



华东師範大學

EAST CHINA NORMAL UNIVERSITY

数据科学与工程算法基础

Algorithm Foundations of Data Science and Engineering

第五章 随机游走及其应用

$$(1+x)^n = 1 + \frac{nx}{1!} + \frac{n(n-1)x^2}{2!} + \dots$$

课程提纲

Content

1 课程引入

2 随机游走

3 网页排名算法PageRank

课程提纲

Content

1 课程引入

2 随机游走

3 网页排名算法PageRank

联合概率分布

- 离散随机变量序列 X_1, \dots, X_n 的联合概率 $P(X_1 = x_1, \dots, X_n = x_n)$
 - 如果 X_1, \dots, X_n 互相独立, $\prod_{i=1}^n P(X_i = x_i)$
 - 否则, $P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_1 = x_1)$
- 链式规则难以扩展到拥有很多随机变量的问题中
 - 至少有 2^{i-1} 个条件分布 $P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_1 = x_1)$
 - 在实际应用中通常简化上述模型
 - ✓ 在语言模型中, 常假设词之间满足一阶相关, 即
$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1})$$
 - ✓ 在金融领域, 假设股票价格满足 t -阶相关, 即
$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_{i-t} = x_{i-t})$$

随机过程

- 一组随机变量序列组成**随机过程**, 记作 $\{X(t) \in S : t \in I\}$
 - **时间集合** t : 表示时间, I 可能是 $\mathbb{R}, \mathbb{Z}, \mathbb{Z}^+$
 - **状态空间** S : 表示随机变量可能取值组成的集合
- **举例**

例子	时间集合	状态空间
浏览网页	步数	所有网页
自然语言	单词位置	字典
股票价格	交易日	价格
赌博	下注次数	筹码数量
液体分子	时间	容器空间

The screenshot shows a Google search results page for the query "seo". The results are filtered by "Everything". The first result is for "Top SEO Specialists" with a link to www.seop.com. The second result is for "Improve Your SEO Ranking" with a link to www.vivahotels.com/seo/. The third result is for "SEO, Web Optimization - Search Engine Optimization Experts" with a link to www.vivahotels.com/seo/. The fourth result is for "Search engine optimization - Wikipedia" with a link to en.wikipedia.org/w/index.php?title=Search_engine_optimization&oldid=3291320. The fifth result is for "Search Engine Optimization (SEO) - Webmaster Tools Help" with a link to support.google.com/webmasters/bin/answer.py?hl=en&answer=70002. The sixth result is for "Web Site SEO Marketing" with a link to www.allwebmarketing.com. The right sidebar features a "People and Pages on Google+" section with profiles for Rand Fishkin and Denny Su. A red box highlights this sidebar area.



抛币游戏

- 游戏规则
 - 假设开始有 \$10, 参加一场公平无限期抛硬币游戏
 - ✓ 正面朝上, 赢得 \$1
 - ✓ 反面朝上, 输掉 \$1
 - 我们用 X_n 表示 n 次抛币后手上剩余的美元
 - 如果 $X_0 = 10$, 则 $\{X_n : n \in N\}$ 构成一个随机序列
 - ✓ 假如 $X_n = 12$, 则 $X_{n+1} = 11$ 或者 13
 - ✓ $(n + 1)$ 次抛币后手上剩余的美元和 X_{n-1} 的值没有关系, 即
$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$
- 抛币游戏中, 手上剩余美元数量是一个随机过程
 - 时间: 抛币次数
 - 状态: 非负整数集合

象棋

- 象棋中帅的位置变化

- 按照规则，帅在每个位置都可以向周围移动
- X_n 表示 n 步之后帅所在的位置

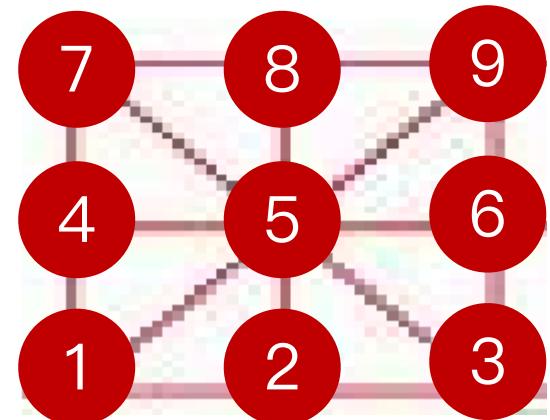
- ✓ 如果 $X_n = 2$, 则 $X_{n+1} = 1, 3, 5$

- ✓ $(n + 1)$ 步帅所在的位置和 X_{n-1} 的值没有关系, 即

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

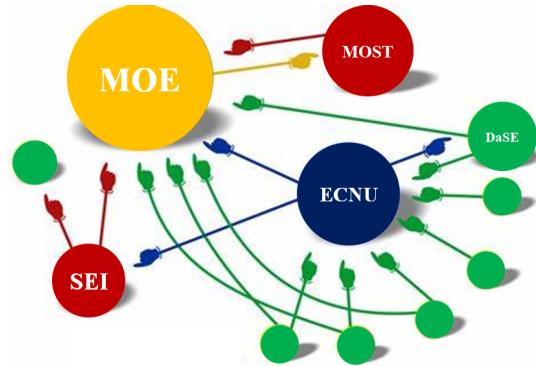
- $\{X_n : n \in N\}$ 构成一个随机过程

- 时间：步数
 - 状态：1 到 9 之间的整数



网页浏览

- 每一个网页可能有对外的锚链接
 - ECNU 指向页面 MOE、DaSE 和 SEI
 - X_n 表示 n 步之后停留的页面
 - ✓ 如果 $X_n = \text{ECNU}$, 则 $X_{n+1} = \text{MOE}$ 、 DaSE 或者 SEI
 - ✓ $(n+1)$ 步停留的页面和 X_{n-1} 的所停留的页面无关, 即
$$P(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$
- $\{X_n : n \in N\}$ 构成一个随机过程
 - 时间: 步数
 - 状态: 所有网页组成的集合



每个页面的重要程度
不同, 如何对网页进
行排序?

课程提纲

Content

1 课程引入

2 随机游走

3 网页排名算法PageRank

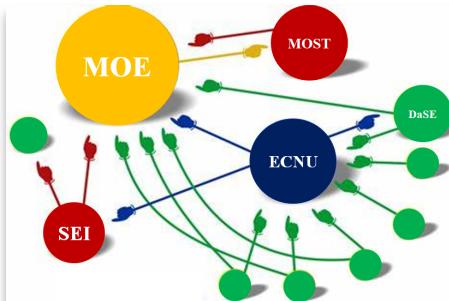
马尔可夫性质

- 如果随机过程 $\{X(t) \in S, t \in I\}$ 满足

$$P(X(t_{n+1}) = x_{n+1} | X(t_0) = x_0, \dots, X(t_n) = x_n) = P(X(t_{n+1}) = x_{n+1} | X(t_n) = x_n)$$

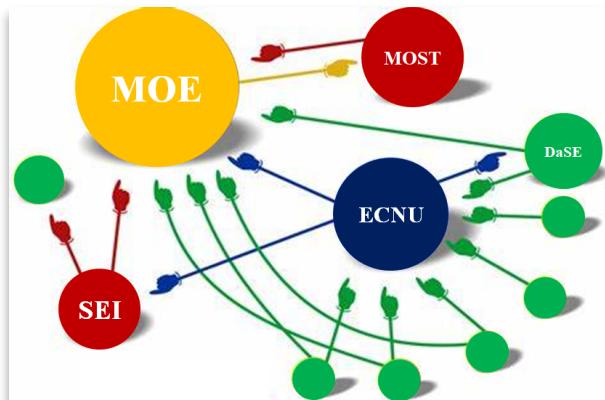
其中 $t_0 < t_1 < \dots < t_n < t_{n+1} \in I, x_i \in S$, 则称该随机过程
满足**马尔可夫性质**

- 如果 t_n 表示当前时间, t_{n+1} 表示将来, t_1, \dots, t_{n-1} 则表示过去
- 可以解释为: “已知现在, 将来与过去独立” (无记忆性)
- 满足马尔科夫性质的例子



随机游走

- 若随机过程 $\{X(t) \in S, t \in I\}$ 满足马尔可夫性质，则称其为**马尔可夫过程**
- 马尔可夫链是特殊的马尔可夫过程，它满足
 - 离散的时间集合：步数、下注次数
 - 离散的状态空间：网页、计数
- 马尔可夫链又被称为**随机游走**
- 例子

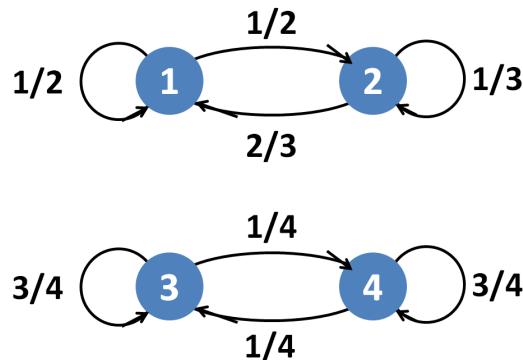


概率转移矩阵

- 假定随机游走的状态空间为 $\Omega = [n]$, 矩阵 $\mathbf{P}^{(t+1)} \in \mathbb{R}^{n \times n}$ 被称为该随机游走的第 $(t + 1)$ 步概率转移矩阵, 如果 $p_{x,y}^{(t+1)} = P(X_{t+1} = y | X_t = x)$

- 其中 $p_{x,y}^{(t+1)}$ 表示在 $(t + 1)$ 步随机游走从状态 x 转移到 y 的概率
- 概率转移矩阵的每行概率和为 1, 即

$$\sum_y p_{x,y}^{(t+1)} = \sum_y \frac{P(X_{t+1} = y | X_t = x)}{P(X_t = x)} = 1$$



$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

$$p_{1,2}^{(t+1)} = \frac{1}{2}$$

$$p_{1,4}^{(t+1)} = 0$$

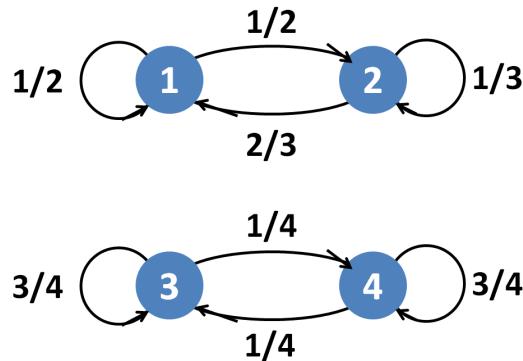
齐次随机游走

- 随机游走是齐次的，如果它满足

$$P(X_{t+s} = y | X_t = x) = P(X_s = y | X_0 = x)$$

- 即经过 t 步从状态 x 走到状态 y 的概率与起始步数 s 无关，则该随机游走是齐次的
- 因此概率 $p_{x,y}^{(t+1)}$ 可以简化为 $p_{x,y}$ ，因为

$$P(X_{t+1} = y | X_t = x) = P(X_1 = y | X_0 = x)$$



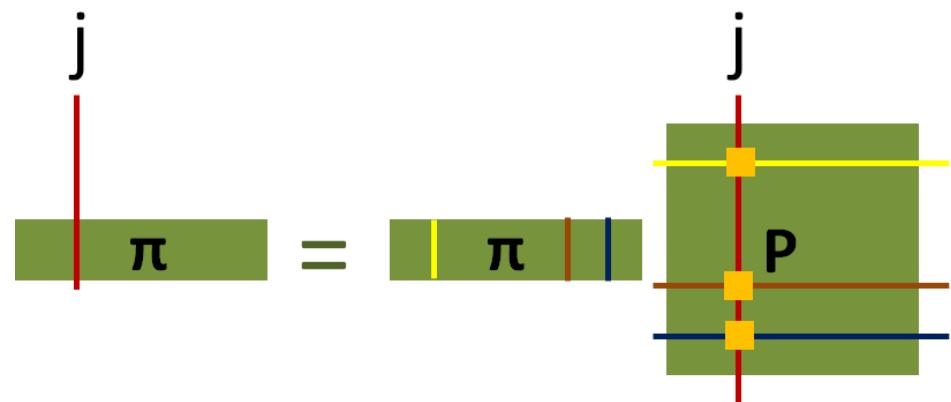
$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

$$p_{1,2} = \frac{1}{2}$$

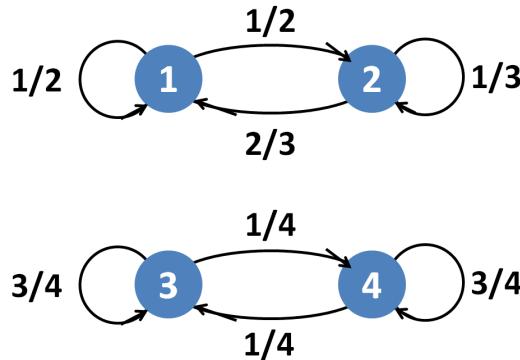
$$p_{1,4} = 0$$

状态分布

- 设 $\pi(t)$ 表示随机游走在时间 t 的状态概率分布，定义为
 $\pi_x^{(t)} = P(X_t = x)$
- 若状态空间为 $\Omega = [n]$ ，则状态概率分布 $\pi_x^{(t)}$ 为 n 维向量，且满足 $\sum_{x \in \Omega} \pi_x^{(t)} = 1$
- 进一步 $\pi_j^{(t+1)} = \sum_i P(X_t = i)P(X_{t+1} = j | X_t = i) = \sum_i \pi_i^{(t)} p_{i,j} = (\pi^{(t)} \mathbf{P})_j$
- 因此， $\pi^{(t+1)} = \pi^{(t)} \mathbf{P}$



状态概率分布的例子



$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

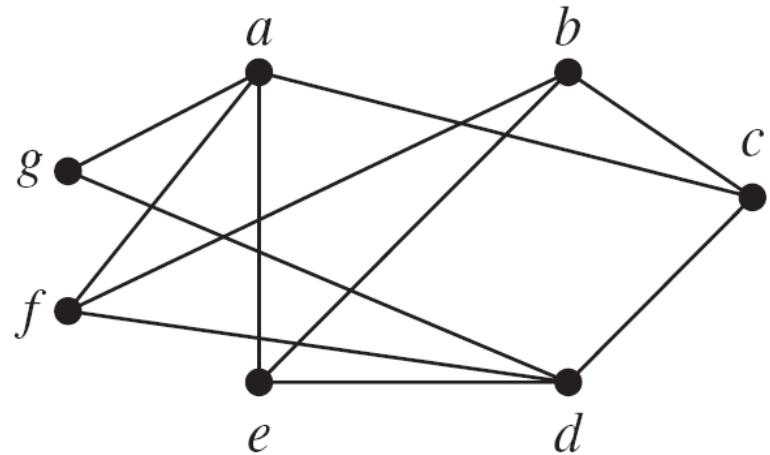
$$(0.4, 0.6, 0, 0) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = (0.4, 0.6, 0, 0).$$

Example $(0, 0, 0.5, 0.5) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = (0, 0, 0.5, 0.5).$

$$(0.1, 0.9, 0, 0) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = (0.35, 0.65, 0, 0).$$

图上的随机游走

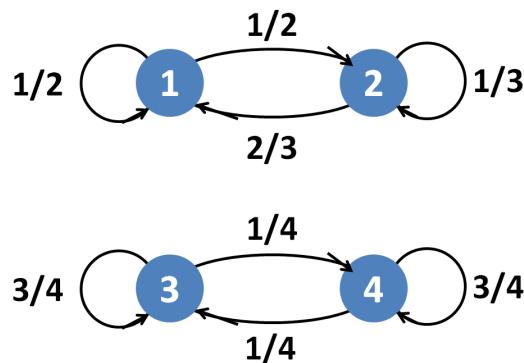
- 图 $G = (V, E)$, 其中 V 是顶点集合, E 为边的集合
 - 转移概率定义为
$$p_{x,y} = \begin{cases} \frac{1}{\deg(x)}, & \text{if } (x, y) \in E \\ 0, & \text{otherwise} \end{cases}$$
 - 定义 $\pi_0(a) = 1$, $\pi_0(x) = 0$ ($x \neq a$) 表示开始我们从顶点 a 出发
- 因此, 图可建模成一个随机游走



$$\deg(a) = 4, \deg(b) = 3$$

平稳分布

- 一个概率转移矩阵为 P 的齐次有限随机游走的**平稳分布**为 π , 如果满足 $\pi P = \pi$
 - 对于某些随机游走经过一段时间后, 无论初始的概率分布是什么, 它的状态概率分布都不会再发生变化
 - 这是因为 $\pi P^n = \pi P^{n-1} = \dots = \pi P^1 = \pi$



$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

状态分布 $(0.4, 0.6, 0, 0)$ 和 $(0, 0, 0.5, 0.5)$ 都是该随机游走的平稳分布

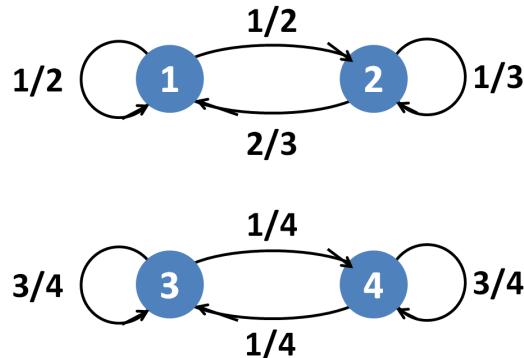
平稳分布的存在性

- 如果一个随机游走是不可约且非周期的，且具有状态转移矩阵 \mathbf{P} 。那么 $\lim_{n \rightarrow \infty} \mathbf{P}_{ij}^n$ 存在且独立于 i ，记为 $\lim_{n \rightarrow \infty} \mathbf{P}_{ij}^n = \pi(j)$ ；同时我们有

$$\lim_{n \rightarrow \infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

其中 π 是方程 $\pi\mathbf{P} = \pi$ 的唯一非负解，即 π 为该随机游走的平稳分布

反例



$$\mathbf{P}^{20} = \begin{pmatrix} 0.4 & 0.6 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{pmatrix}$$

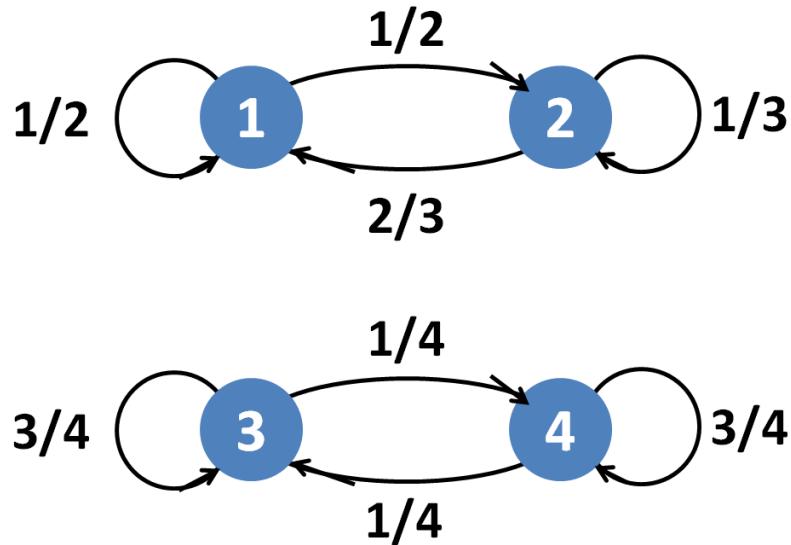
- 为什么和定理描述的不同呢?

- 状态分布 $(0.4, 0.6, 0, 0)$ 和 $(0, 0, 0.5, 0.5)$ 都是该随机游走的平稳分布
- 该马尔可夫链实际上可以拆成两个相互独立的马尔可夫链
 - ✓ 每个马尔可夫链都有自己的稳态分布
 - ✓ 我们把这种情况称为是**可约的**

不可约性

- 如果能从状态 x 出发到达状态 y , 则称状态 y 是从状态 x 是可达的, 换句话说 $\exists n, P_{x,y}^n > 0$ 。如果状态 x 和状态 y 互相可达, 则称它们是连通的
- 当马尔可夫链的所有状态均连通时, 称此马尔可夫链是不可约的
 - y 从 x 可达意味着有一条从 x 指向 y 的路径
 - x 与 y 是连通的意味着 x 与 y 是强连通的
 - 一个有限的马尔可夫链是不可约的当且仅当转换图是强连通的
 - 连通关系满足:
 - ✓ 反身性: x 与 x 是连通的
 - ✓ 对称性: x 与 y 是连通的当且仅当 y 与 x 是连通的
 - ✓ 传递性: 如果 x 与 y 是连通的且 y 与 z 是连通的, 则 x 与 z 是连通的

可约马尔可夫链 I



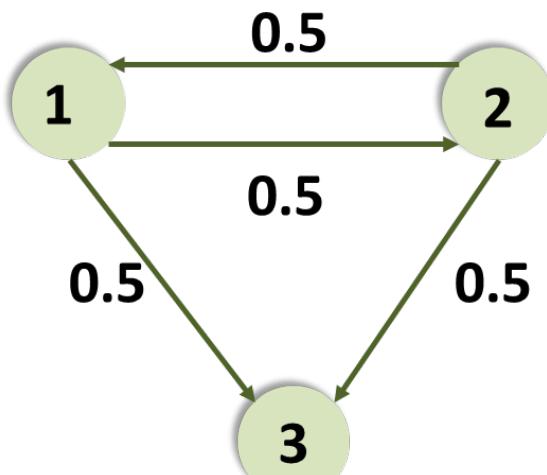
$$(0.4, 0.6, 0, 0) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = (0.4, 0.6, 0, 0)$$

$$(0, 0, 0.5, 0.5) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} & \frac{3}{4} \end{pmatrix} = (0, 0, 0.5, 0.5)$$

- 可约马尔可夫链

- 由两个独立的联通分量组成
- $(0.4, 0.6, 0, 0)$ 和 $(0, 0, 0.5, 0.5)$ 都是该随机游走的平稳分布
- 因此，可约马尔可夫链平稳分布可能不唯一

可约马尔可夫链 II



$$\mathbf{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 \\ 0 & 0 & 0 \end{pmatrix}$$

- 非强连通图

- 顶点 3 没有出边

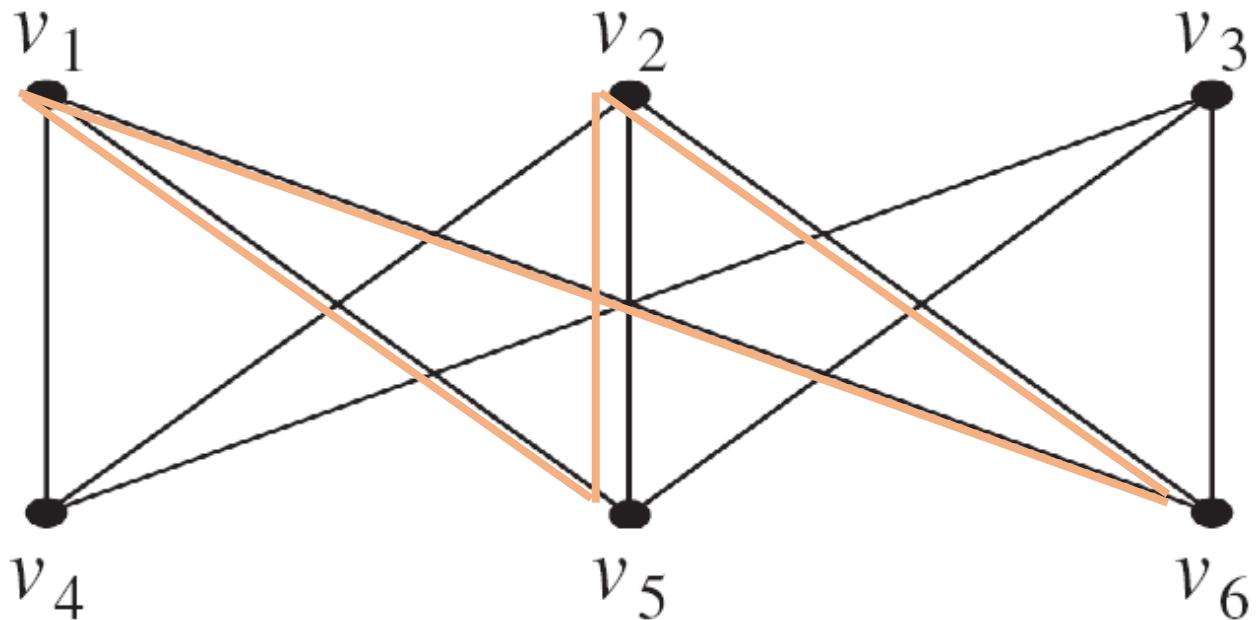
- $\lim_{n \rightarrow \infty} \mathbf{P}^n = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

- 因此，可约马尔可夫链也可能不存在平稳分布

周期性

- 马尔可夫链状态 x 的周期为最大公因数 (gcd)，使得 $d_x = \text{gcd}\{n \mid \mathbf{P}_{x,x}^n > 0\}$ 。如果 $\forall n \geq 1, \mathbf{P}_{x,x}^n = 0$ ，则 $d_x = \infty$ 。
 - 在一个有限马尔可夫链中， $\mathbf{P}_{x,x}^n > 0$ 意味着有一个经过状态 x 且长度为 n 的环
 - x 的周期是所有通过 x 的环的长度的最大公因数
- 某个状态的周期是 1，则该状态是**非周期的**
- 如果一个马尔可夫链的所有状态都是**非周期的**，则称该马尔可夫链是**非周期的**

例子

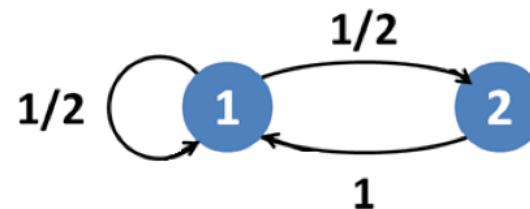


- 状态 v_1 是非周期的吗?
 - 所有经过状态 v_1 的环的长度为偶数
 - 状态 v_1 的周期为 2, 因此状态 v_1 是周期的
- 二分图是周期的
- 周期马尔可夫链一定不存在平稳分布

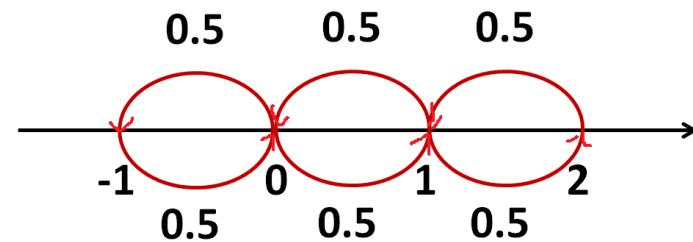
周期的性质

- 如果状态 x 和状态 y 是连通的, 则 $d_x = d_y$
 - 如果一个马尔可夫链是不可约的, 那么所有状态都具有相同的周期
- 如果 $n \bmod d_x \neq 0$, 那么 $\mathbf{P}_{x,x}^n = 0$

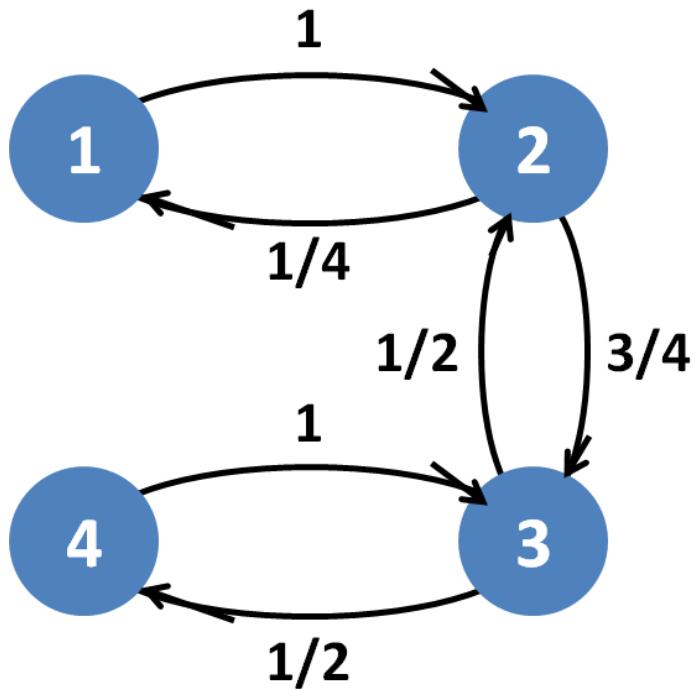
- 右图是否是周期的?



- 右图是否是周期的?



例子



- 这个马尔可夫链是否可约?
- 是否是非周期的?
- 周期是多少?
- 是否存在平稳分布?

课程提纲

Content

1 课程引入

2 随机游走

3 网页排名算法PageRank

网页排序

Microsoft Bing

国内版 国际版

PageRank algorithm

ALL IMAGES VIDEOS

7,760,000 Results Any time ▾

Page Rank Algorithm and Implementation - GeeksforGeeks

<https://www.geeksforgeeks.org/page-rank-algorithm-implementation> ▾

Aug 30, 2017 - PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages.

User rating: 3.5/5

如何对满足检索要求的网页进行排序？

<https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank> ▾

- 5.1.1. History and explanation. PageRank is named after Google co-founder Larry Page, and is used ...
- 5.1.2. Use-cases - when to use the PageRank algorithm. PageRank can be applied across a wide ...
- 5.1.3. Constraints - when not to use the PageRank algorithm. If there are no links from within a ...

The PageRank algorithm - Cornell University

pi.math.cornell.edu/~web6140/TopTenAlgorithms/PageRank.html ▾

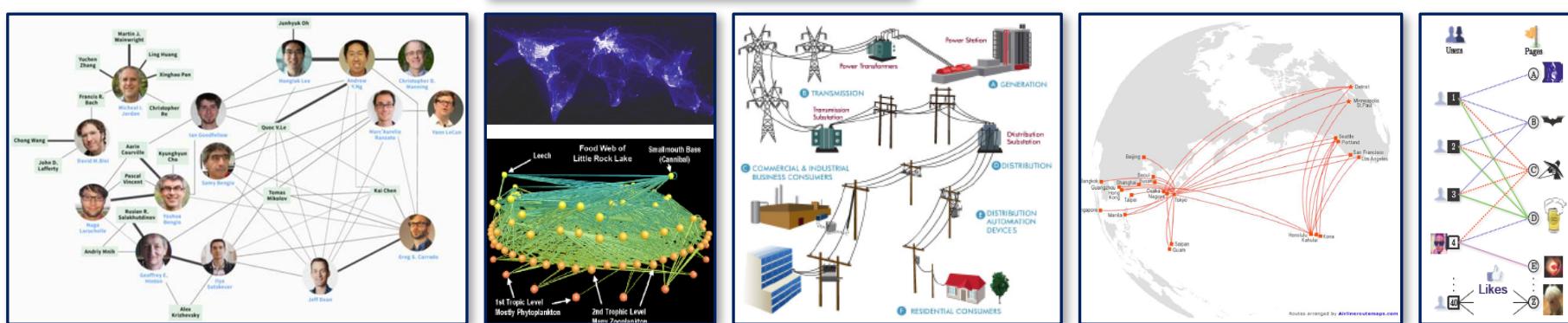
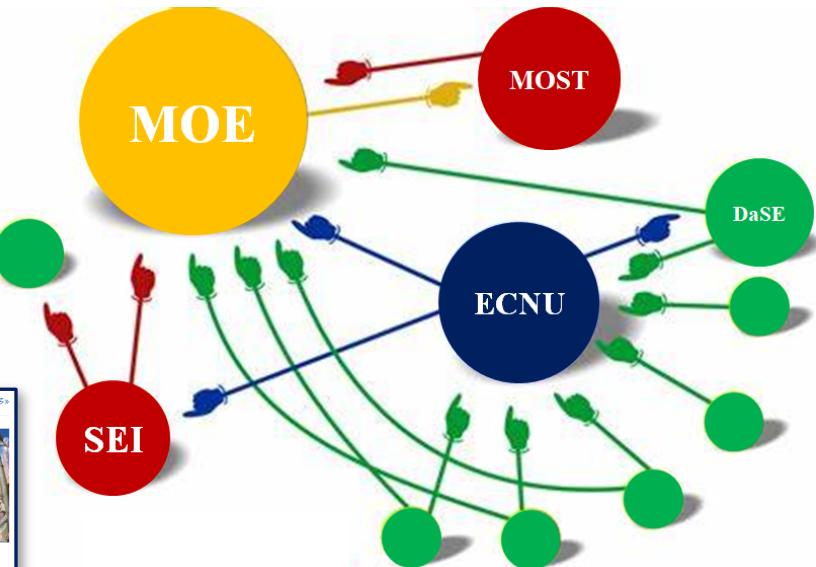
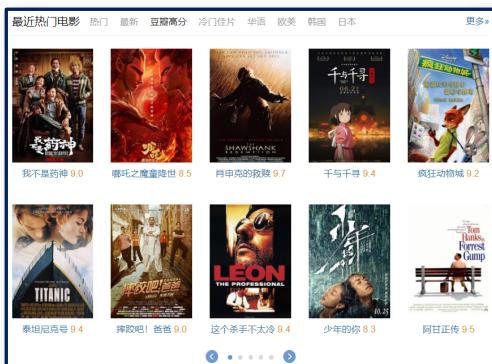
The PageRank algorithm ¶ As the internet rapidly grew in the 1990s, it became increasing difficult to find the right webpage or nugget of information. The internet was missing a homepage that could be a portal to the rest of the web. The simple idea ¶

PageRank (Google) Algorithm Explained | Global Software ...

<https://www.globalsoftwaresupport.com/pagerank-algorithm-explained> ▾

图顶点排序问题

- 网页排序是图顶点排序的特例
- 图在我们日常生活中无处不在
 - 合作网络
 - 食物链
 - 电力网络
 - 航空运输网络
 - 服务订阅
 - 商品购买
 - 影评/书评

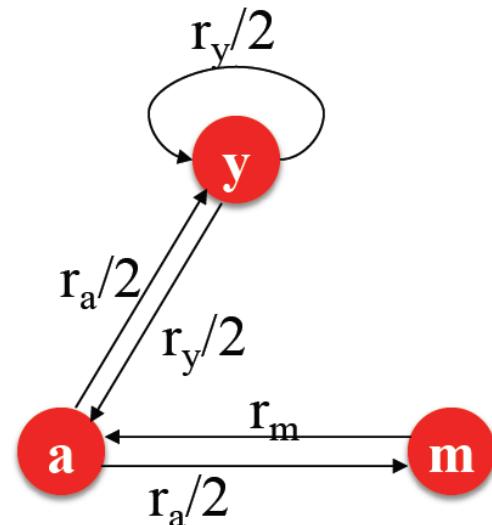


网页重要性

- 很多人访问的网页自然是重要的，但无法获知每个人的行为
- 若用户浏览网页是随机的点击行为
 - 从一个随机网页开始，通过网页上的锚链接随机进入下一个页面（状态转移）
 - 所有用户越可能停留（访问）的页面越重要
 - $\text{PageRank} = \text{停留在某个网页上概率}$ （平稳分布）
- 假设
 - 并非所有网页都同等“重要”
 - 网页拥有更多入链，那么该网页更重要

递归公式

- 若网页 v 的重要性是 r , 且有 n 条出链, 每个出链对其他顶点的投票权重为 $\frac{r}{n}$, 网页 v_a 的重要程度是它入边的投票权重之和 $r_a = \frac{1}{2}r_y + r_m$
- 得到方程组
$$\begin{cases} r_a = \frac{1}{2}r_y + r_m \\ r_y = \frac{1}{2}r_y + \frac{1}{2}r_a \\ r_m = \frac{1}{2}r_a \end{cases}$$
- 缺少常数项, 不存在唯一解
- 因为是平稳分布, 所以增加约束 $r_a + r_y + r_m = 1$
- 得到 $r_a = \frac{2}{5}, r_y = \frac{2}{5}, r_m = \frac{1}{5}$



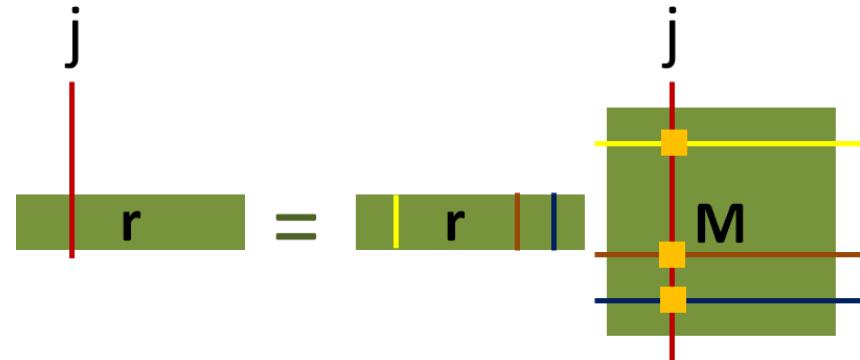
矩阵形式

- 图可以建模成随机游走
- 图 $G = (V, E)$, 其中 V 是顶点集合, E 为边的集合, 转移概率矩阵定义为 $M_{x,y} = \begin{cases} \frac{1}{\deg(x)}, & \text{if } (x,y) \in E \\ 0, & \text{otherwise} \end{cases}$
- 若平稳分布为 r , 根据平稳分布的定义 $r = rM$
- r 为转移概率矩阵 M 最大特征值 1 对应的左特征向量

PageRank 算法

- 矩阵形式的算法

```
1: Set  $r^{(0)} = (\frac{1}{n}, \dots, \frac{1}{n})$ ;  
2: For  $k = 1, 2, \dots$ ;  
3: Let  $r^{(k)} = r^{(k-1)} \cdot M$ ;
```



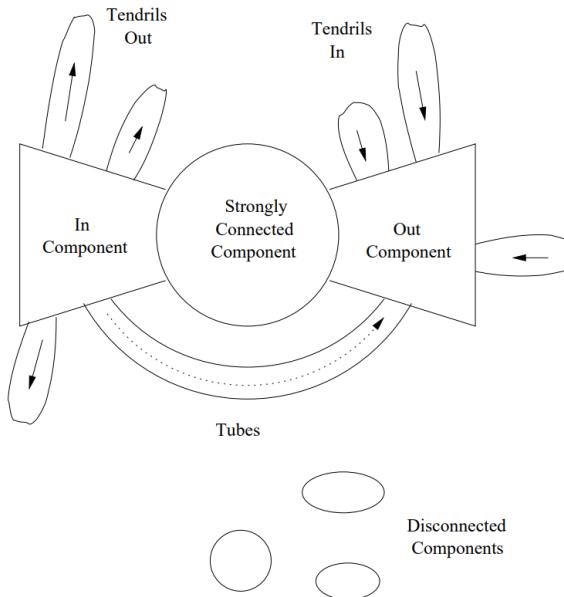
- 稀疏矩阵形式的算法

```
1: Set  $r^{(0)} = (\frac{1}{n}, \dots, \frac{1}{n})$ ;  
2: For  $k = 1, 2, \dots$ ;  
3: For  $j = 1, 2, \dots, n$ ;  
4: Let  $r_j^{(k)} = \sum_{i \rightarrow j} \frac{r_i}{n_i}$ ;
```

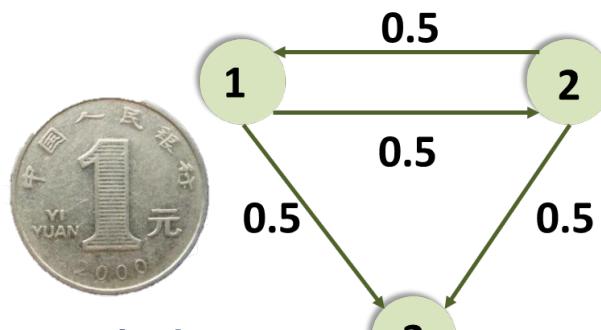
- 算法会收敛吗?

- 存在唯一平稳分布吗?
 - ✓ 是否是不可约的?
 - ✓ 是否是非周期的?

PageRank 算法改进 I



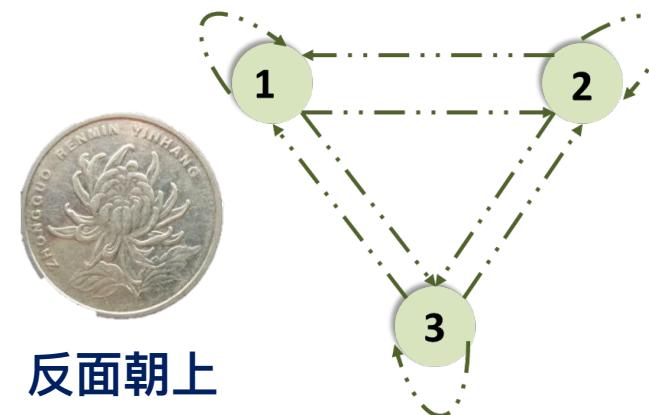
- 修改转移概率矩阵 $\widetilde{M} = \beta M + (1 - \beta)[\frac{1}{n}]_{n \times n}$
- 修改的合理性
 - 以概率 β 按原来的图结构进行网页浏览
 - 以概率 $1 - \beta$ 从一个新的网页开始浏览
 - 实践中, β 一般被设置为 0.85
- 强连通 + 非周期



正面朝上
的概率为 β

or

反面朝上
的概率为 $1 - \beta$



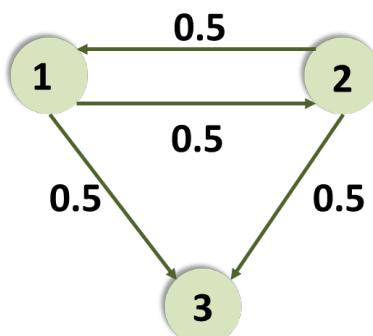
存在的问题

- 原始矩阵 M 是稀疏的，而 \widetilde{M} 是非常密集的
- 简化

$$\begin{aligned}r_j &= \sum_{i=1}^n \widetilde{M}_{ij} r_i = \sum_{i=1}^n (\beta M_{ij} + \frac{1-\beta}{n}) r_i \\&= \sum_{i=1}^n \beta M_{ij} r_i + \sum_{i=1}^n \frac{1-\beta}{n} r_i \\&= \sum_{i=1}^n \beta M_{ij} r_i + \frac{1-\beta}{n}\end{aligned}$$

- 即 $r = \beta rM + [\frac{1-\beta}{n}]_n$
- 此外，矩阵 M 可能不是一个概率转移矩阵

解决泄露问题



$$M = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0 & 0 & 0 \end{pmatrix}$$

$$M^3 = \begin{pmatrix} 0 & 0.125 & 0.125 \\ 0.125 & 0 & 0.125 \\ 0 & 0 & 0 \end{pmatrix}$$

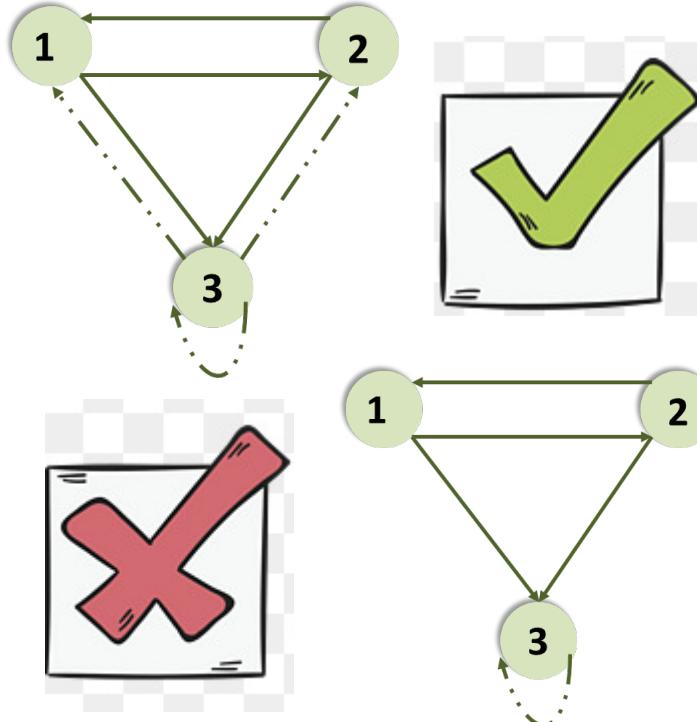
$$M^2 = \begin{pmatrix} 0 & 0.25 & 0.25 \\ 0.25 & 0 & 0.25 \\ 0 & 0 & 0 \end{pmatrix}$$

$$M^4 = \begin{pmatrix} 0 & 0.0625 & 0.0625 \\ 0.0625 & 0 & 0.0625 \\ 0 & 0 & 0 \end{pmatrix}$$

- 由于顶点 3 (dead end) 的存在造成 PageRank 值的泄露
- 再次修改
 - 顶点 3 可以随意跳转到其他任意一个顶点

$$\widetilde{M} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$

成为转移概率矩阵



Google PageRank 算法

- 完整版 PageRank 算法

1: set $\mathbf{r}^{old} = (\frac{1}{n}, \dots, \frac{1}{n})$;

2: repeat until convergence: $\sum_j |r_j^{new} - r_j^{old}| < \epsilon$;

3: $\forall j : r_j'^{new} = \sum_{i \rightarrow j} \frac{r_i^{old}}{n_i}$;

$r_j'^{new} = 0$ if in-degree of j is 0;

Now revise the random walk:

4: $\forall j : r_j^{new} = \beta r_j'^{new} + \frac{1-\beta s}{n}$, where $s = \sum_j r_j'^{new}$;

5: $\mathbf{r}^{old} = \mathbf{r}^{new}$;

- 一轮迭代中泄露的 PageRank 值为 $1 - \sum_j r_j'^{new} = 1 - s$

- 每个顶点的 PageRank 值来自三部分

$$\beta \left(\sum_{i \rightarrow j} \frac{r_i^{old}}{n_i} + \frac{1-s}{n} \right) + \frac{1-\beta}{n} = \beta r_j'^{new} + \frac{1-\beta s}{n}$$

总结

- Centrality 是一种度量图顶点重要性的方法
 - PageRank 只是其中的一种度量
 - 其他还有 Betweenness, Closeness, Cluster Coefficient 等
- 二分图是一类很重要的图
 - 对于随机游走是周期的，不存在平稳分布
 - 如何度量二分图顶点的重要程度？
- 马尔可夫链有很多经典的应用
 - 隐马尔科夫模型：HMM
 - 条件随机场：CRF
 - 马尔可夫链 Monte-Carlo 采样：MCMC