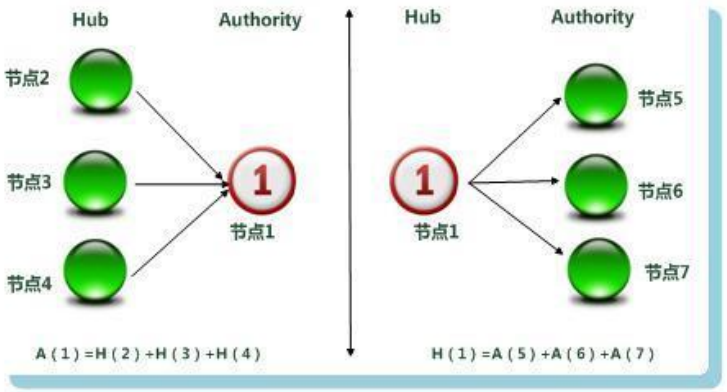
  - 基本假设2：好的Hub会指向很多个好的Authority

$$\forall p, h(p) = \sum_{i=1}^{n} a(i)$$

  - 因此，在HITS算法中，每个节点需要计算两个权威值

- **HITS算法的计算过程**

- 基于先前的基本假设，HITS的计算过程如下：
- 首先，根据关键字获取与查询最相关的少数节点，及与这些页面有链接关系的节点，作为待选集合
- 其次，对所有节点的 a(p) 与 h(p) 进行初始化，可都设为1
- 最后，迭代计算两个步骤，即基本假设所对应的两个公式
  - 重复这一步，直到最终收敛为止
    - 注意每一步中的归一化过程！
  - 输出Authority或Hub值较高的节点



- **HITS算法的计算过程**

- 基于先前的基本假设， HITS的计算过程如下：
  - 假定邻接矩阵为M，Authority向量为a，Hub向量为h
  - 则有如下迭代式：
    - $a_{k+1} = M^T h_k$， $h_{k+1} = M a_{k+1}$
  - 或者，可采用如下迭代式：
    - $a_{k+1} = (M^T M)^k M^T a_0$， $h_{k+1} = (M M^T)^{k+1} h_0$
    - 其中，a_0, h_0为Authority/Hub向量的初始值，可设为全1向量

HITS算法的优缺点

- 相比于PageRank，HITS算法是一种能够区分节点功能的排序算法
- 优点
  - 更好地区别描述不同节点的不同功能
  - 主题相关，因此可以单独用于面向特定主题的节点排序
- 缺点
  - 需要在线计算，时间代价较大
  - 对链接结构变化敏感，且依然可能受到“链接作弊”的影响

本章小结

节点地位与节点权力

- **网络中的议价与权力**
  - 网络中的权力交换
  - 纳什议价模型
  - 网络交换的稳定性问题
- **网络节点的地位**
  - 基于中心性的启发式衡量
  - PageRank及其衍生模型、HITS算法

tongxu@ustc.edu.cn