

2022-2023秋季课程:数据科学与大数据导论

Introduction to Data Science and Big data


# Chapter 8: Graph Data Analytics

曹劲舟 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2022年11月



# 图数据入门

## □我们身边有很多复杂的系统

- 截止到2019年4月1日，我们的**社会**包含7,579,185,859个**个体**，个体与个体之间有着频繁的**交流**
- 截止到2018年，中国的**手机用户**数量突破15亿，这些用户的移动设备之间通过复杂的通信网络彼此**连接**
- 截止到2011年10月，**全球网站**的总量突破5亿，这些网站上汇集了海量的彼此**关联**的信息与知识
- 我们的**大脑**由**860亿**个彼此相连、传递信息的**神经元**组成
- 我们的生命体由**基因与蛋白质**的复杂**交互关系**构成
- .....

## □思考：这些复杂的系统有什么**共性**？

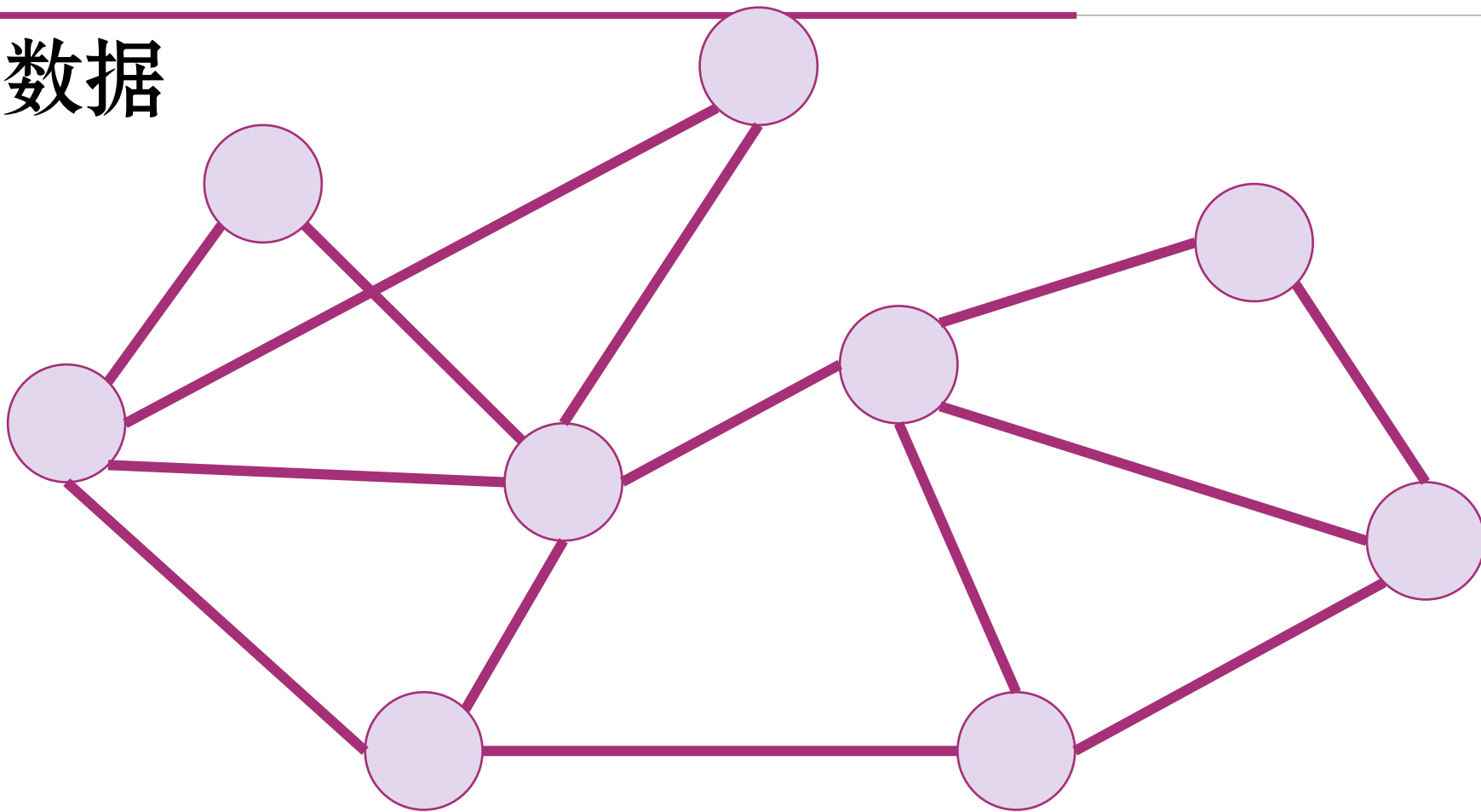


关系

# 图数据入门

---

## □图数据



图模型：表征事物之间的相互关联

# 图数据入门

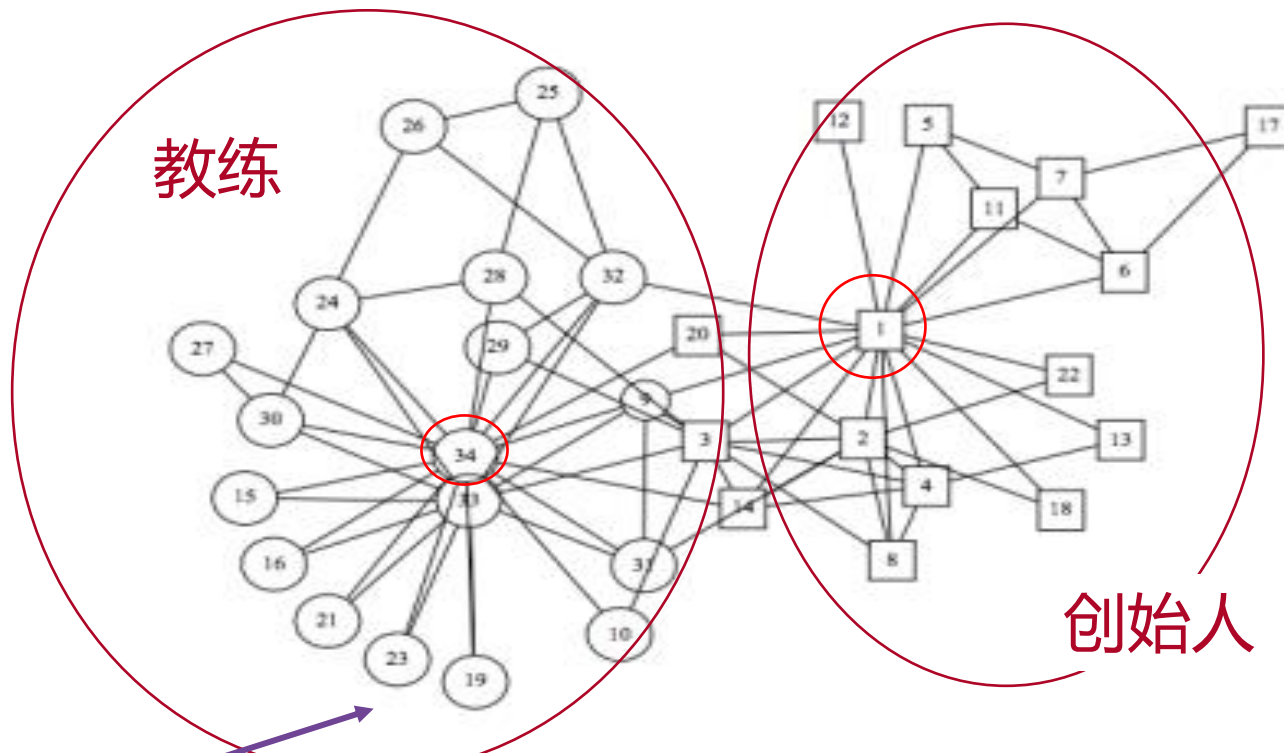
□ 为什么使用图模型对数据建模？

■ 图提供了一种观察数据**结构**特征的视角

一个空手道俱乐部中34个成员之间朋友关系形成的图  
你能发现什么特点？

最终这个俱乐分裂成两个对立的空手道俱乐部

结构平衡理论



<https://petterhol.me/2018/01/28/zacharys-zachary-karate-club/>

# 图数据入门

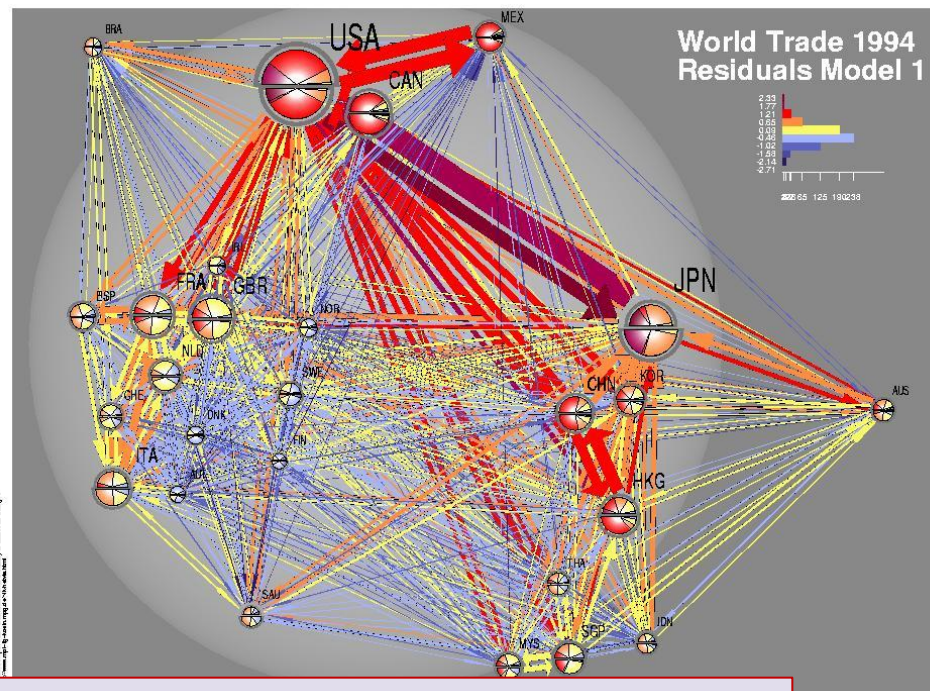
## □为什么使用图模型对数据建模？

- 图提供了一种观察数据结构特征的视角
- 图提供了一种理解**个体行为**的分析工具

国家之间贸易网络，你能发现什么结构特点？

节点大小：贸易总额  
边的粗细：所连接两个国家之间的贸易额

反映了参与贸易的机会与限制



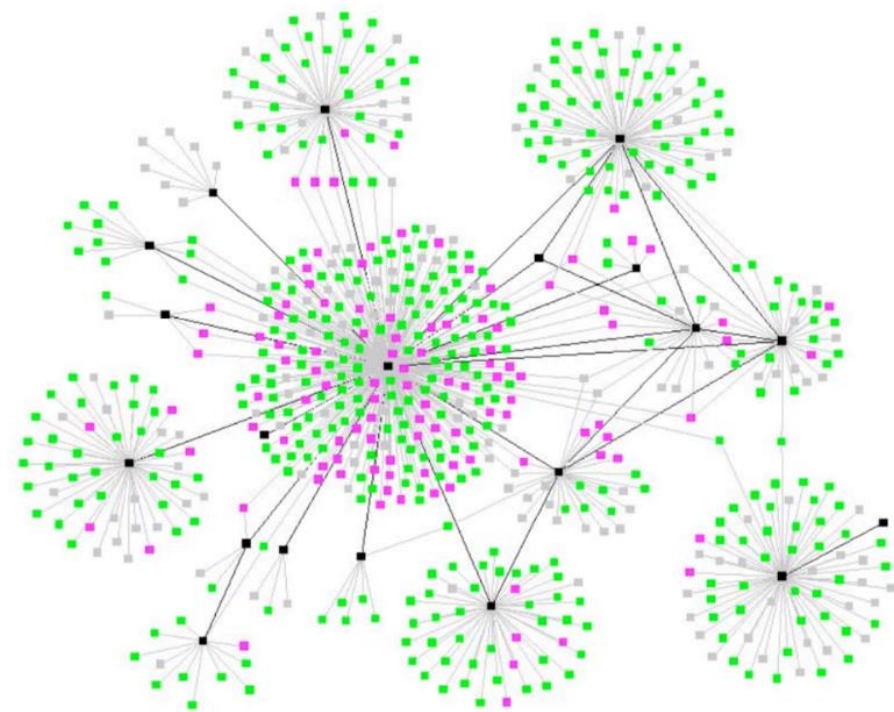


# 图数据入门

## □为什么使用图模型对数据建模？

- 图提供了一种观察数据结构特征的视角
- 图提供了一种理解个体行为的分析工具
- 图提供了一种解释**信息传播**的直观方法

一次**肺结核**爆发的**扩散**过程，  
与信息传播很类似



[Am J Public Health](#). 2007 March; 97(3): 470–477.

doi: [10.2105/AJPH.2005.071936](https://doi.org/10.2105/AJPH.2005.071936)

PMCID: PMC1805030

PMID: [17018825](https://pubmed.ncbi.nlm.nih.gov/17018825/)

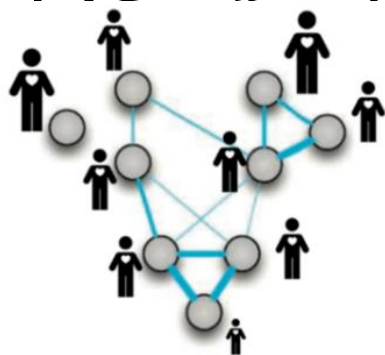
### Transmission Network Analysis to Complement Routine Tuberculosis Contact Investigations

[McKenzie Andre](#), MD, [Kashef Ijaz](#), MD, [Jon D. Tillinghast](#), MD, [Valdis E. Krebs](#), MLIR, [Lois A. Diem](#), BS,  
[Beverly Metchock](#), DrPH, [Theresa Crisp](#), MPH, and [Peter D. McElroy](#), PhD

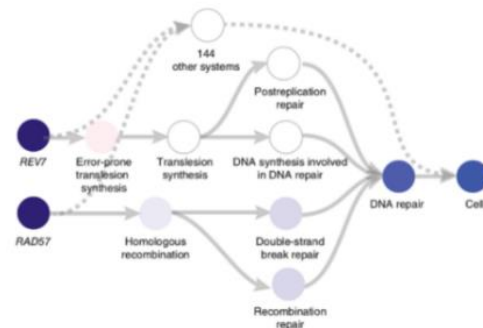


# 图数据入门

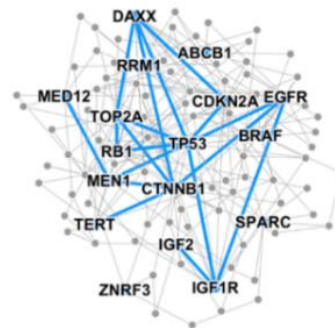
## □丰富多彩的图数据



Patient networks



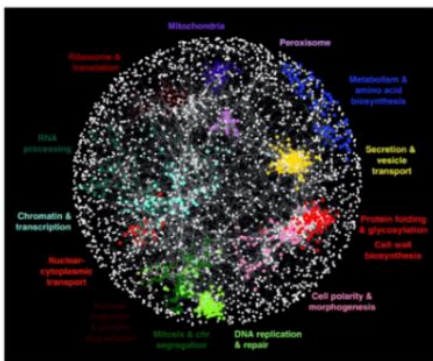
Hierarchies of cell systems



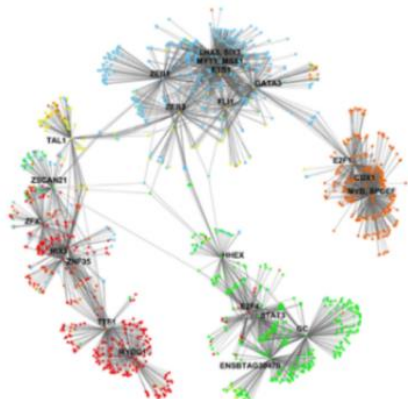
Disease pathways

pathways 英 ['pɑːθweɪz] 美 ['pæθweɪz]

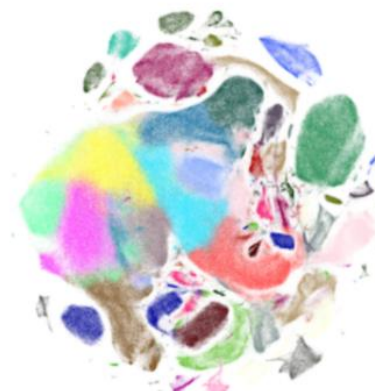
n. 小路; 路径; 途径; 行动路线;



Genetic interaction networks



Gene co-expression networks



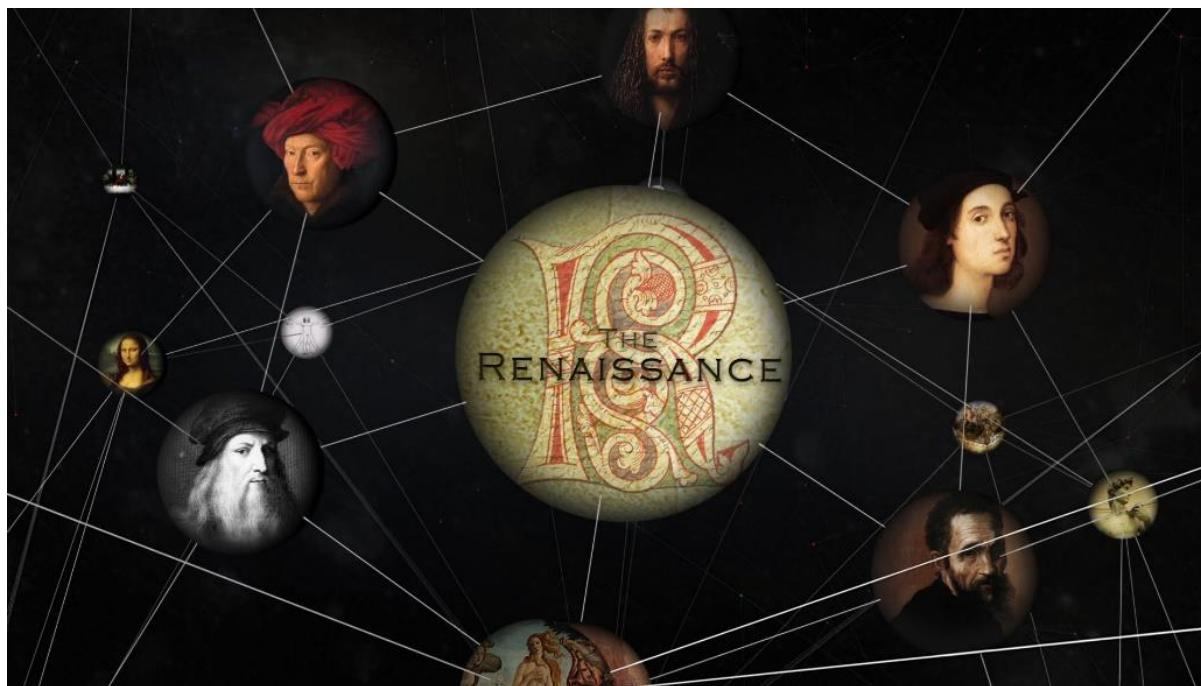
Cell-cell similarity networks



# 图数据入门

## □ 知识图谱

■ 知识图谱：语义关联、机器可读的知识表示技术



节点  
关系  
节点的属性  
关系的属性  
.....

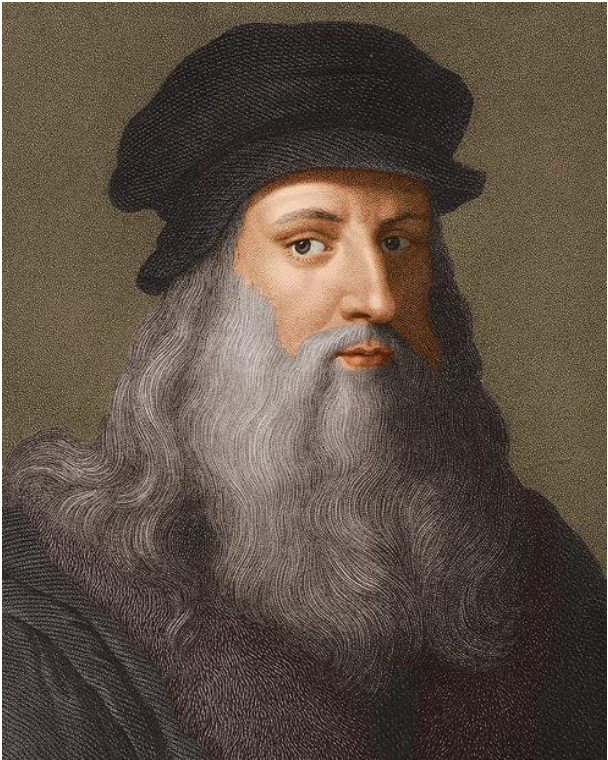
# 图数据入门

## □ 知识图谱



Mona\_Lisa

painted-by



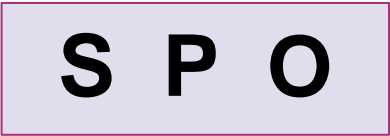
Leonardo\_da\_Vinci

IsA

Painting

IsA

Artist

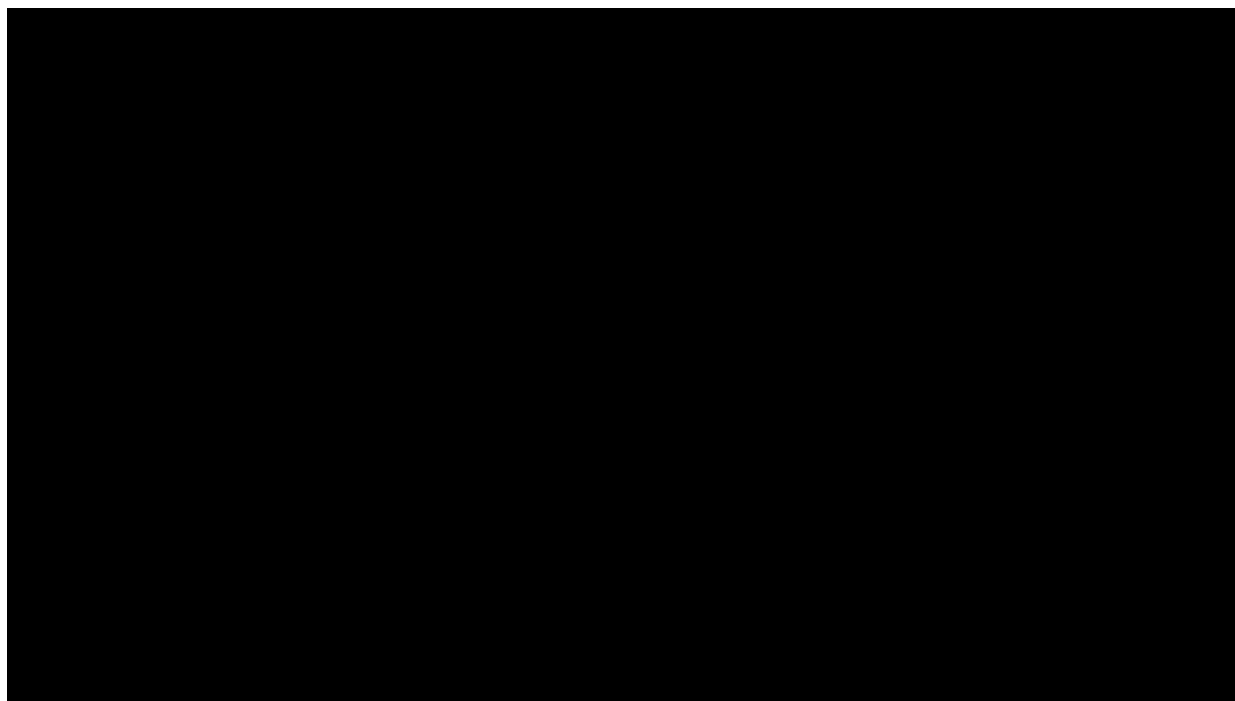


# 图数据入门

---

## □ 知识图谱的应用

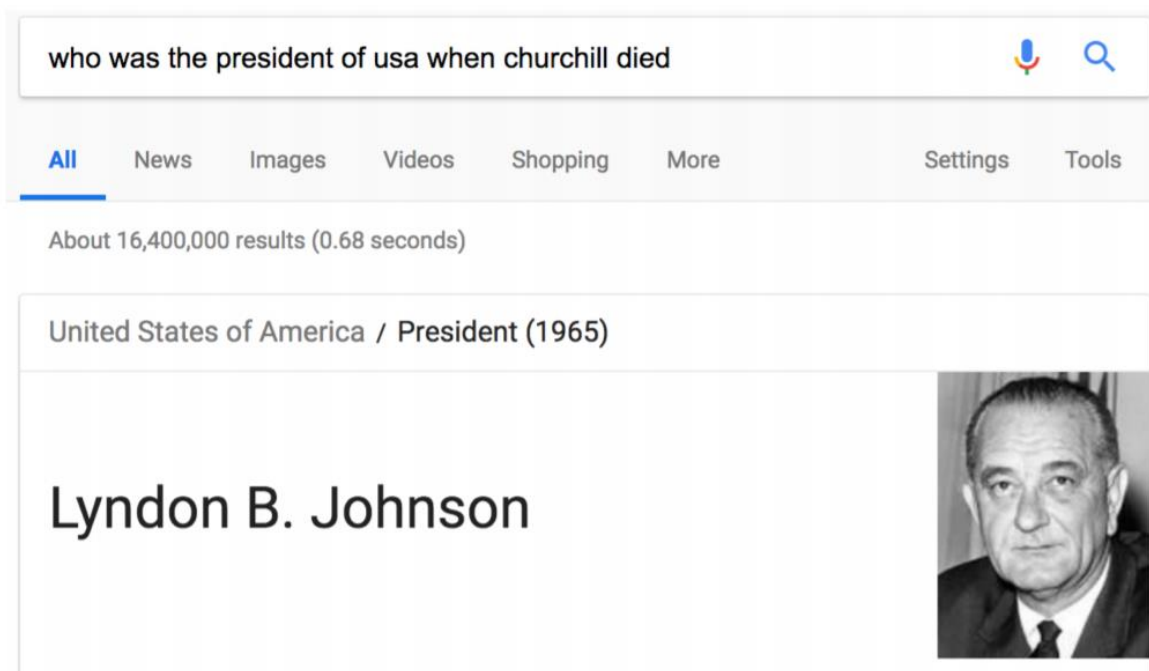
- 为AI系统提供领域知识



# 图数据入门

## □ 知识图谱的应用

### ■ 问答系统

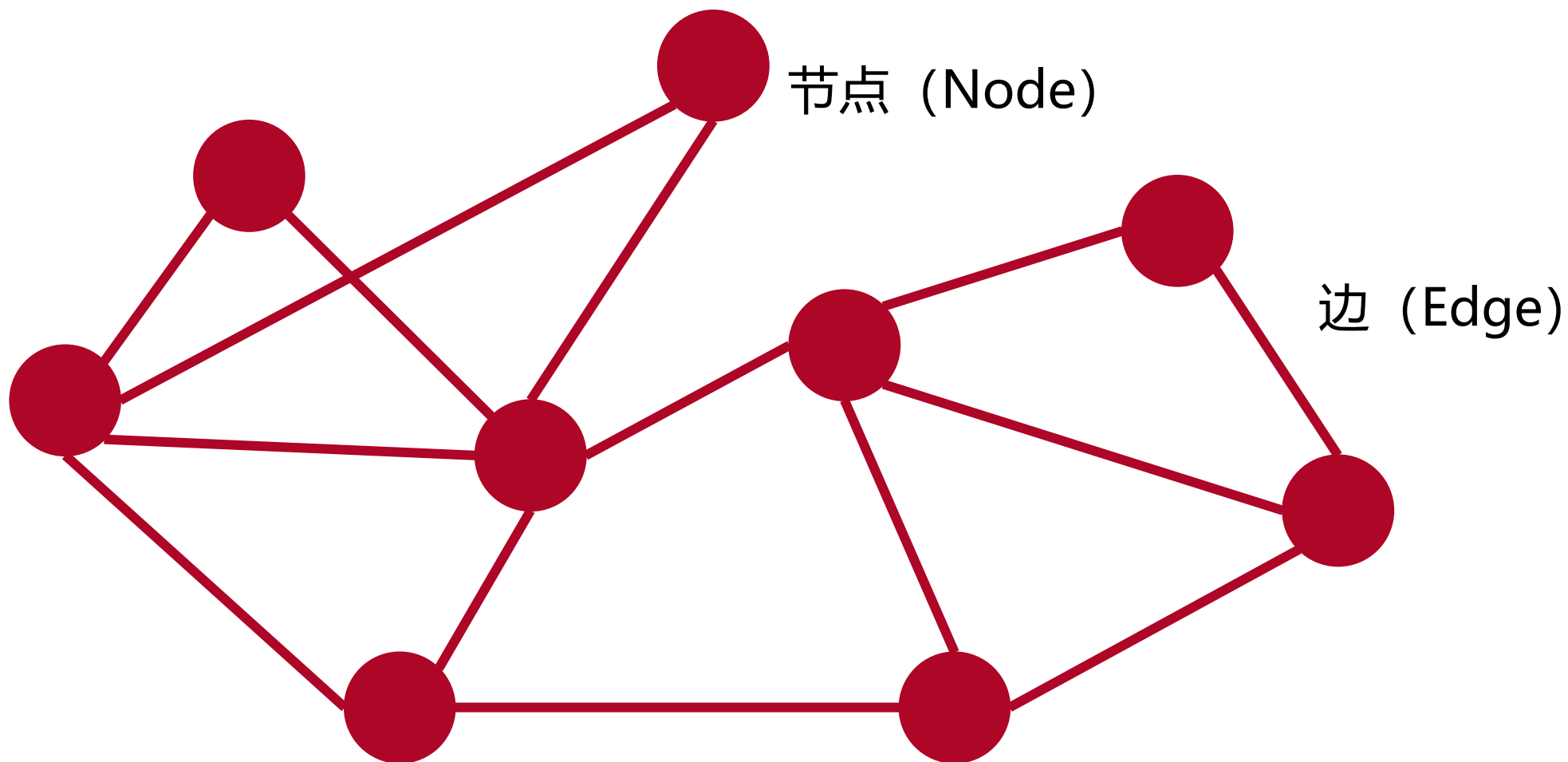


<http://www-bcf.usc.edu/~xiangren/NAACL18-KB-full.pdf>

# 图数据入门

## □图的形式化定义

$$G = (V, E)$$





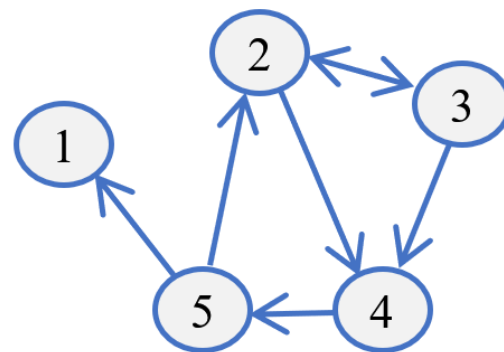
# 图数据入门

---

## □图的表示

## □可以使用三种方式进行表示(Representation)

- 邻接矩阵(Adjacency Matrix)
- 边列表(Edge List)
- 邻接关系列表(Adjacency List)



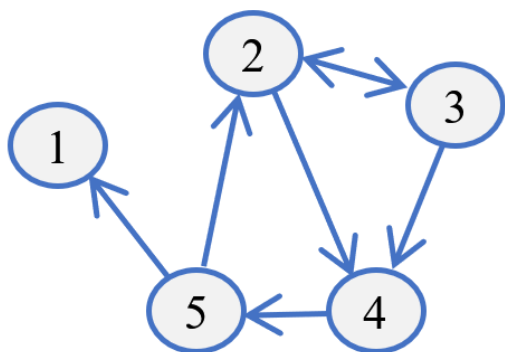
一个简单的有向图

# 图数据入门

## □图的表示

## □可以使用三种方式进行表示(Representation)

- 邻接矩阵(Adjacency Matrix)
- 边列表(Edge List)
- 邻接关系列表(Adjacency List)



一个简单的有向图

“邻接矩阵”表示

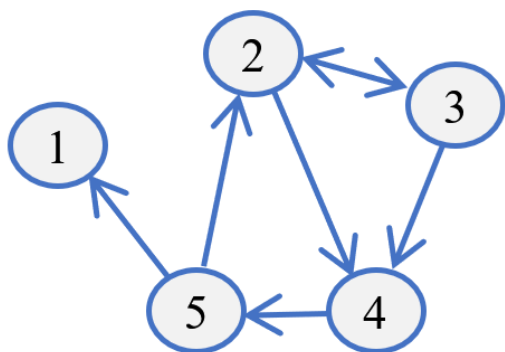
$$\text{ADJ} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

# 图数据入门

## □图的表示

## □可以使用三种方式进行表示(Representation)

- 邻接矩阵(Adjacency Matrix)
- 边列表(Edge List)
- 邻接关系列表(Adjacency List)



一个简单的有向图

“边列表”表示

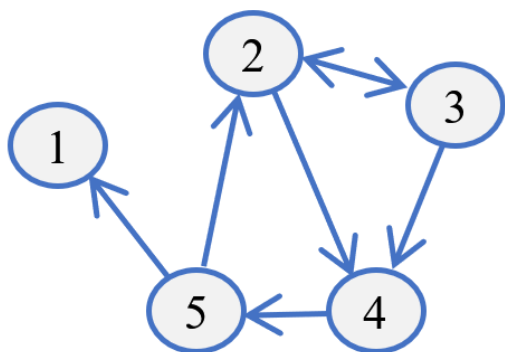
2,3
2,4
3,2
3,4
4,5
5,1
5,2

# 图数据入门

## □图的表示

## □可以使用三种方式进行表示(Representation)

- 邻接矩阵(Adjacency Matrix)
- 边列表(Edge List)
- 邻接关系列表(Adjacency List)



一个简单的有向图

“邻接关系列表”表示

```
1:  
2: 3 4  
3: 2 4  
4: 5  
5: 1 2
```

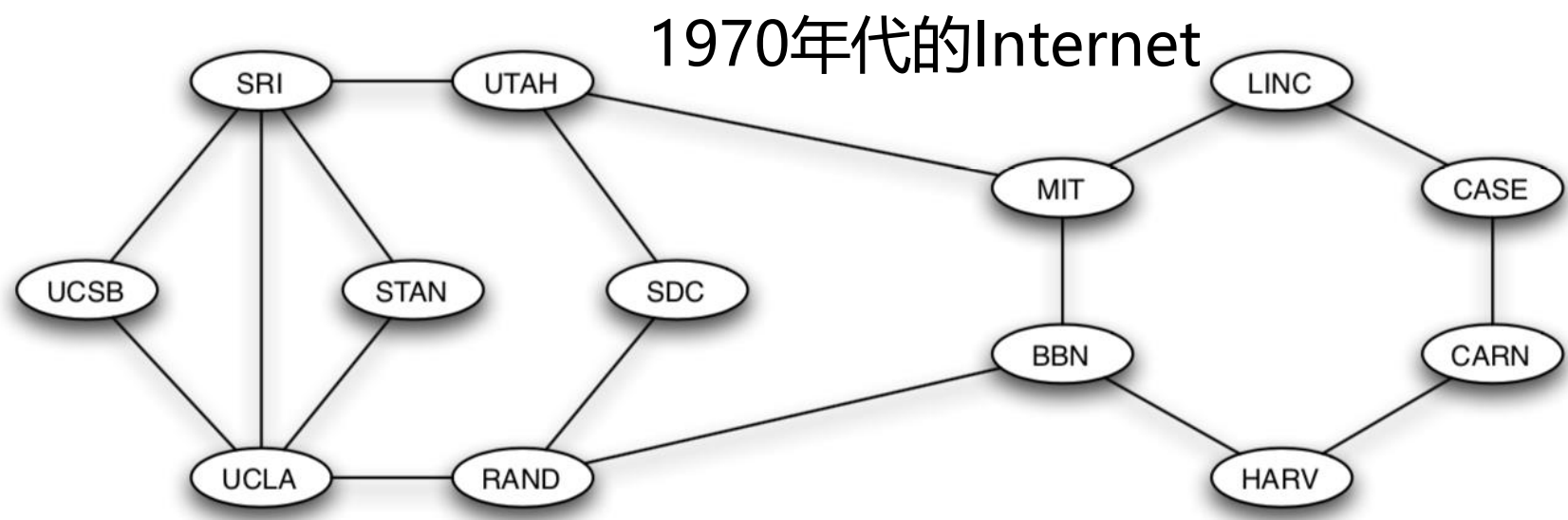




# 图数据入门

## 图的表示

练习：请写出下图的邻接矩阵、边列表与邻接关系列表（写出邻接矩阵的一行）

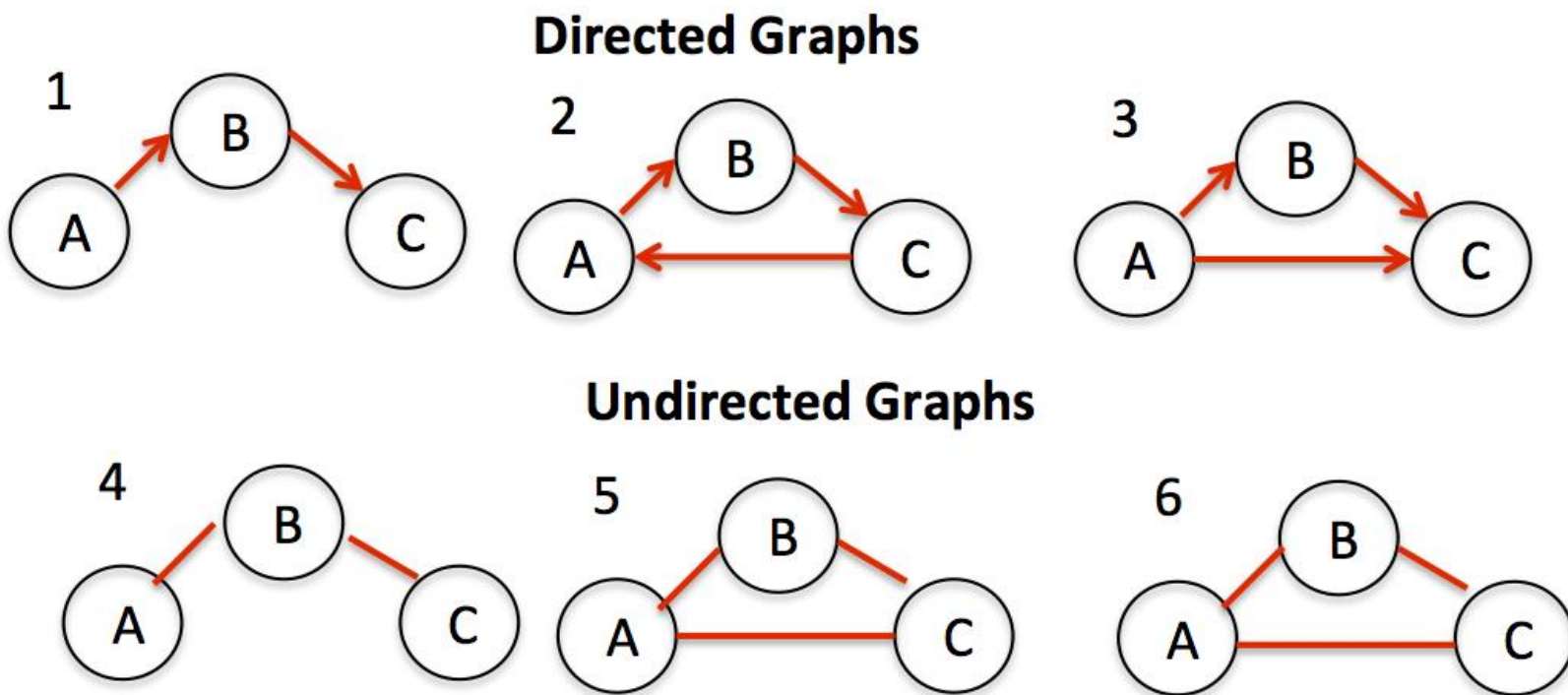


	UCSB	SRI	UCLA	STAN	UTAH	SDC	RAND	MIT	BBN	LINC	HARV	CASE	CARN
SDC	0	0	0	0	1	0	1	0	0	0	0	0	0

# 图数据入门

## □ 无向图与有向图 (Undirected Graph vs. Directed Graph)

■ 一条边两端的两个节点是否具有对称关系



微博

微信

# 图数据入门

□边的权重

□图上的每一条边 $e$ 关联一个数字 $w(e)$ ，用来表示边的**重要性或成本**

□例：道路网中的权重

节点：路口

边：道路

权重：**拥堵**情况



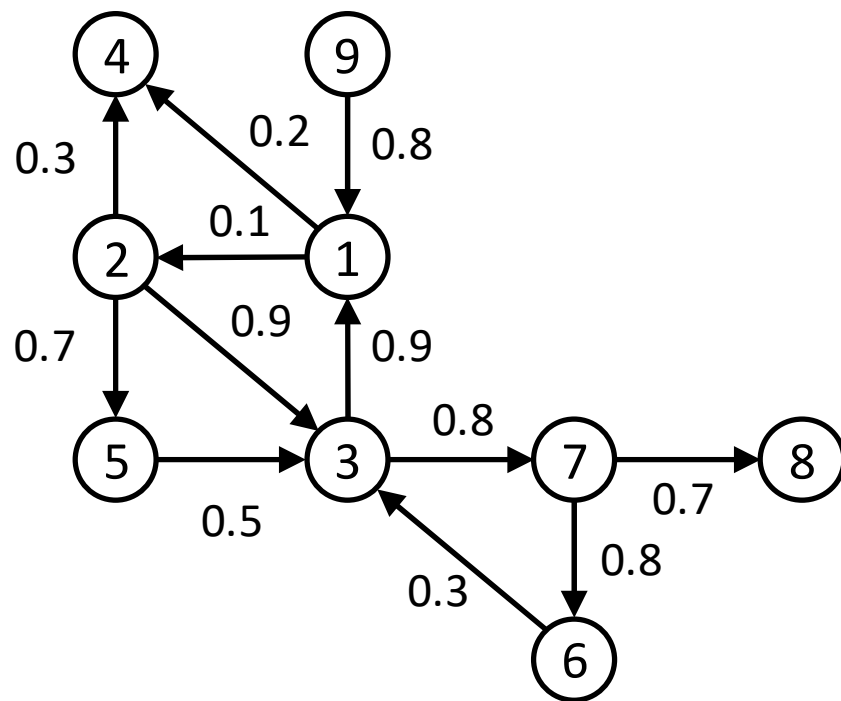
# 图数据入门

## □边的权重

□图上的每一条边 $e$ 关联一个数字 $w(e)$ ，用来表示边的**重要性或成本**

## □例：信息传播中的权重

- Independent Cascade (IC) 模型
- 假设用户 $u$ 到 $v$ 之间有一个概率值，称为影响概率
- 概率值越大， $v$ 相信 $u$ 传播消息的可能性就越大
- 概率值可以通过 $u$ 和 $v$ 的互动历史学习得到



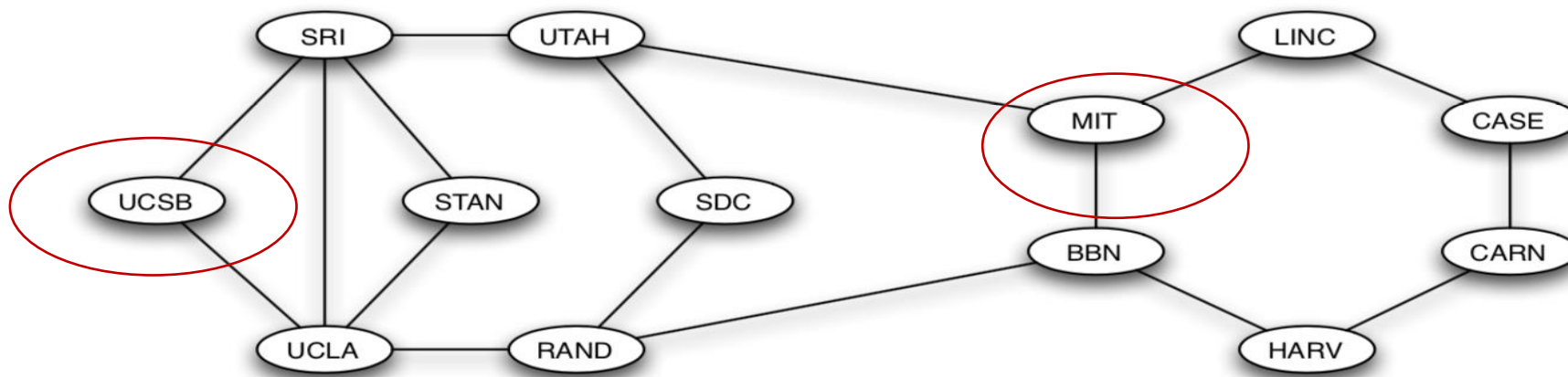
# 图数据入门

□ 路径：图上节点的序列，序列中任意两个相邻的节点都有边相连

■ Path  $p = (v_0, v_1, \dots, v_m)$  where any  $(v_i, v_{i+1}) \in E$

■ 简单路径：不包含重复节点的路径

■ 环：起点与终点相同的路径  $p = (v_0, v_1, \dots, v_m, v_0), (v_i, v_{i+1}) \in E$



- 请写出UCSB到MIT的路径
- 该图是否存在环？



# 图数据入门

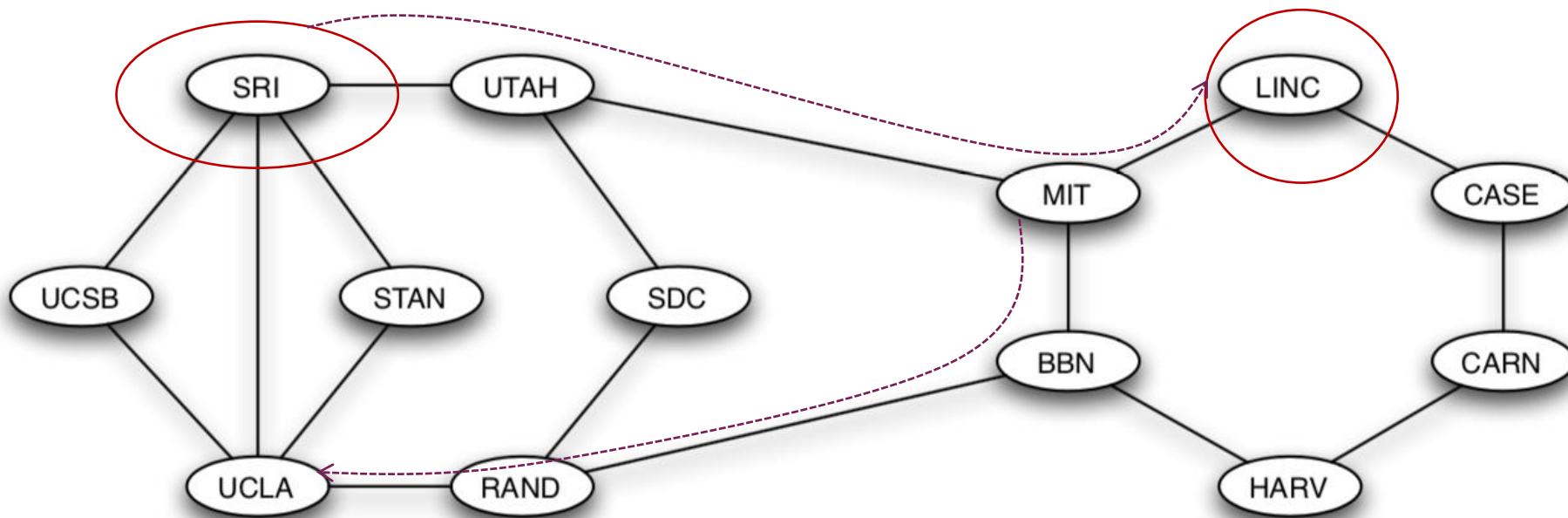
## □最短路径与距离

## □定义某条路径 $p$ 的长度为它所包含边的个数

■例：路径MIT, BBN, RAND, UCLA的长度为3

## □定义图上两点的距离为它们之间最短路径的长度

■例：LINC与SRI之间的距离为3



# 图数据入门

---

## □最短路径与距离

□思考题：如果节点Z在节点X和Y所有的最短路径上，则称Z为X和Y的**关键节点（Pivot node）**，其中Z与X和Y均不重合

■**请构造一个图**：每个节点均为至少一对节点的关键节点

■**请构造一个图**：该图中至少包含四个节点，并存在一个节点X，它是图中所有节点对的关键节点（不包含X）

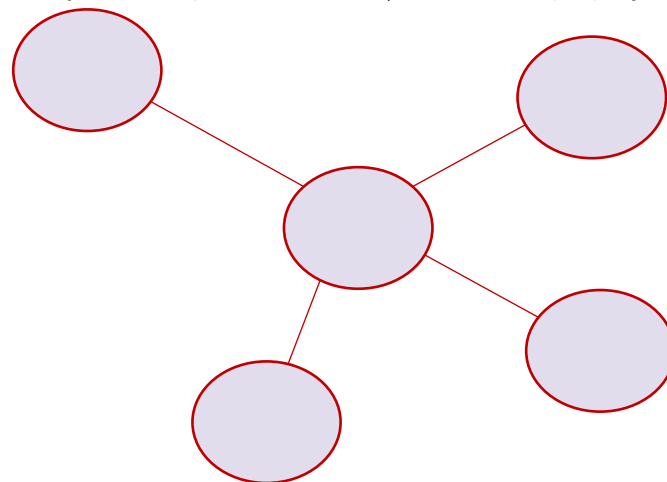
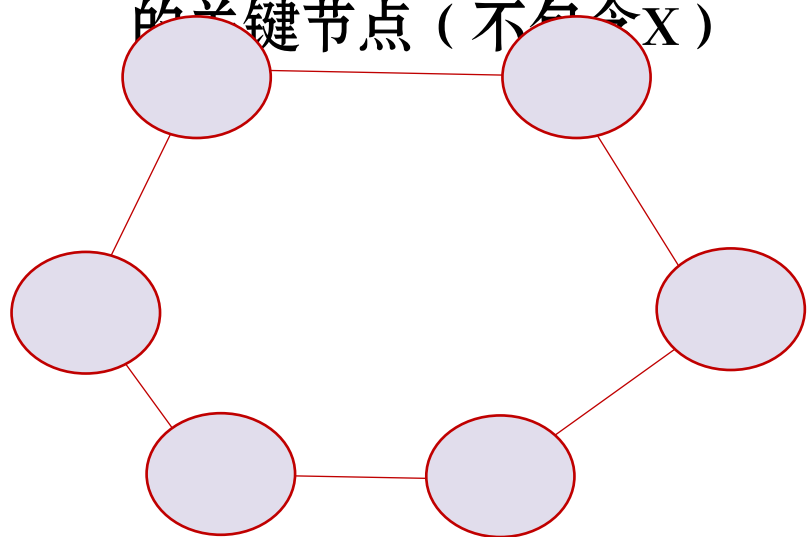
# 图数据入门

□图的基本概念：最短路径与距离

□思考题：如果节点Z在节点X和Y所有的最短路径上，则称Z为X和Y的**关键节点（Pivot node）**，其中Z与X和Y均不重合

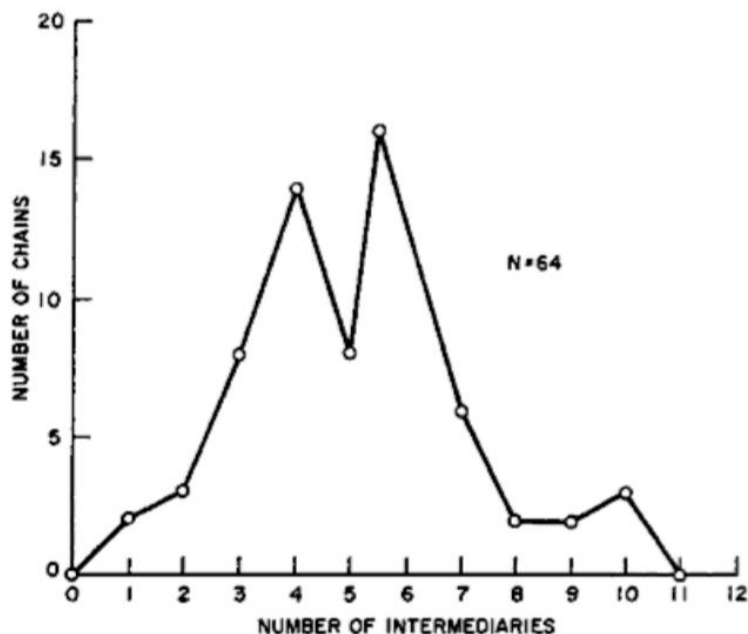
■请构造一个图：每个节点均为至少一对节点的关键节点

■请构造一个图：该图中至少包含四个节点，并存在一个节点X，它是图中所有节点对的关键节点（不包含X）



# 图数据入门

- 图的基本概念：六度分隔现象
- 在这个世界上，任意两个人之间，**只隔着六个人**
  - 六度分隔



小世界现象（又称小世界效应），也称六度分隔理论（英文：xxx）

# 图数据入门

## □图的基本概念：小世界现象



<https://www.cs.uic.edu/~cornelia/kdsin16/lectures/sna1.pdf>



# 图数据入门

---

## □图的基本概念：三元闭包（Triadic Closure）

- 如果两个人在网络中有共同的好友，他们成为好友的几率也会提升

## □几点原因：如果B和C都有共同的好友A，那么

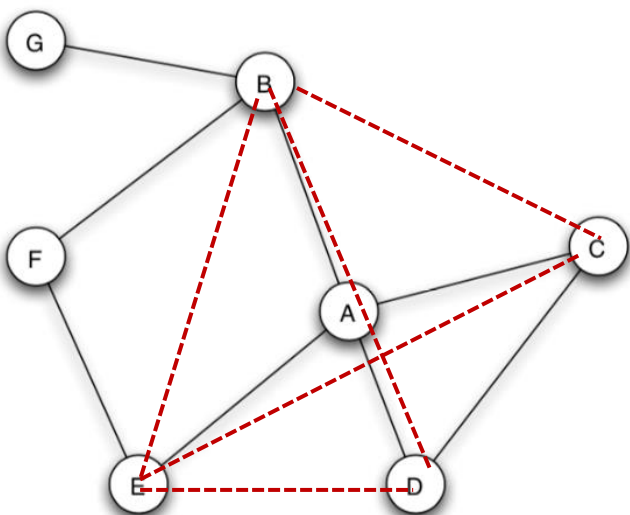
- B更有**可能遇到**C——因为他们都与A有交集
- B和C更有**可能互相信任**——因为他们有共同的好友
- A更有**可能介绍**B和C认识

# 图数据入门

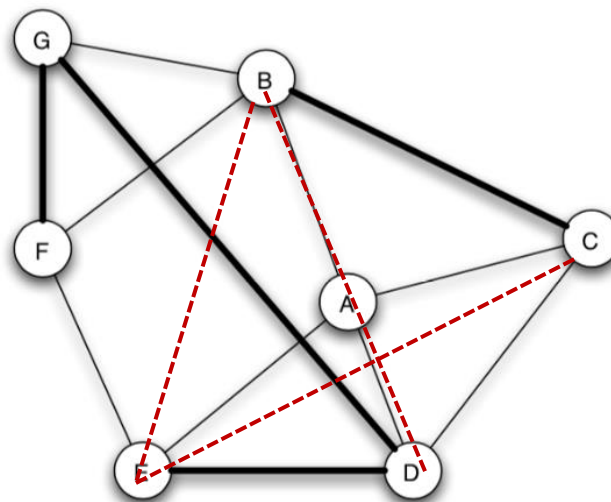
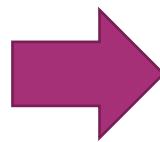
□图的基本概念：三元闭包（Triadic Closure）

□量化指标：clustering coefficient

■节点A的clustering coefficient：A的任意两个邻居是好友的概率



clustering coefficient of A =  $1/6$



clustering coefficient of A =  $1/2$

[Bearman and Moody]青少年女孩调查：clustering coefficient越低越易自杀

# 图数据入门

---

## □ Graph模块的基本内容

## □ 基本知识点：

- Centrality：图里的哪些节点更重要？
- Community：图是否能够划分为不同的社区
- Influence：信息如何在图上传播，如何度量人与人之间的影响力
- Query：如果利用图回答一些基本的问题

## □ 基本技能：

- Statistical Thinking：统计思维
- Optimization Thinking：优化思维
- System Thinking：系统思维

# 中心度

□ 节点中心度（Node Centrality）分析

□ 在网络中，不同节点的“地位”是不平等的

■ 例子：美国高中生恋爱关系图

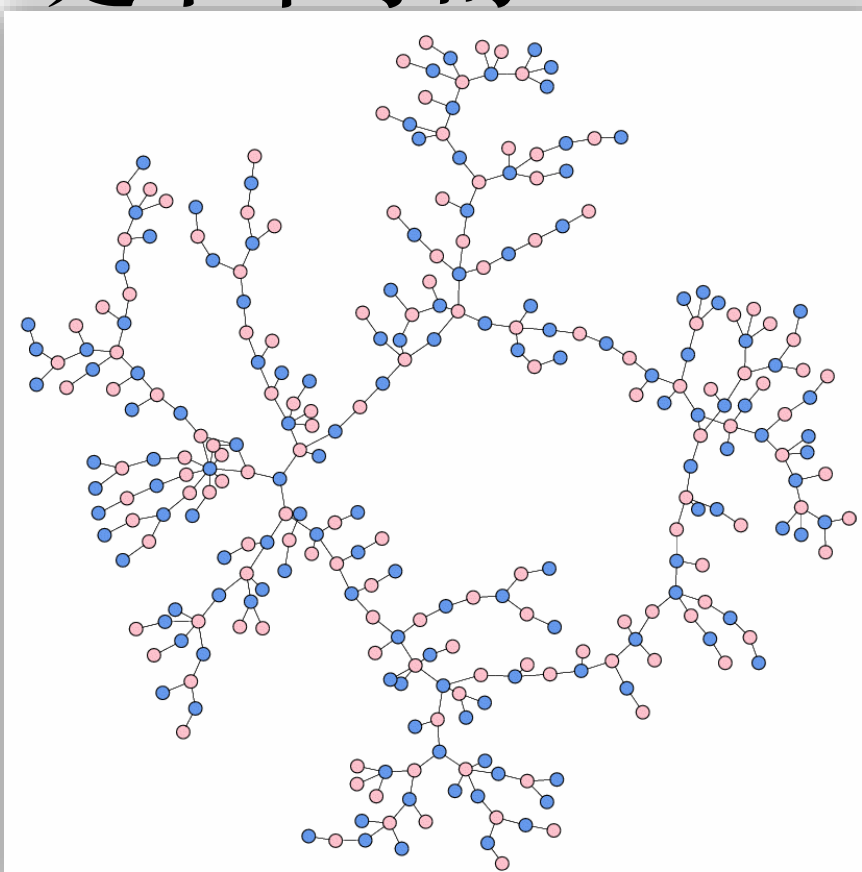
■ 边表示18个月内谈过恋爱



无向图

• 思考：

- 你觉得哪些节点更重要？
- 你怎么解释这种重要性？



# 中心度

□ 节点中心度（Node Centrality）分析

□ 在网络中，不同节点的“地位”是不平等的

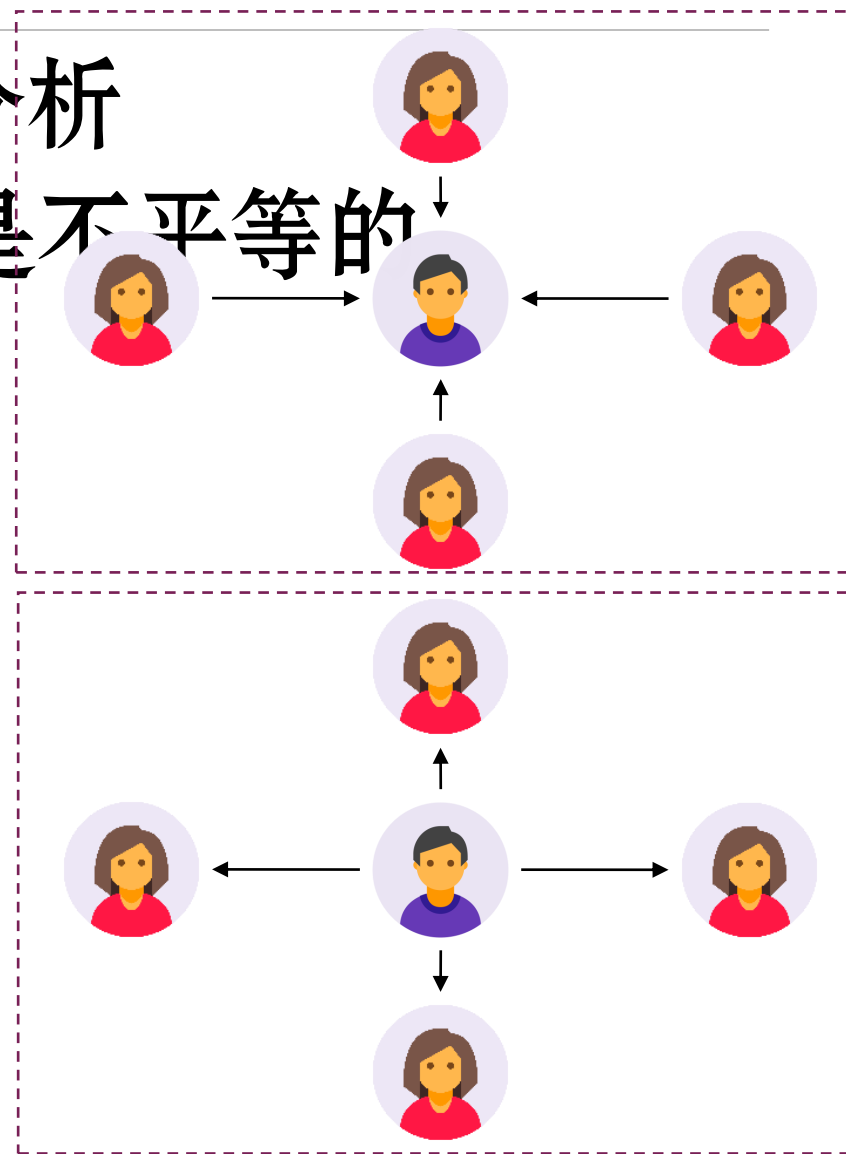
■ 例子：美国高中生恋爱关系图

■ 如果定义有向边：“追求”关系



• 思考：

- 右边两图中男生的重要性一样吗？
- 你怎么解释这种重要性？



# 中心度

## □ 节点中心度（Node Centrality）

■ 给定一个图，哪些节点更重要或更有影响力？

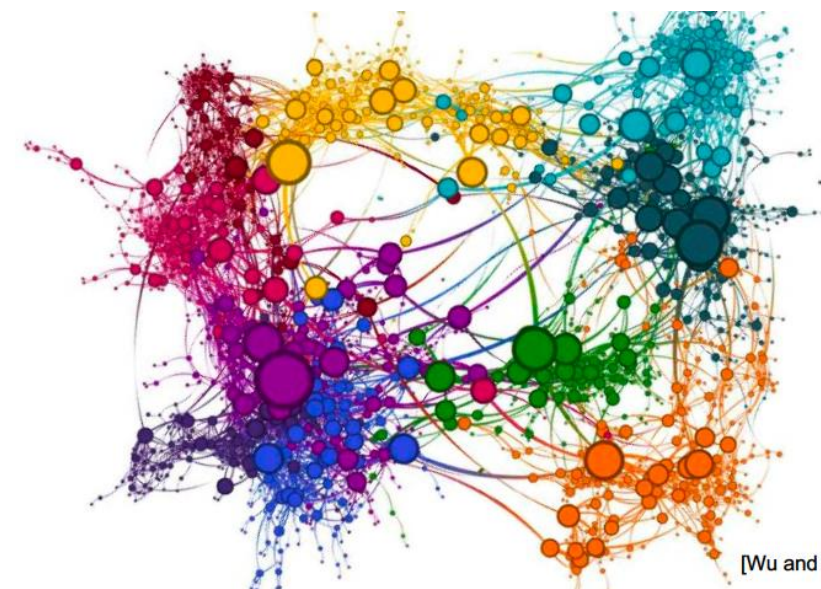
## □ 为什么要研究节点中心度（Node Centrality）？

■ 在社交网络中，每个人的影响力如何？

■ 在道路网络中，有哪些“关键节点”？

■ 在Web网络中，哪些网页更加重要？

■ 你能想到其它吗？





# 中心度

---

## □节点中心度（ Node Centrality ）

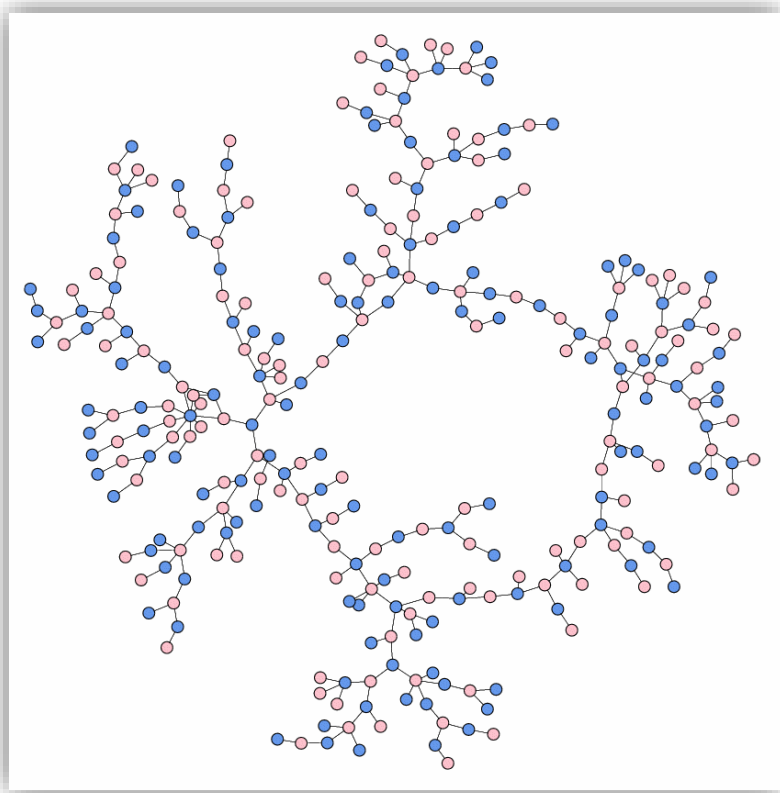
- 1.基于几何图形的度量方法
- 2.基于路径的度量方法
- 3.PageRank算法

# 中心度

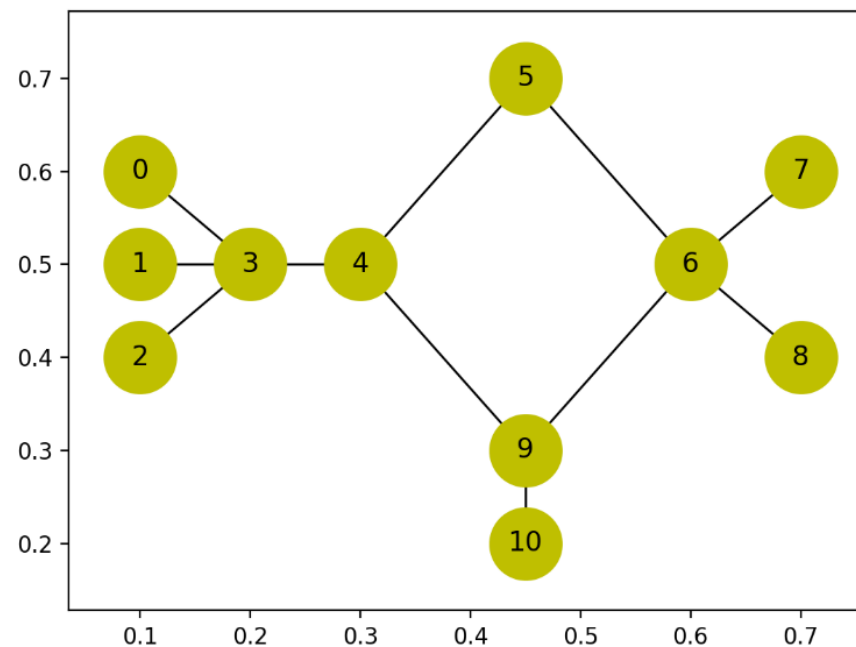
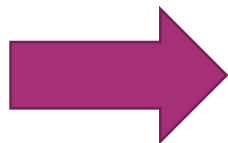
□ 我们考虑一个小例子

□ 演示示例

■ 简化版的美国高中生恋爱关系图



简化



# 中心度

## □基于几何图形的度量方法

## □基本思想

■节点 $v$ 的Centrality是该节点到其它节点的距离的函数

## □(In-)Degree Centrality

■节点 $v$ 的Centrality取决于它的度

• (如果是有向图, 则为入度)

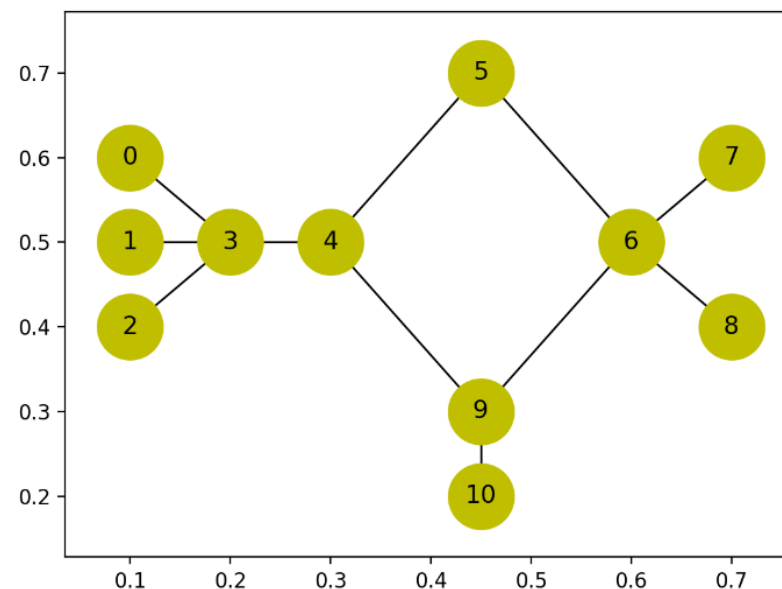
■即到节点 $v$ 距离为1的节点的个数

$$C_D(v) = \frac{\deg(v)}{n-1}$$

## □请计算右图中节点的Degree Centrality

{0: 0.1, 1: 0.1, 2: 0.1, 3: 0.4, 4: 0.3, 5: 0.2, 6: 0.4, 7: 0.1, 8: 0.1, 9: 0.3, 10: 0.1}

课堂练习, 计算一下节点4的Degree中心度

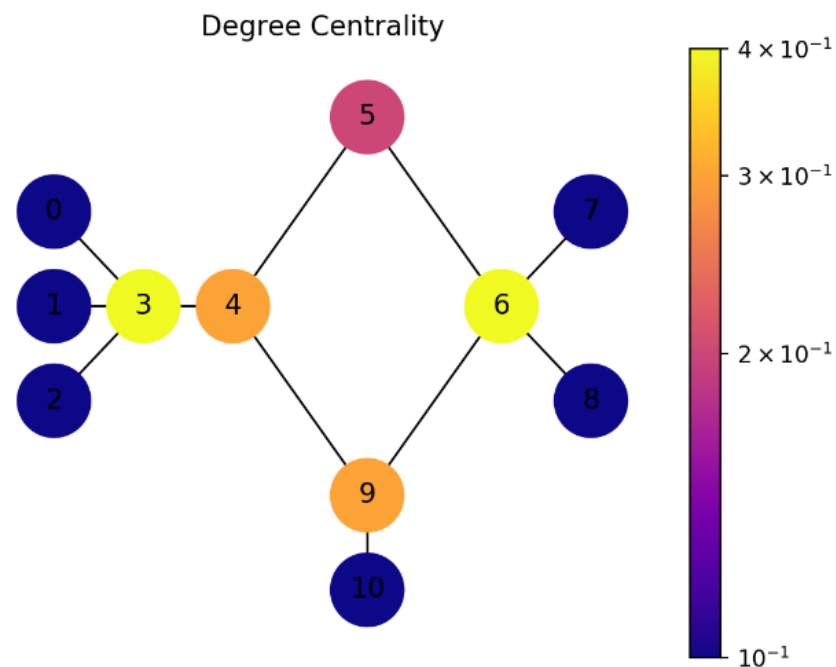


问题: 公式中为什么除以 $n-1$ ? 某个节点最多与 $n-1$ 个其它节点有关系

# 中心度

## □如何解释Degree Centrality?

- 恋爱网络：哥的情史很丰富 😏
- 微博网络：大V



# 中心度

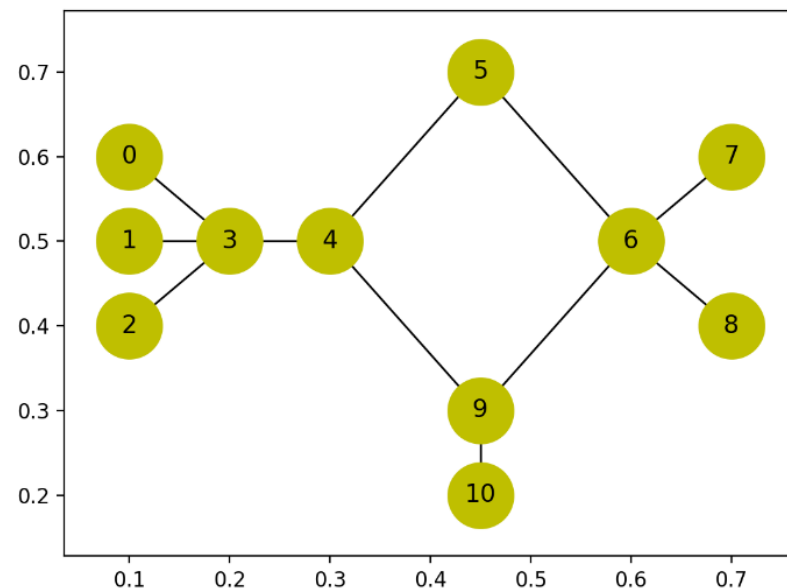
## □ Closeness 中心度

- 如果一个顶点到其他顶点的距离都比较短，那么它处于网络的中心
- 节点  $v$  的 Centrality 取决于其它节点到它的距离是否接近
- 其它节点到  $v$  的距离越接近， $v$  越重要

$$C_C(v) = \frac{n - 1}{\sum_{u \in V - \{v\}} d(u, v)}$$

$d(u, v)$  为  $u$  到  $v$  的最短距离

Closeness Centrality {0: 0.32, 1: 0.32, 2: 0.32, 3: 0.45, 4: 0.53, 5: 0.45, 6: 0.43, 7: 0.31, 8: 0.31, 9: 0.5, 10: 0.34}



# 图数据入门、中心度

## □ Closeness 中心度

■ 如果一个顶点到其他顶点的距离都比较短，那么该顶点就是网络的中心

■ 节点  $v$  的 Centrality 取决于其它节点到它的距离是否接近

■ 其它节点到  $v$  的距离越接近， $v$  越重要

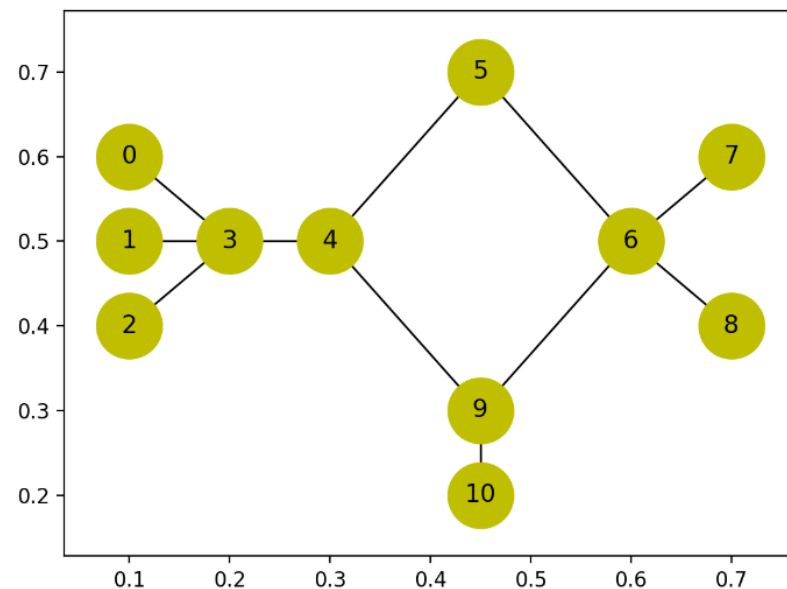
$$C_C(v) = \frac{n-1}{\sum_{u \in V-\{v\}} d(u, v)}$$

$d(u, v)$  为  $u$  到  $v$  的最短距离

Closeness Centrality {0: 0.32, 1: 0.32, 2: 0.32, 3: 0.45, 4: 0.53, 5: 0.45, 6: 0.43, 7: 0.31, 8: 0.31, 9: 0.5, 10: 0.34}

课堂练习，计算一下节点4的 Closeness 中心度

$$(11-1)/(2+2+2+1+1+1+2+2+3+3) = 10/19 = 0.5263$$

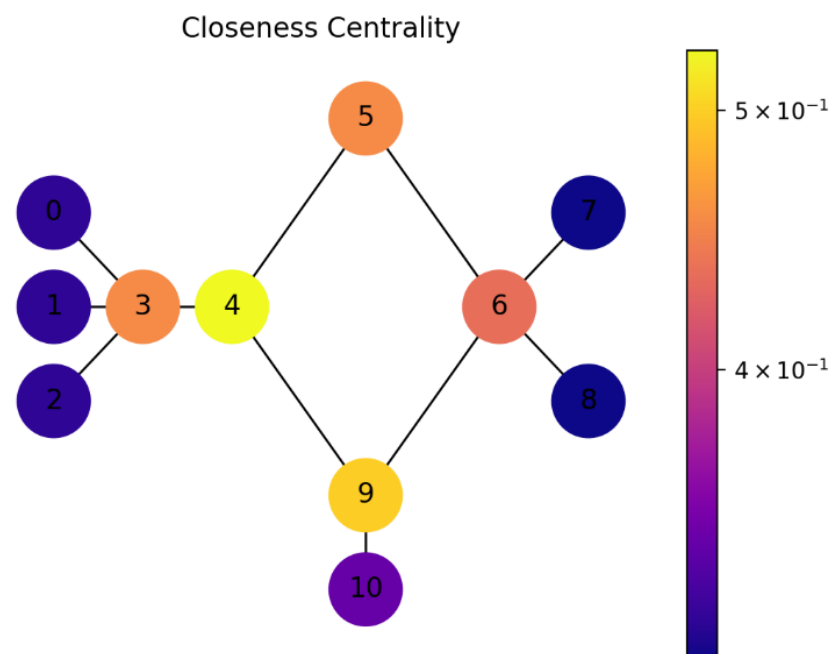




# 图数据入门、中心度

## □如何解释Closeness Centrality?

- 恋爱网络：哥不是她的男朋友，就是他女友的前男友 😞
- 道路网络：中心地标建筑



# 图数据入门、中心度

## □ Betweenness Centrality

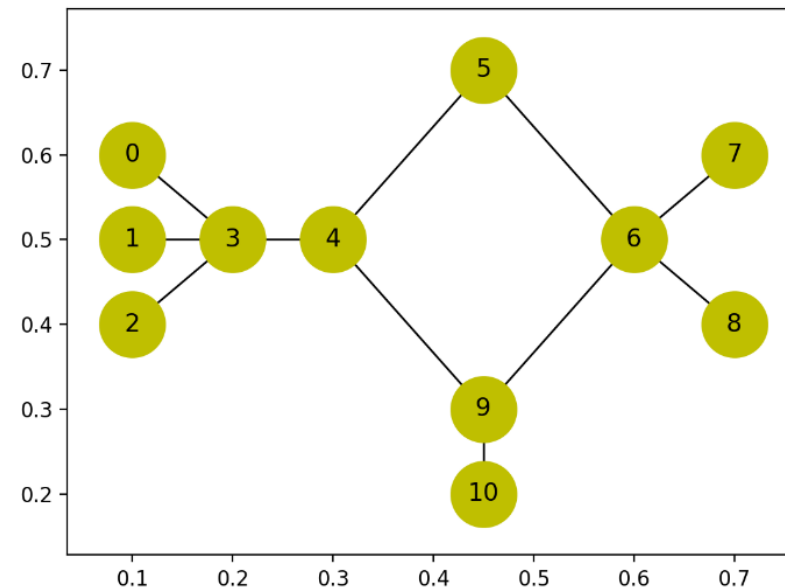
- 节点  $v$  的 Centrality 取决于它是否经常出现在其它节点的最短路径上
- $v$  出现在其它节点最短路径上次数越多,  $v$  越重要

$$C_C(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

$\sigma(s,t)$  为  $s$  到  $t$  的最短路径个数;  $\sigma(s,t|v)$   $s$  到  $t$  经过  $v$  的最短路径个数

Betweenness  
Centrality

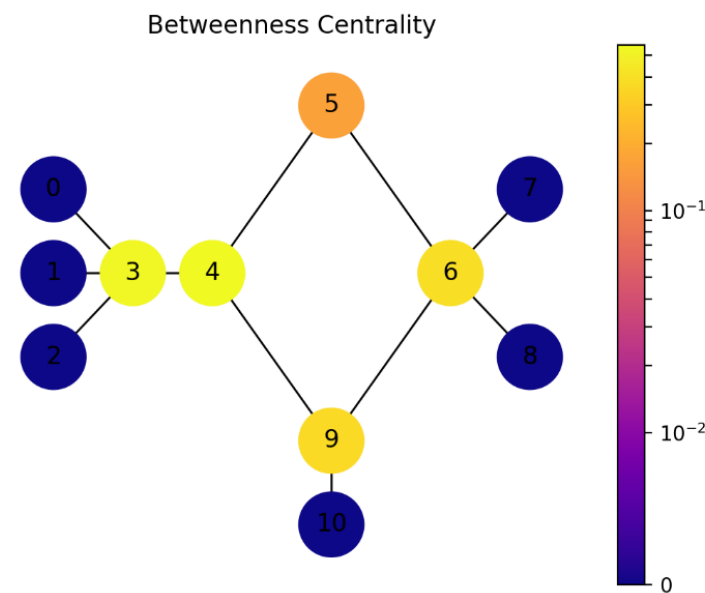
{0: 0.0, 1: 0.0, 2: 0.0, **3: 0.53**, **4: 0.56**, 5: 0.17, **6: 0.4**, 7: 0.0, 8: 0.0, 9: 0.37, 10: 0.0}



# 图数据入门、中心度

## □如何解释Betweenness Centrality?

- 恋爱网络：没有哥，那些妹子们这辈子也不会有什么关联 😞
- 贸易网络：贸易枢纽hubs

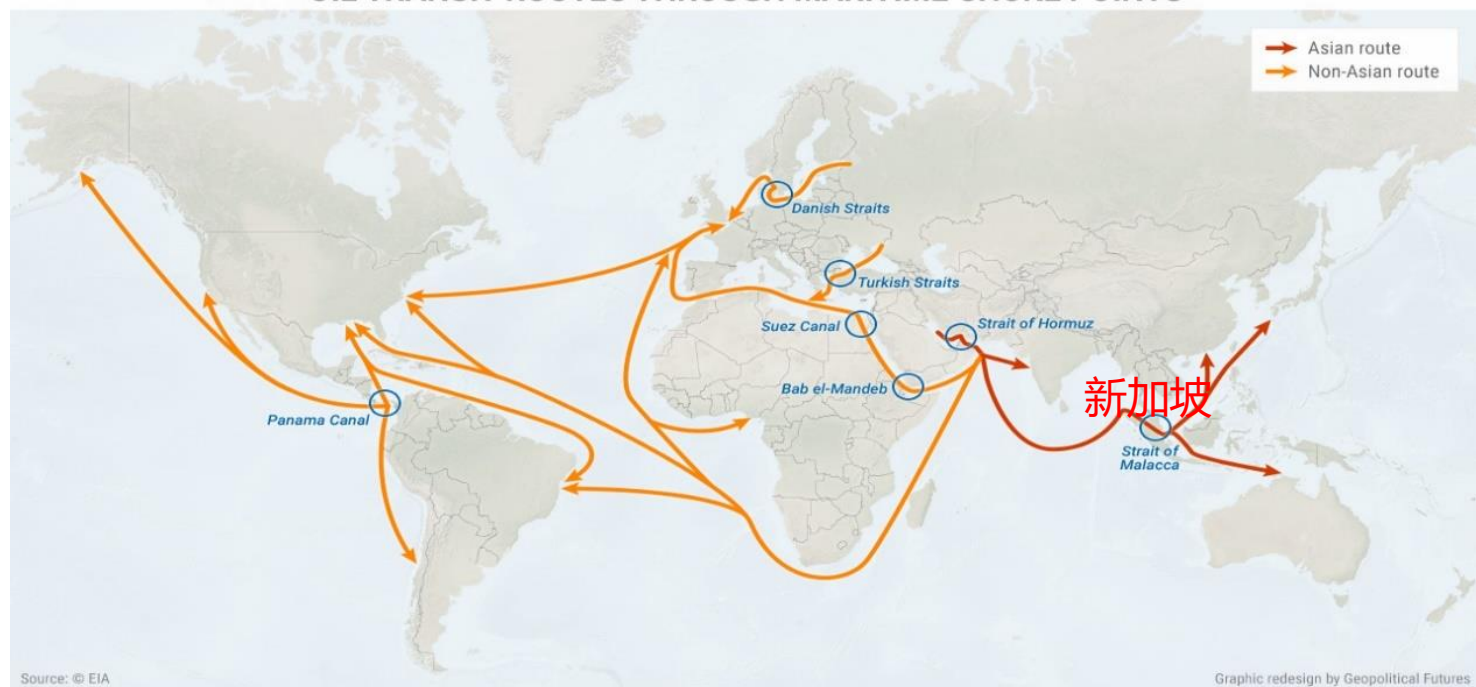


# 图数据入门、中心度

□ Betweenness 中心度

□ 思考：新加坡为什么成为全球贸易港？

OIL TRANSIT ROUTES THROUGH MARITIME CHOKES



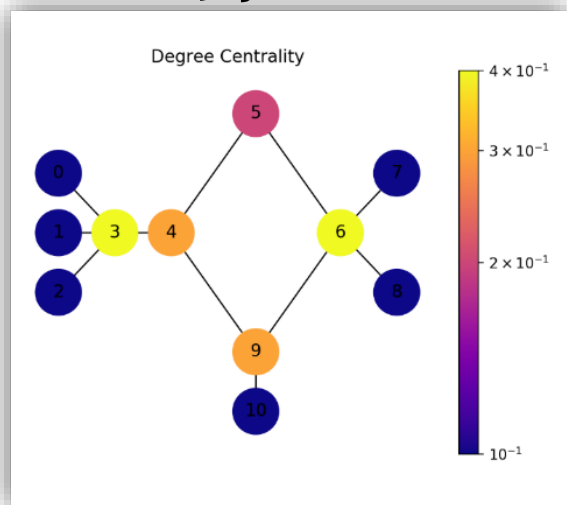
波斯湾石油运输航路图

The Persian Gulf is a leading oil-producing region, accounting for 30% of global supply. Meanwhile, **East Asia** is a major oil-consuming region and accounts for **85%** of the Persian Gulf's exports, according to the Energy Information Administration (EIA)

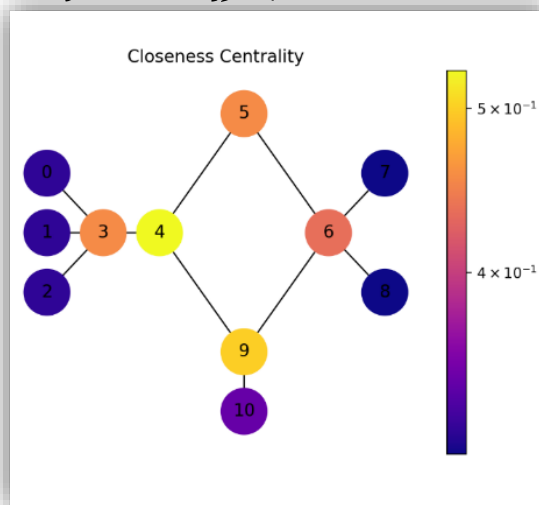
地缘政治与  
Betweenness Centrality

# 图数据入门、中心度

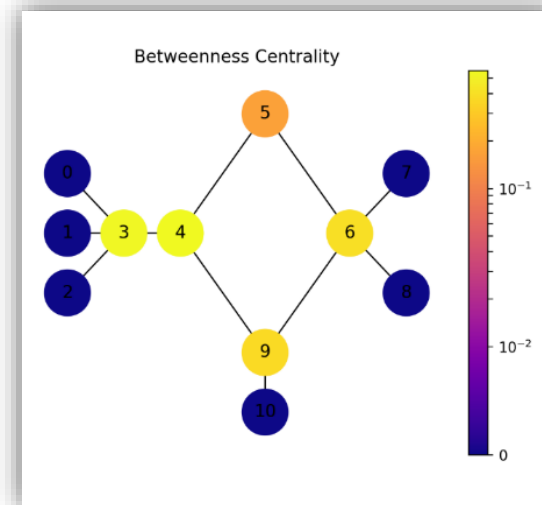
## □三种Centrality度量的比较



Degree Centrality  
3/6/4/9



Closeness Centrality  
4/9/5/3/6



Betweenness Centrality  
4/3/6/9

Degree

{0: 0.1, 1: 0.1, 2: 0.1, 3: 0.4, 4: 0.3, 5: 0.2, 6: 0.4, 7: 0.1, 8: 0.1, 9: 0.3, 10: 0.1}

Closeness

{0: 0.32, 1: 0.32, 2: 0.32, 3: 0.45, 4: 0.53, 5: 0.45, 6: 0.43, 7: 0.31, 8: 0.31, 9: 0.5, 10: 0.34}

Betweenness

{0: 0.0, 1: 0.0, 2: 0.0, 3: 0.53, 4: 0.56, 5: 0.17, 6: 0.4, 7: 0.0, 8: 0.0, 9: 0.37, 10: 0.0}

# 图数据入门

## □Python实例分析

```
$ pip install networkx
```

### NetworkX

### Software for complex networks

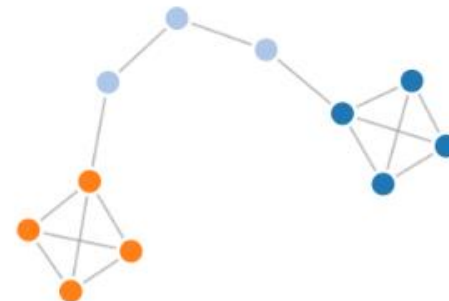
#### Stable (notes)

2.2 — September 2018  
[download](#) | [doc](#) | [pdf](#)

#### Latest (notes)

2.3 development  
[github](#) | [doc](#) | [pdf](#)

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



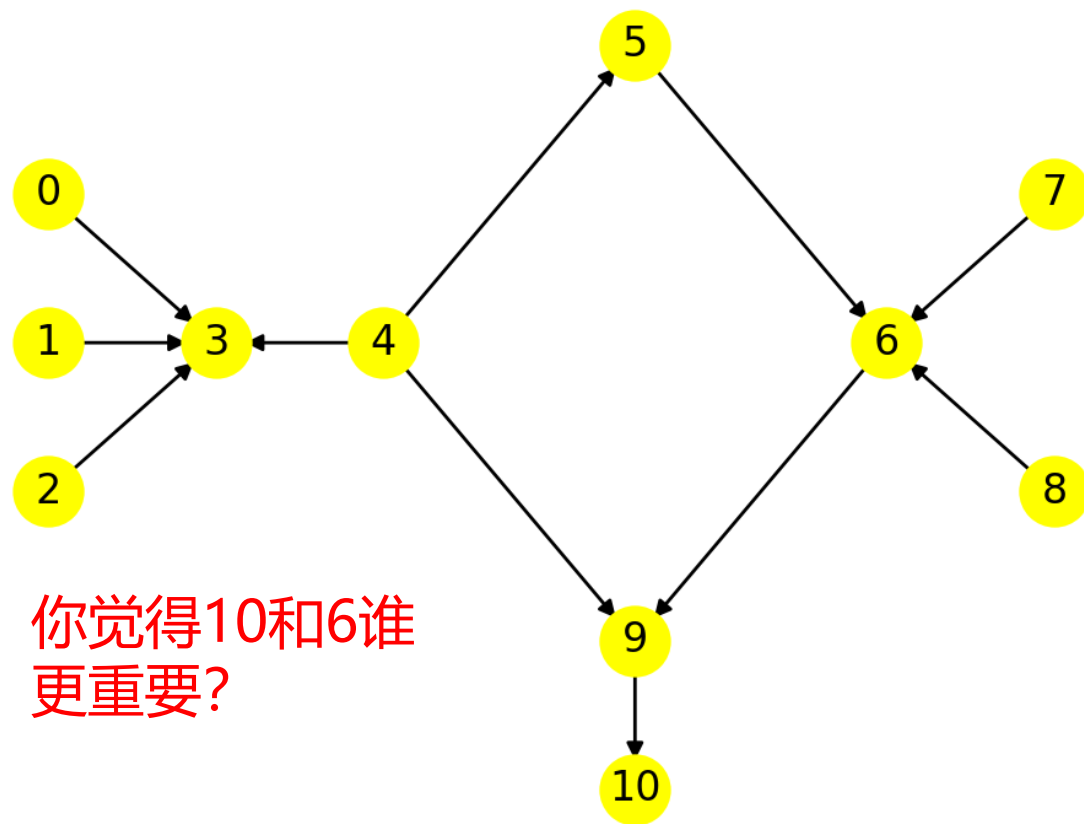


# Page rank

- 度量有向图节点的重要性
- 示例：简易版恋爱关系有向图
  - 定义有向边：“追求”关系



- 基于投票的思路
  - 将每个入边看作一次投票
  - 得到的票数越多，越重要



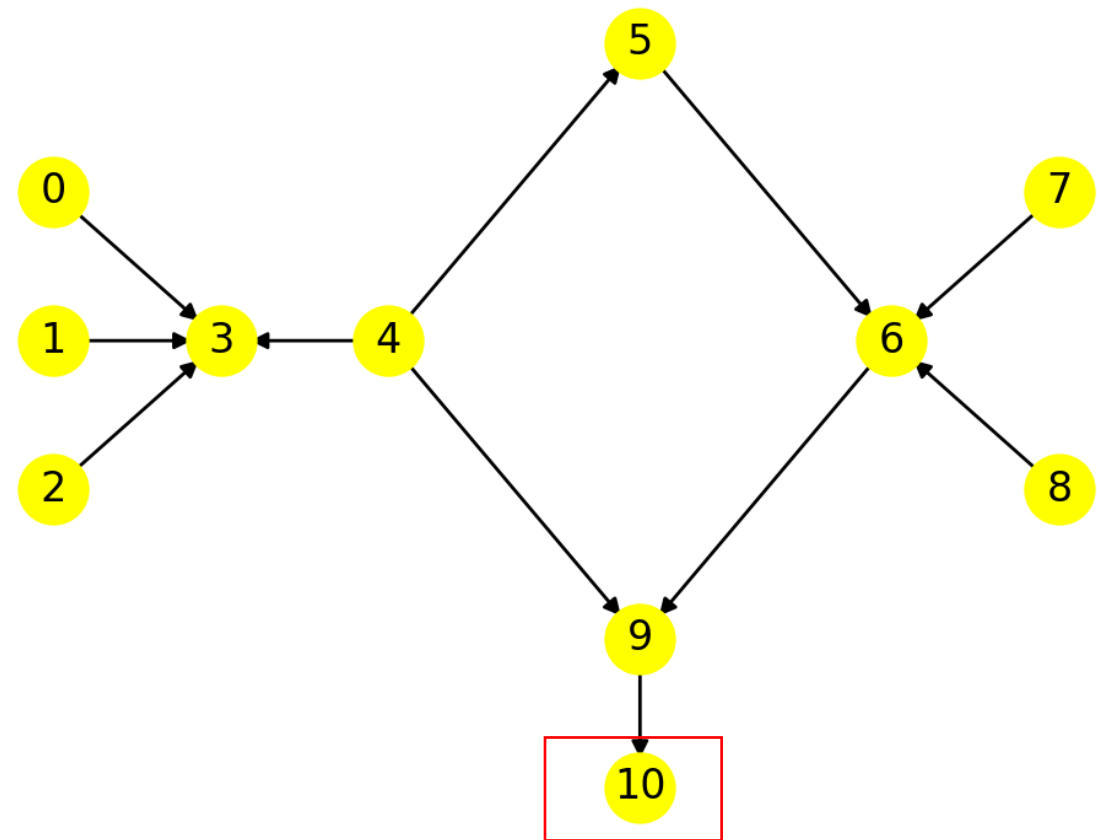
# Page rank

□PageRank的基本思想：给不同的人边赋上不同的权重

- 考虑某个节点 $v$
- 指向 $v$ 的节点的PageRank值越高，相应入边的权重越高
- 指向 $v$ 的节点指向其它节点的数目越多，分摊越多，对 $v$ 相应入边的权重越低



In-Degree Centrality不能表达!

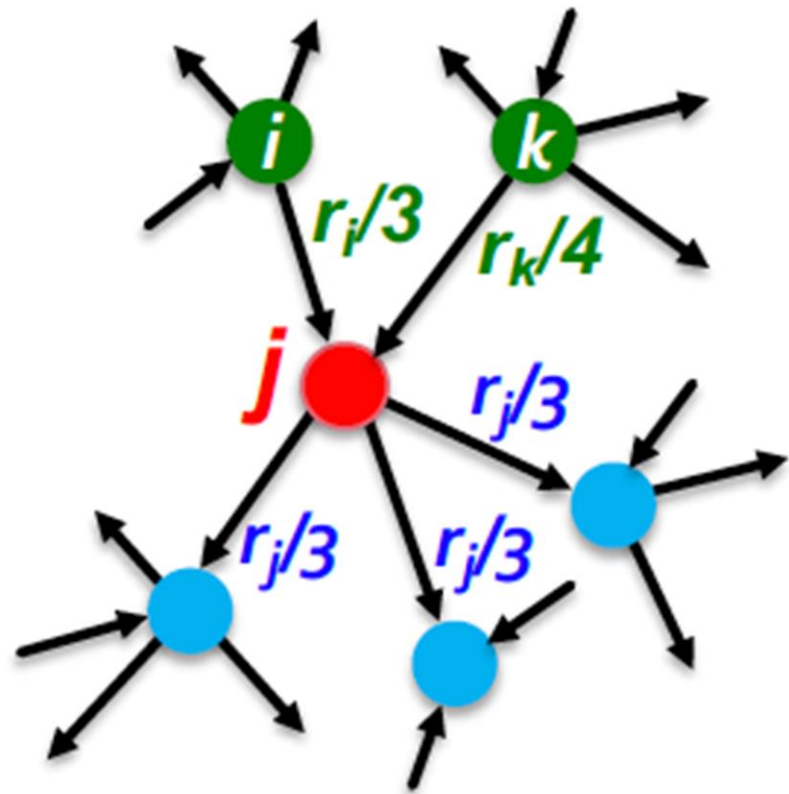


你觉得10和6谁更重要？  
还真难说；因为6得到了5/7/8的投票；  
但是10得到了9的投票，而9得到了6的投票。

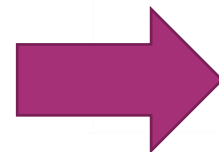
# Page rank

□ PageRank的基本思想：给不同的人边赋上不同的权重

- 考虑某个节点  $v$
- 指向  $v$  的节点的PageRank值越高，相应入边的权重越高
- 指向  $v$  的节点指向其它节点的数目越多，分摊越多，对  $v$  相应入边的权重越低



- 1,  $i$  指向  $j$
- 2,  $k$  指向  $j$
- 3,  $i$  的出度为3, 以  $1/3$  分摊
- 4,  $k$  的出度为4, 以  $1/4$  分摊



$$r_j = r_i/3 + r_k/4$$

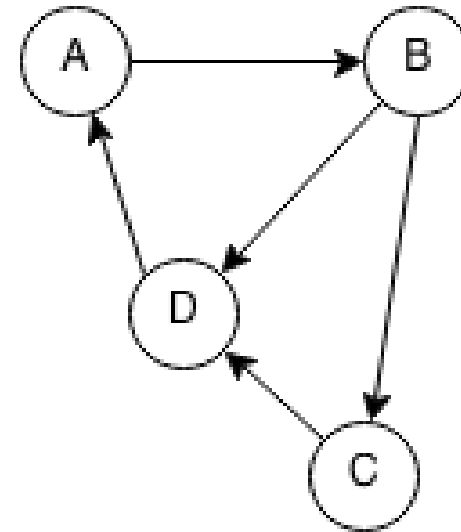
# Page rank

## □如何用数学表达上述想法

- 定义有向图的邻接矩阵 $A = \{L_{ij}\}$
- 其中 $L_{ij} = 1$ 表示*i到j有边*， $L_{ij} = 0$ 表示无边
- 以下图为例

		To			
		A	B	C	D
From	A	0	1	0	0
	B	0	0	1	1
	C	0	0	0	1
	D	1	0	0	0

$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$



# Page rank

## □如何用数学表达上述想法

■定义每个节点的出度为 $m_i$ ，则有

$$m_i = \sum_{j=1}^n L_{ij}$$

■构造M矩阵如下

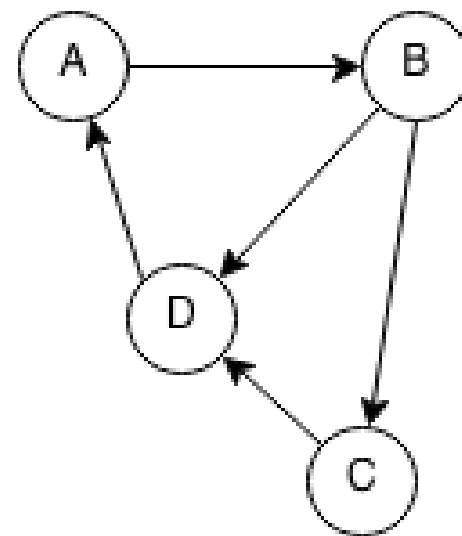
■对角线上的元素的值为

- A的某一行的1的sum，即某个节点的出度

A  
B  
C  
D

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



# Page rank

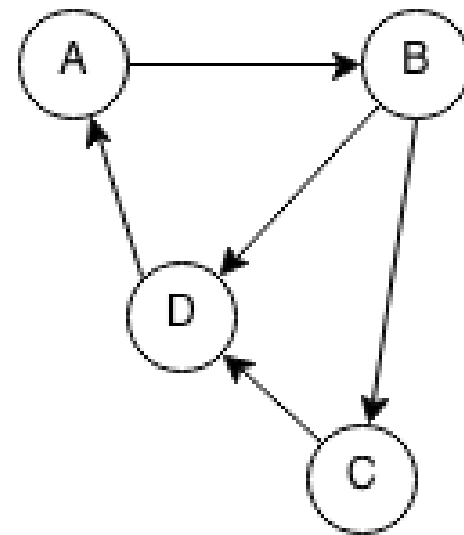
## □如何用数学表达上述想法

### ■计算 $M^{-1}A$

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$M^{-1}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$



B的出度为2，它的重要性按照1/2进行分摊，如何分摊看后文

D的出度为1，它的重要性按照1/1进行分摊...



# Page rank

## □ 如何用数学表达上述想法

■ 假设已有结点A, B, C, D的初始page rank为

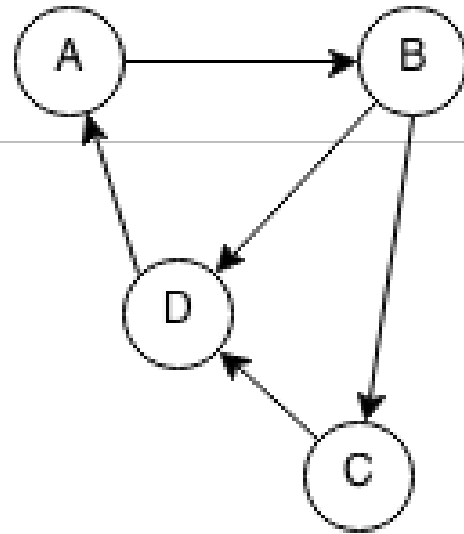
■  $[p_1 \quad p_2 \quad p_3 \quad p_4]$

■ 利用矩阵乘法进行Page Rank值的迭代结算

■  $[p_1 \quad p_2 \quad p_3 \quad p_4] = [p_1 \quad p_2 \quad p_3 \quad p_4] M^{-1} A$

■  $= [p_1 \quad p_2 \quad p_3 \quad p_4] \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

■  $= \begin{bmatrix} p_4 & p_1 & \frac{1}{2}p_2 & \frac{1}{2}p_2 \end{bmatrix}$



这里  $\frac{1}{m_1}L_{11}$ 、 $\frac{1}{m_2}L_{21}$ 、 $\frac{1}{m_3}L_{31}$  都是0  
 $\frac{1}{m_4}L_{41}=1$

看看新的  $p_1$

$$= p_1 \frac{1}{m_1} L_{11} + p_2 \frac{1}{m_2} L_{21} + p_3 \frac{1}{m_3} L_{31} + p_4 \frac{1}{m_4} L_{41}$$

它是其它各个节点的重要度，根据是否有它们到本节点的连接，分摊到本节点的重要度，累加

$p_2, p_3, p_4$  做类似理解

# Page rank

□ 写出PageRank值 $p_i$ 的递推公式

$$p_i = \sum_{j \rightarrow i} \frac{p_j}{m_j} = \sum_{j=1}^n \frac{L_{ji}}{m_j} p_j$$

看看新的 $p_i$

$$= p_1 \frac{1}{m_1} L_{1i} + p_2 \frac{1}{m_2} L_{2i} + p_3 \frac{1}{m_3} L_{3i} + p_4 \frac{1}{m_4} L_{4i}$$

它是其它各个节点的重要度，根据是否有它们到 $p_i$ 的连接，分摊到本节点的重要度的累加

将上面的公式写成矩阵形式

# Page rank

## □ Page Rank迭代过程的一般形式

■ 写出PageRank值 $p_i$ 的递推公式

$$\mathbf{p} = (p_1, p_2, \dots, p_n)$$

$$\mathbf{A} = \begin{bmatrix} L_{11} & L_{12} & \dots & L_{1n} \\ L_{21} & L_{22} & \dots & L_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ L_{n1} & L_{n2} & \dots & L_{nn} \end{bmatrix}$$

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & m_n \end{bmatrix}$$

$$\mathbf{p} \leftarrow \mathbf{p}(\mathbf{M}^{-1}\mathbf{A})$$

$$\text{Let } \mathbf{L} = \mathbf{M}^{-1}\mathbf{A}$$
$$\mathbf{p}^{t+1} \leftarrow \mathbf{p}^t \mathbf{L}$$

$$p_i = \sum_{j \rightarrow i} \frac{p_j}{m_j} = \sum_{j=1}^n \frac{L_{ji}}{m_j} p_j$$

看看新的 $p_i$

$$= p_1 \frac{1}{m_1} L_{1i} + p_2 \frac{1}{m_2} L_{2i} + p_3 \frac{1}{m_3} L_{3i} + p_4 \frac{1}{m_4} L_{4i}$$

它是其它各个节点的重要度，  
根据是否有它们到 $p_i$ 的连接，分  
摊到本节点的重要度的累加

将上面的公式写成矩阵形式

# Page rank

---

- PageRank分值稳定代表了什么?
- 分值稳定为什么重要?
  - 度量节点重要性需要**分值稳定**
- 分值会稳定到什么状态?
  - 分值稳定意味着  $p^{t+1} = p^t$   **$p = pL$**
- 这说明稳定状态时
  - $p$ 是矩阵 $L$ 对应**特征值为1**的特征向量!
- 可是……
  - 1.我们怎么能确定 $L$ **有**为1的特征值?
  - 2.就算有, 特征向量 $p$ **唯一**吗?

# Page rank

□真正的PageRank算法

□在前面计算的公式的基础上做了“微小”改动

$$p = \alpha pL + \frac{1-\alpha}{n} pE, \text{ **E is the } n \times n \text{ matrix of 1s}**$$

α: Damping parameter, 经验上取0.85

$$p = p(0.85 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} + 0.15 \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix})$$

以0.15的比例，在每个节点上按照1/4跳转到本节点和另外3个节点

# Page rank

□真正的PageRank算法

□在前面计算的公式的基础上做了“微小”改动

$$p = \alpha pL + \frac{1-\alpha}{n} pE, \text{ **E is the } n \times n \text{ matrix of 1s**}$$

α: Damping parameter, 经验上取0.85

展开

$$p = p \cdot 0.85 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} + p \cdot 0.15 \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

$$\begin{aligned} p \cdot 0.15 \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} &= [p_1 \quad p_2 \quad p_3 \quad p_4] \cdot 0.15 \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \\ &= 0.15 \begin{bmatrix} \frac{p_1+p_2+p_3+p_4}{4} & \frac{p_1+p_2+p_3+p_4}{4} & \frac{p_1+p_2+p_3+p_4}{4} & \frac{p_1+p_2+p_3+p_4}{4} \end{bmatrix} \\ &= 0.15 \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} = \frac{0.15}{4} [1 \quad 1 \quad 1 \quad 1] \end{aligned}$$

# Page rank

---

□真正的PageRank算法

□在前面计算的公式的基础上做了“微小”改动

$$p = \alpha pL + \frac{1-\alpha}{n} e, e \text{ 为元素都为1的行向量}$$



$\alpha$ : Damping parameter, 经验上取0.85

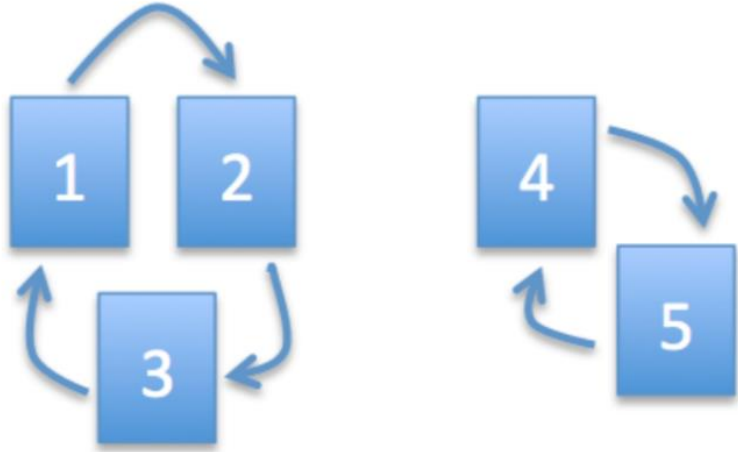
PageRank计算的过程也称随机游走 (Random Walk)



# Page rank

□ 再次考虑之前的反例……考虑  $\alpha = 0.85$

$$p = \alpha pL + \frac{1-\alpha}{n} pE$$



$$L = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$= \frac{0.15}{5} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} + 0.85 \cdot \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0.03 & 0.03 & 0.88 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.88 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \end{pmatrix}.$$

Now **only one** eigenvector of  $A$  with eigenvalue 1:  $p =$

$$\begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}.$$

比较合理

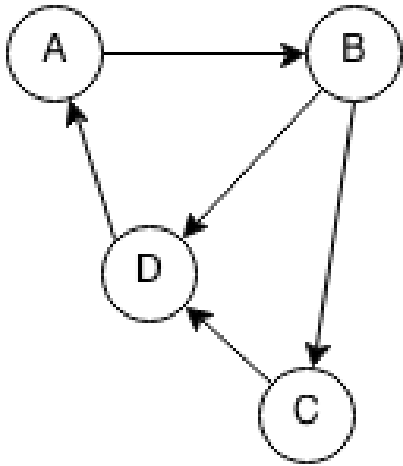
# Page rank

□ 计算该图结构中节点的page rank分值

□ 考虑  $\alpha = 0.8$  且  $p^{(0)} = (1, 0, 0, 0)$

■ 计算  $p^{(1)}$  和  $p^{(2)}$

■ 计算收敛时的  $p$



$$p = \alpha pL + \frac{1-\alpha}{n} pE \rightarrow$$

$$p = p(\alpha L + \frac{1-\alpha}{n} E), E \text{ is the } n \times n \text{ matrix of 1s}$$

$$L = M^{-1}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$p = p(0.80 \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} + \frac{0.20}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}) = p \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.45 & 0.45 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.85 & 0.05 & 0.05 & 0.05 \end{bmatrix}$$

$p^{(0)} = (1, 0, 0, 0)$  代入即可

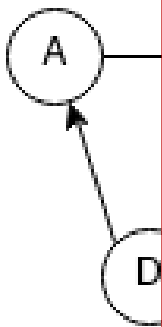
# Page rank

□ 计算该

□ 考虑  $\alpha$

■ 计算  $p$

■ 计算收



[0.05 0.85 0.05 0.05]  
[0.09 0.09 0.39 0.43]  
[0.394 0.122 0.086 0.398]  
[0.3684 0.3652 0.0988 0.1676]  
[0.18408 0.34472 0.19608 0.27512]  
[0.270096 0.197264 0.187888 0.344752]  
[0.3258016 0.2660768 0.1289056 0.279216 ]  
[0.2733728 0.31064128 0.15643072 0.2595552 ]  
[0.25764416 0.26869824 0.17425651 0.29940109]  
[0.28952087 0.25611533 0.1574793 0.29688451]  
[0.2875076 0.2816167 0.15244613 0.27842957]  
[0.27274365 0.28000608 0.16264668 0.28460358]  
[0.27768287 0.26819492 0.16200243 0.29211978]  
[0.28369582 0.27214629 0.15727797 0.28687992]  
[0.27950393 0.27695666 0.15885852 0.28468089]  
[0.27774471 0.27360315 0.16078266 0.28786948]  
[0.28029558 0.27219577 0.15944126 0.28806739]  
[0.28045391 0.27423647 0.15887831 0.28643132]  
[0.27914505 0.27436313 0.15969459 0.28679723]  
[0.27943779 0.27331604 0.15974525 0.28750092]



,  $E$  is the  $n \times n$  matrix of 1s

$$p = \begin{bmatrix} 0 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \\ 0 & 0 \end{bmatrix} + \alpha \begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.45 & 0.45 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.85 & 0.05 & 0.05 & 0.05 \end{bmatrix}$$

# Page rank

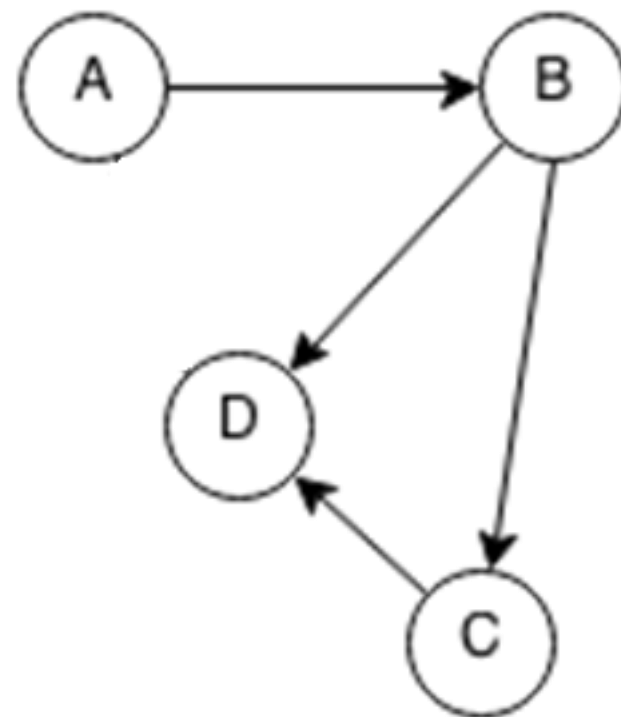
---

## □ 练习

## □ 请计算以下图的PageRank值

■ 请写出邻接矩阵, 设  $\alpha = 0.8$

■ 假设初值为  $p^{(0)} = (1, 0, 0, 0)$

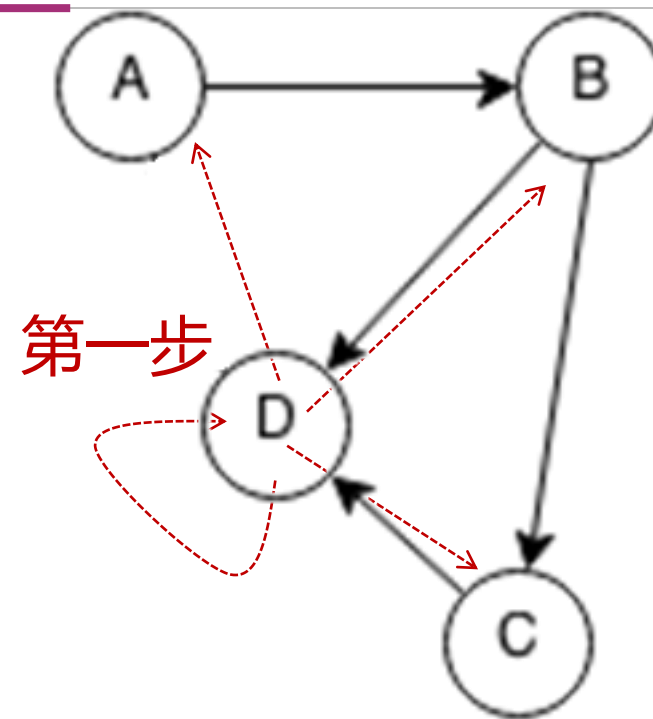


# Page rank练习

□请计算以下图的PageRank值

■请写出邻接矩阵, 设 $\alpha = 0.8$

■假设初值为 $p^{(0)} = (1, 0, 0, 0)$



□解决方案

■将Dangling节点

■与所有节点都建立一条边

■修改邻接矩阵

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\text{注意 } M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$$L = M^{-1}A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

# Page rank

## □练习

## □请计算以下图的PageRank值

■请写出邻接矩阵, 设 $\alpha = 0.8$

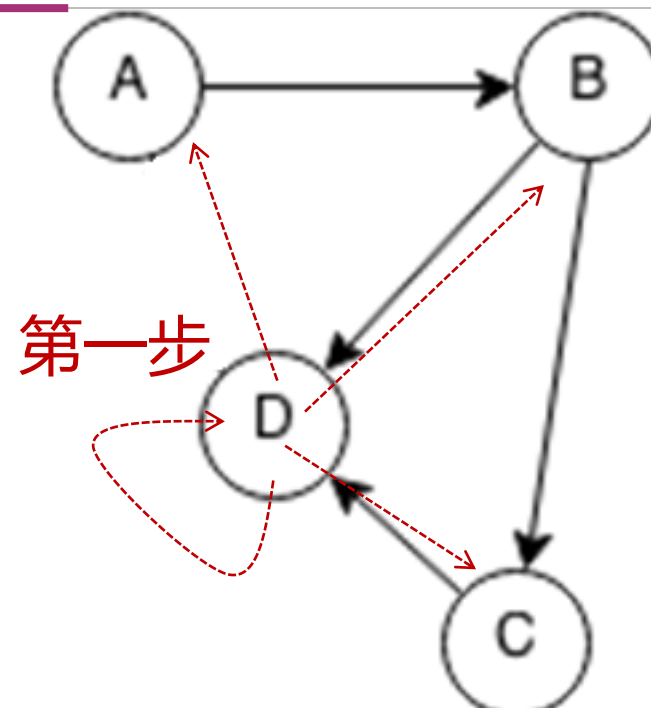
设初值为 $p^{(0)} = (1,0,0,0)$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$L = M^{-1}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

$$p = p(0.80) \left( \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} + \frac{0.20}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \right)$$

$p^{(0)} = (1,0,0,0)$ 代入即可



$$\begin{bmatrix} 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.4 & 0.4 \\ 0 & 0 & 0 & 0.8 \\ 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix} + \begin{bmatrix} 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.05 & 0.05 \end{bmatrix} =$$
$$\begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.45 & 0.45 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$

# 练习题

## □ 前几次迭代的结果如下

- $p^{(1)} = [0.05 \ 0.85 \ 0.05 \ 0.05] \rightarrow \text{sum to } 1$
- $p^{(2)} = [0.06 \ 0.1 \ 0.4 \ 0.44] \rightarrow \text{sum to } 1$
- $p^{(3)} = [0.138 \ 0.186 \ 0.178 \ 0.498] \rightarrow \text{sum to } 1$
- $p^{(4)} = [0.1496 \ 0.26 \ 0.224 \ 0.3664] \rightarrow \text{sum to } 1$
- $p^{(5)} = [0.12328 \ 0.24296 \ 0.22728 \ 0.40648] \rightarrow \text{sum to } 1$
- .
- .
- .

```
17 max_iter=5
18 p = [1,0,0,0]
19 p = np.asarray(p)
20 L = [ [0.05,0.85,0.05,0.05],
21       [0.05,0.05,0.45,0.45],
22       [0.05,0.05,0.05,0.85],
23       [0.25,0.25,0.25,0.25]]
24 for i in range(max_iter):
25     p = p @ L
26     print (p)
27
```

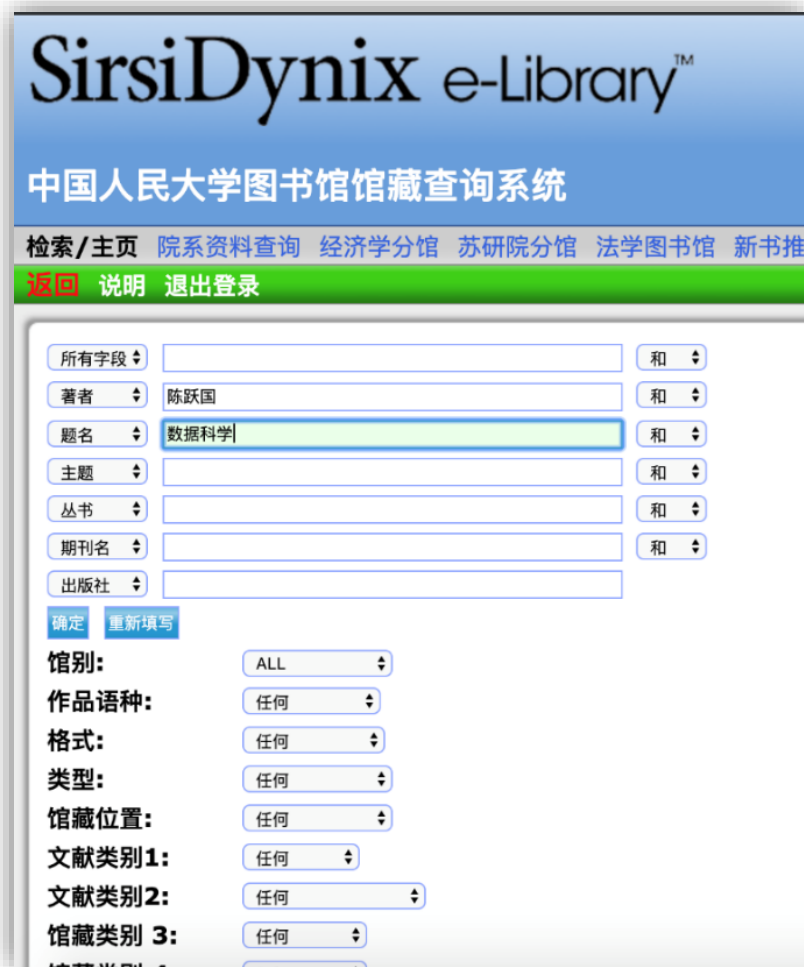
$$p^{(0)} = (1,0,0,0)$$

$$\begin{bmatrix} 0.05 & 0.85 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.45 & 0.45 \\ 0.05 & 0.05 & 0.05 & 0.85 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}$$



# Page rank

## □ PageRank在Web Search中的应用



SirsiDynix e-Library™

中国人民大学图书馆馆藏查询系统

检索/主页 院系资料查询 经济学分馆 苏研院分馆 法学图书馆 新书推

返回 说明 退出登录

所有字段 和

著者 陈跃国 和

题名 数据科学 和

主题 和

丛书 和

期刊名 和

出版社 和

确定 重新填写

馆别: ALL

作品语种: 任何

格式: 任何

类型: 任何

馆藏位置: 任何

文献类别1: 任何

文献类别2: 任何

馆藏类别3: 任何



- 覆盖主题：单一 vs. 多元
- 内容源：专家学者 vs. 普罗大众
- 质量评估标准：清晰 vs. 复杂
- 用户查询：结构化（精确但有门槛）、关键词（易用却可能有歧义）

# Page rank

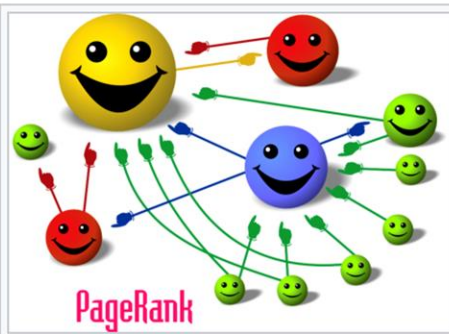
□ PageRank在Web Search中的应用

□ PageRank由谷歌公司的两个创始人Larry Page和Sergei Brin提出，主要解决Web Page的排序问题

PageRank

is a [link analysis](#) algorithm and it assigns a numerical [weighting](#) to each element of a [hyperlinked set](#) of

documents, such as the [World Wide Web](#), with the purpose of "measuring" its relative importance within the set. The [algorithm](#) may be applied to any collection of entities with [reciprocal](#) quotations and



Cartoon illustrating the basic principle of PageRank. The size of each face is proportional to the total size of the other faces which are pointing to it.

```
Elements Console Sources Network Performance Memory Application
"PageRank is a "
<a href="/wiki/Network theory#Link analysis" title="Network theory">link analysis</a>
" algorithm and it assigns a numerical "
<a href="/wiki/Weighting" title="Weighting">weighting</a>
" to each element of a "
<a href="/wiki/Hyperlink" title="Hyperlink">hyperlinked</a>
<a href="/wiki/Set (computer science)" class="mw-redirect" title="Set (computer science)">set</a>
" of documents, such as the "
...
<a href="/wiki/World Wide Web" title="World Wide Web">World Wide Web</a> == $0
", with the purpose of "measuring" its relative importance within the set. The "
<a href="/wiki/Algorithm" title="Algorithm">algorithm</a>
" may be applied to any collection of entities with "
<a href="/wiki/Reciprocal link" class="mw-redirect" title="Reciprocal link">reciprocal</a>
" quotations and references. The numerical weight that it assigns to any given element "
<i>E</i>
" is referred to as the "
<i>PageRank of E</i>
" and denoted by "
```

html body #content #bodyContent div#mw-content-text.mw-content-ltr div.mw-parser-output p a

# Page rank

---

## □ Node Centrality

### □ 1. 基于几何图形的度量方法

- Degree Centrality

- Closeness Centrality

### □ 2. 基于路径的度量方法

- Betweenness Centrality

### □ 3. PageRank算法

- 矩阵运算形式（为什么要有 **damping factor?**）

- 马尔科夫链的数学性质

- 个性化PageRank算法（后文介绍）

$L$  有为1的特征值  
特征向量  $p$  唯一



马尔科夫链存在唯一的稳态分布