Data Mining 2025

Link Analysis

Dept. of Computer Science and Information Engineering

National Cheng Kung University

Kun-Ta Chuang

ktchuang@mail.ncku.edu.tw

Objectives

- To review common approaches to link analysis
- To calculate the popularity of a site based on link analysis
- To model human judgments indirectly

Outline

1. Motivation
2. Early Approaches to Link Analysis
3. Hubs and Authorities: **HITS**
4. **Page Rank**
5. Other issues and Limitation of Link Analysis

Motivation

- Human knowledge is real, convincing and trustable information
  - E.g., classification by human in yahoo
- Hyperlinks contain information about the *human judgment*
- Social sciences
  - Nodes: persons, organizations
  - Edges: social interaction
- **Easy job?** *Counting in-links for popularity*

An example: scientific literature

- **Impact factor** (**http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/**)
  - for journal evaluation
  - *Garfield (Science 1955, 1972)*
  - The average number of citations per recently published item
  - C / N
  - **C**: the total number of citations in a given time interval [t, t + t1] to articles published by a given journal during [t – t2, t]
  - **N**: the total number of articles published by that journal in [t – t2, t]
- Issues
  - The number of citation base
  - Normalization?

Mark	Rank	Abbreviated Journal Title (linked to journal information)	ISSN	JCR Data			
				Total Cites	Impact Factor	5-Year Impact Factor	Im
☐	1	NAT REV MOL CELL BIO	1471-0072	29222	39.123	42.508	
☐	2	CELL	0092-8674	171297	32.403	34.774	
☐	3	CANCER CELL	1535-6108	19726	26.566	28.174	
☐	4	CELL STEM CELL	1934-5909	10145	25.421	27.494	
☐	5	NAT MED	1078-8956	54228	22.462	26.418	
☐	6	NAT CELL BIOL	1465-7392	29959	19.488	20.116	
☐	7	ANNU REV CELL DEV BI	1081-0706	8399	15.836	19.733	
☐	8	MOL CELL	1097-2765	44493	14.178	14.202	
☐	9	DEV CELL	1534-5807	18481	14.030	14.202	
☐	10	CELL METAB	1550-4131	9907	13.668	17.770	
☐	11	CURR OPIN CELL BIOL	0955-0674	13795	12.897	12.594	
☐	12	NAT STRUCT MOL BIOL	1545-9993	22401	12.712	12.114	

ISI impact factor: **http://isiknowledge.com/**

Early Approaches

Basic Assumptions

- Hyperlinks contain information about the human judgment of a site
- The more incoming links to a site, the more it is judged ***important***

Bray 1996 *(Measuring the Web, WWW)*

- The **visibility** of a site is measured by the number of other sites pointing to it (indegree)
- The **luminosity** of a site is measured by the number of other sites to which it points (outdegree)
- ***Limitation***: **failure to capture the relative importance of different parents (children) sites**
- *But works in some reports!*

Early Approaches

Mark *(Commun ACM, 1988)*

- To calculate the score S of a document at vertex v

$$S(v) = s(v) + \frac{1}{|\,ch[v]\,|} \sum_{w \in |ch(v)|} S(w$$

v: a vertex in the hypertext graph G = (V, E)
S(v): the global score
s(v): the score if the document is isolated
ch(v): children of the document at vertex v

- Limitation:

- Require G to be a directed acyclic graph (DAG)

- If v has a single link to w, S(v) > S(w)

- If v has a long path to w and s(v) < s(w), then S(v) > S (w)

→***Unreasonable***, *users need go through the long path from the irrelevant document (v) to reach the important document (w)*

→*But show the message passing schemes*

Early Approaches

Marchiori *(WWW, 1997)*

- ***Hyper information*** should complement textual information to obtain the overall information

  $S(v) = s(v) + h(v)$

  - S(v): overall information
  - s(v): textual information
  - h(v): hyper information

Can't handle real world cases → a cyclic graph

- $h(v) = \sum_{w \in |ch[v]|} F^{r(v, w)} S(w)$
  - **F**: a fading constant, $F \in (0, 1)$
  - **r(v, w)**: the rank of w after sorting the children of v by S(w)

→ a remedy of the previous approach (Mark 1988)

HITS - Kleinberg's Algorithm

- HITS – **H**ypertext **I**nduced **T**opic **S**election
- For each vertex v Є V in a subgraph of interest:

a(v) - the authority of v

h(v) - the hubness of v

- A site is very **authoritative** if it receives many citations. *Citation from important sites **weight more** than citations from less-important sites*
- Hubness shows the **importance** of a site. A good hub is a site that links to many authoritative sites

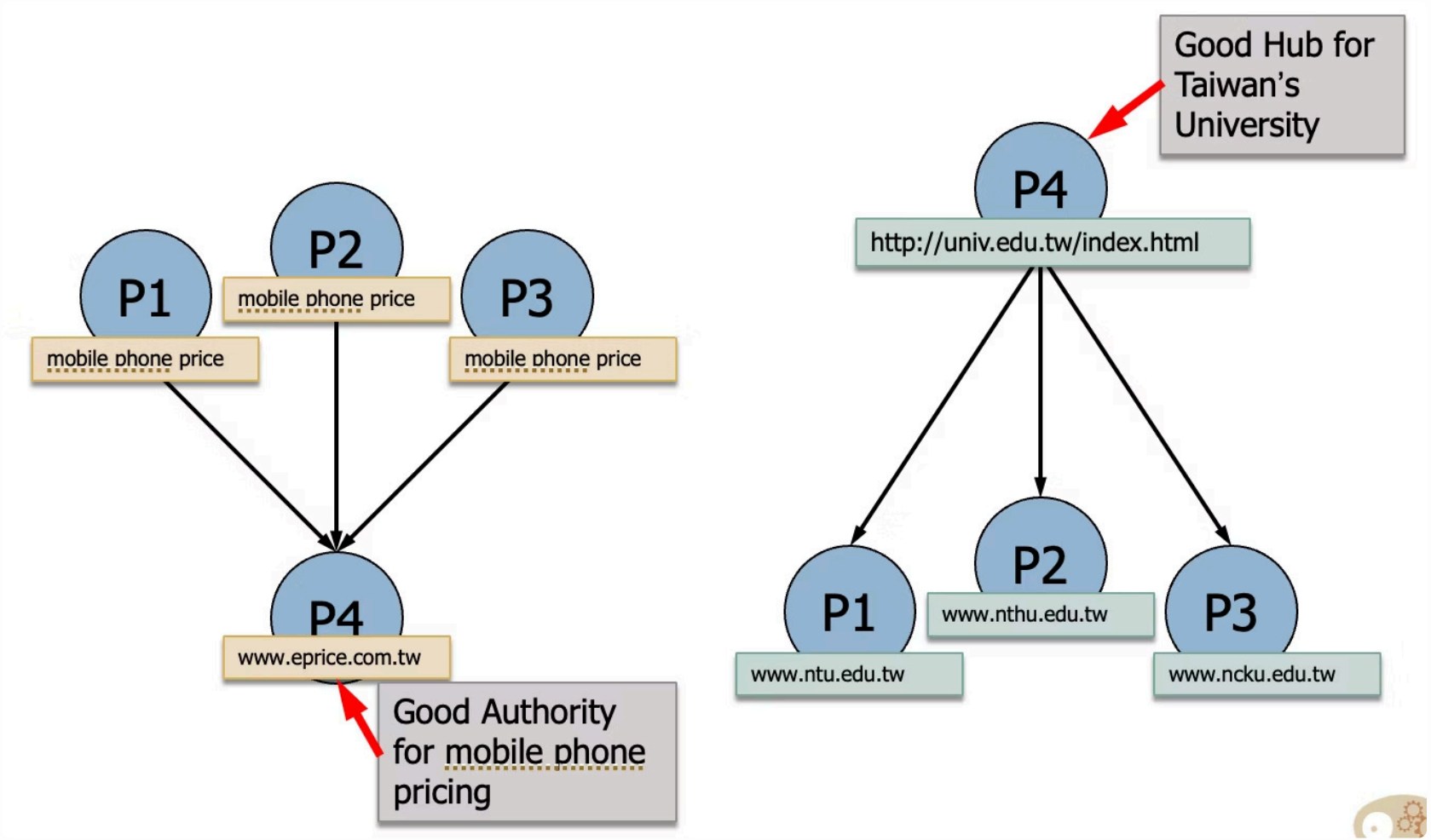
雞生蛋，蛋生雞?

Twin relation v.s. triple relation or more

Motivation

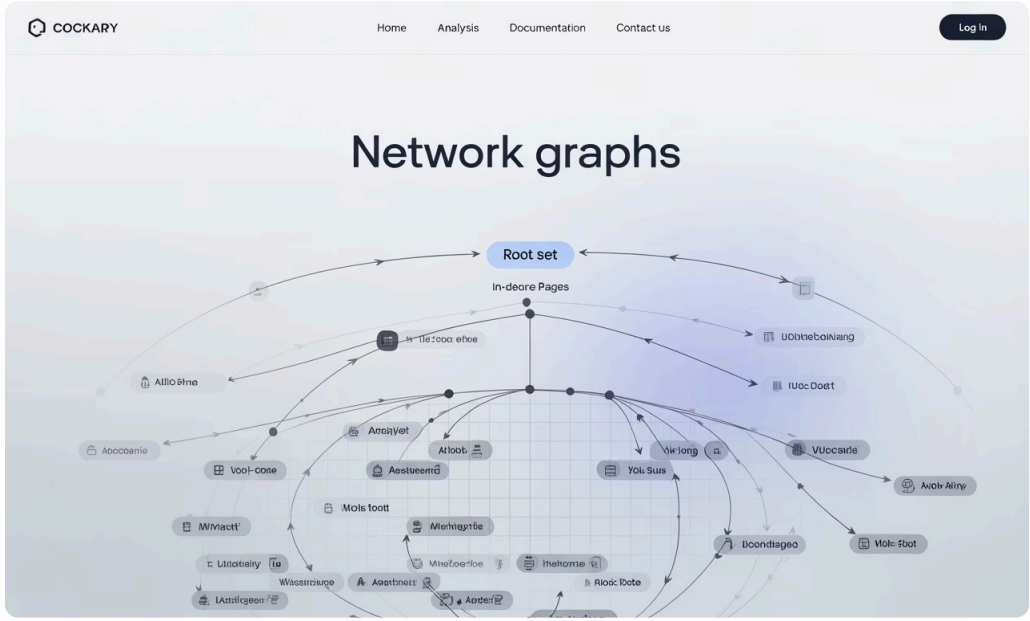
- For a given query, which pages are the answer set?
- Results of search engines
  - Rank manually
  - Rank by similarity
  - Rank by hit rate *(need usage log)*
  - Rank by link analysis (HITS, PageRank,...)
- Relevant v.s. Authoritative
  - Intra-page v.s. inter-page
- **Users need authoritative pages among relevant pages.**

Authorities and Hubs



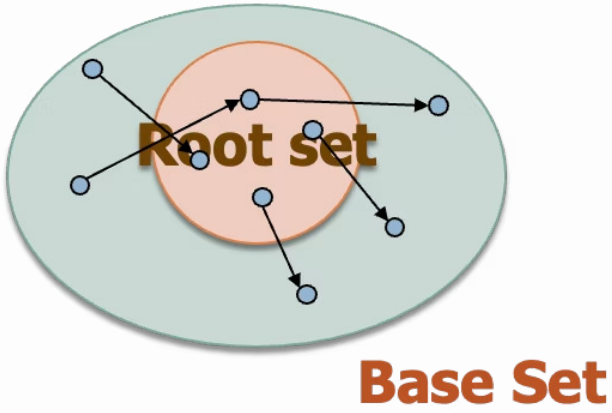
Introduction

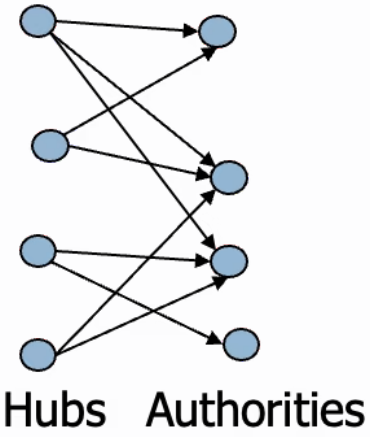
- How to find authoritative pages for queries
  - Step I: rank pages according to their **in-degree** in the **sub-graph** induced by the **root set S**
    - root set: top k pages indexed by search engines
    - Problems
      - very few edges, a large fraction of the nodes will be isolated
      - real authoritative pages are not included in the root set



Introduction

- Step II: **extend** the root set to **base set**
  - Problems
    - Unrelated page of large in-degree
- New approach (kleinberg '97)
  - There should also be considerable overlap in the sets of pages that point to authoritative pages.
    - Hub pages
    - *mutually reinforcing relationship*





Authority and Hubness Convergence

- Recursive dependency:

$a(v) \approx \Sigma w \in pa[v]\ h(w)$

$h(v) \approx \Sigma w \in ch[v]\ a(w)$

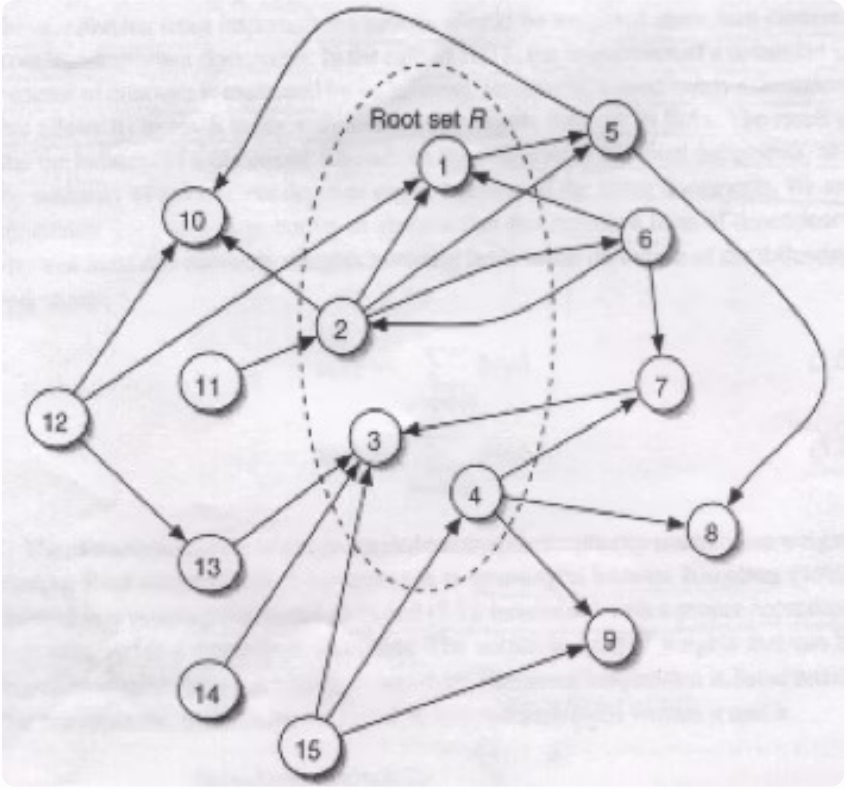
- Using Linear Algebra, we can prove:

a(v) and h(v) converge

HITS Example

Find a base subgraph:

- Start with a root set R {1, 2, 3, 4}
- {1, 2, 3, 4} - nodes relevant to the topic
- Expand the root set R to include all the children and a fixed number of parents of nodes in R
    - *Indegree v.s. outdegree*



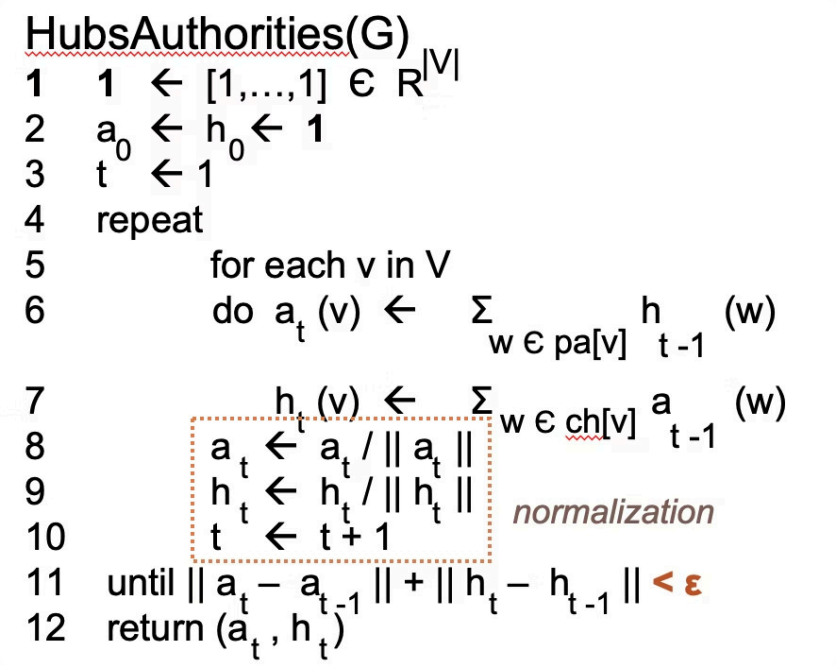
A new set S (base subgraph)

HITS Example

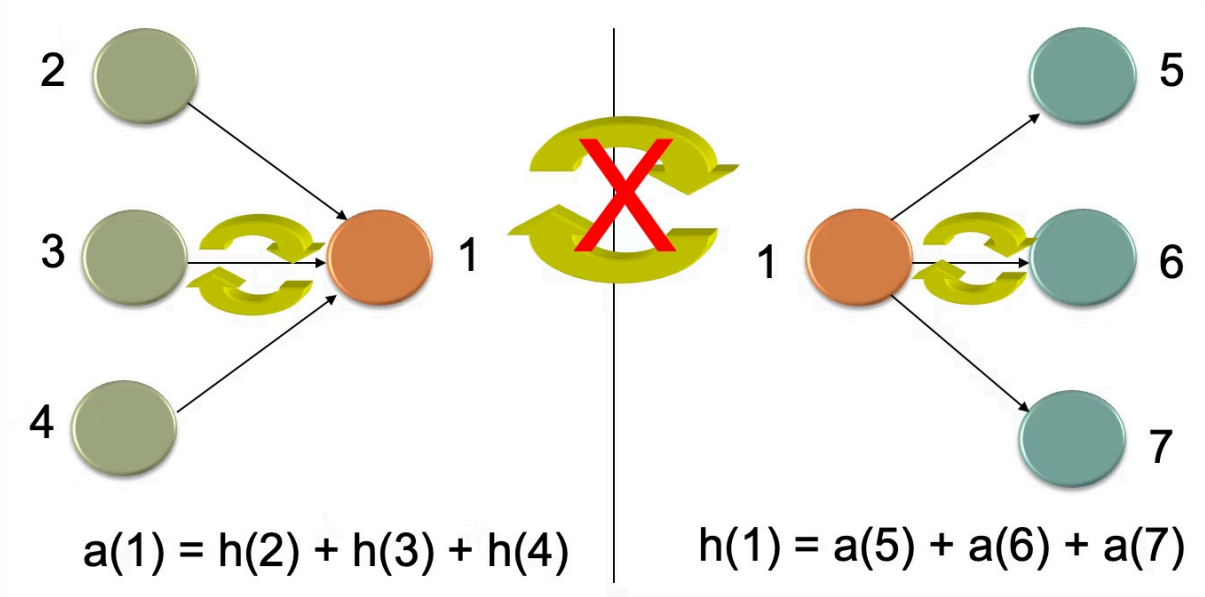
```
BaseSubgraph( R, d)
1.  S <- r
2.  for each v in R
3.      do S <- S U ch[v]
4.         P <- pa[v]
5.         if |P| > d
6.            then P <- arbitrary subset of P having size d
7.                 S <- S U P
8.  return S
```

HITS Example

Hubs and authorities: two n-dimensional **a** and **h**



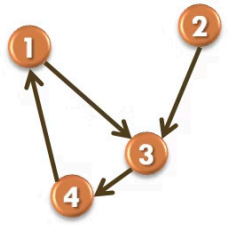
Authority and Hubness



Basic Link Analysis

- Let A denote the **adjacency matrix** of the graph, $\boldsymbol{a}_t \leftarrow A^t \boldsymbol{h}_{t-1}$, $\boldsymbol{h}_t \leftarrow A\, \boldsymbol{a}_{t-1}$
  - $\boldsymbol{a}_n$ is the unit vector in the direction of $(A^tA)^{n-1}A^tz$
  - $\boldsymbol{h}_n$ is the unit vector in the direction of $(AA^t)^nz$
- $\boldsymbol{a}^*$ is the principal eigenvector of A^tA, and $\boldsymbol{h}^*$ is the principal eigenvector of AA^t

Adjacency matrix



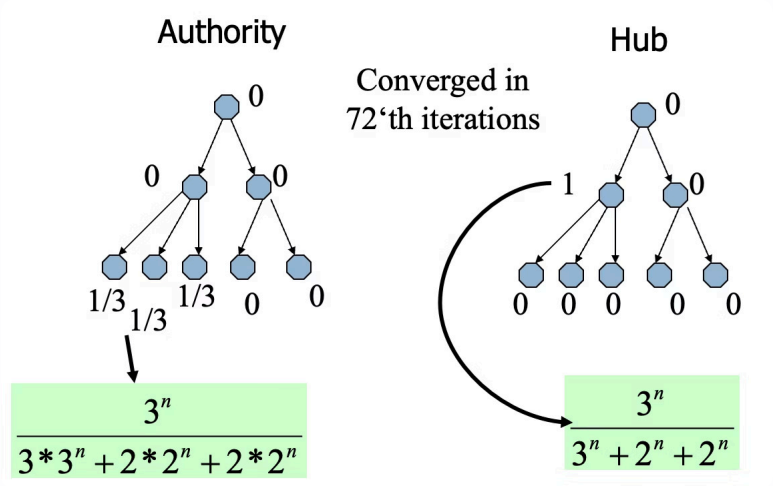
$$A = \begin{bmatrix} 0010 \\ 0010 \\ 0001 \\ 1000 \end{bmatrix} \qquad A^t A = \begin{bmatrix} 1000 \\ 0000 \\ 0020 \\ 0001 \end{bmatrix} \qquad AA = \begin{bmatrix} 0001 \\ 0001 \\ 1000 \\ 0010 \end{bmatrix}$$

In-Out

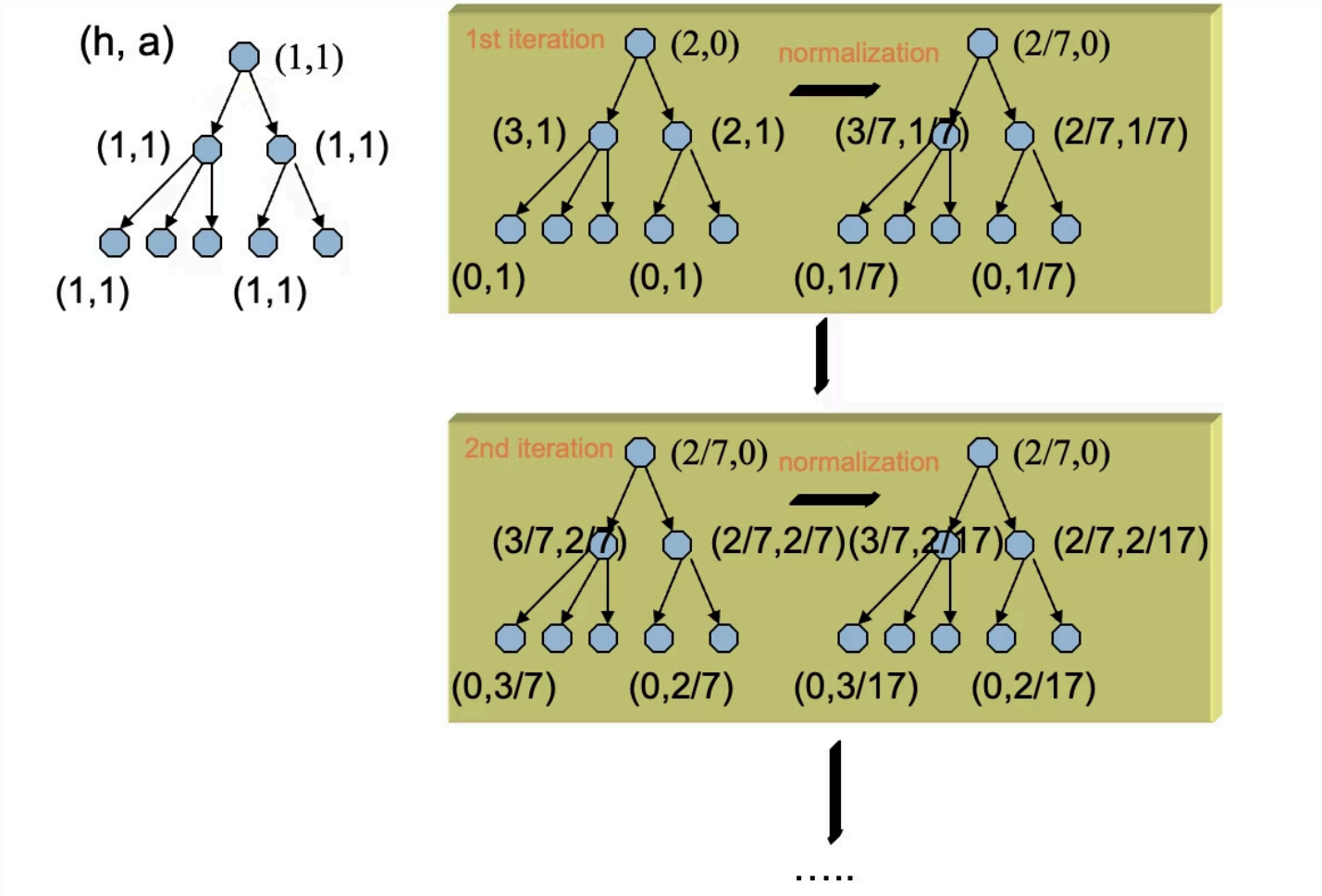
$$A^t = \begin{bmatrix} 0001 \\ 0000 \\ 1100 \\ 0010 \end{bmatrix} \qquad AA^t = \begin{bmatrix} 1100 \\ 1100 \\ 0010 \\ 0001 \end{bmatrix}$$

Out-In

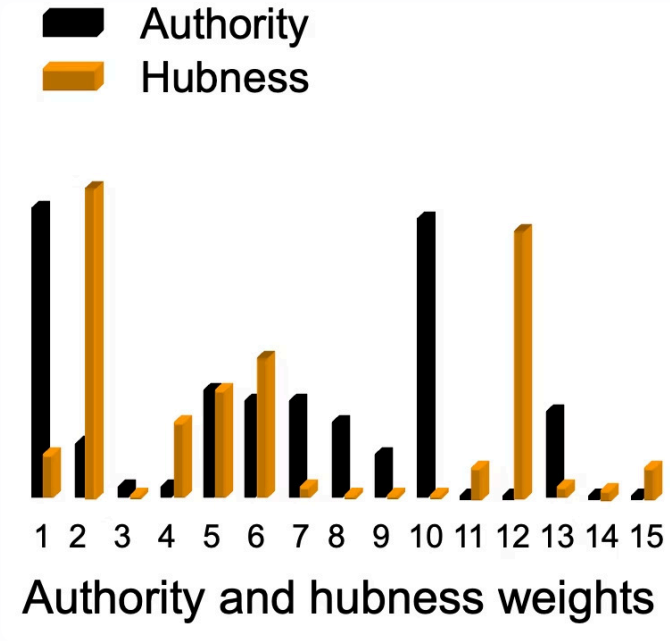
Example (1-norm normalization)

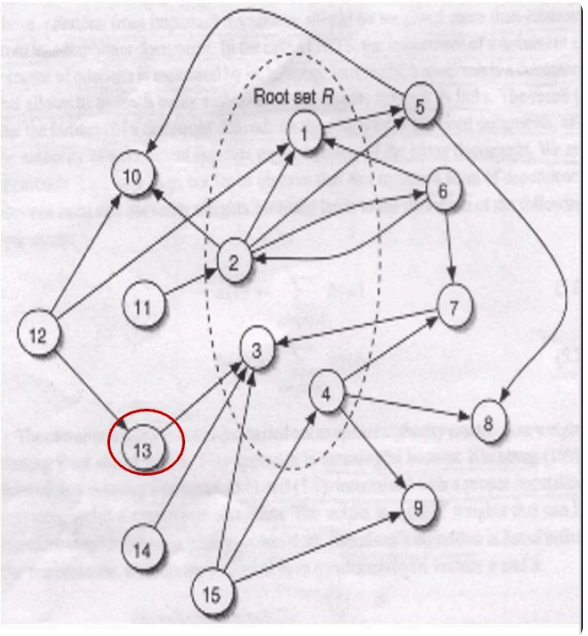


Example (1-norm normalization)



HITS Example Results



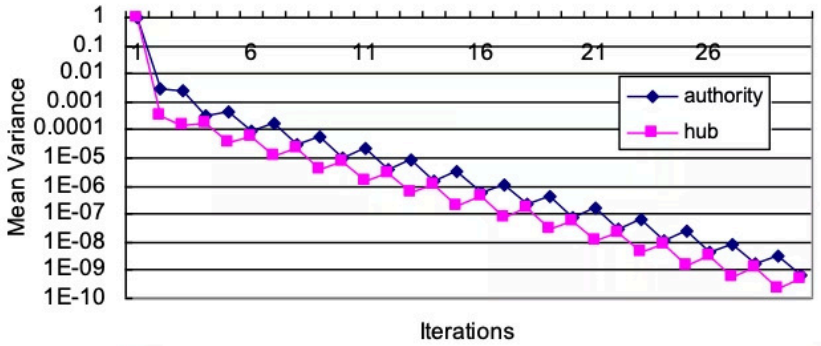


Issues for HITS

- Mutually reinforcing relationships between hosts
- Nepotistic links cancellation
  - Nepotistic links: links between pages that are present for reasons other than merit
    - Menu links
    - Link-based spam
- Link normalization

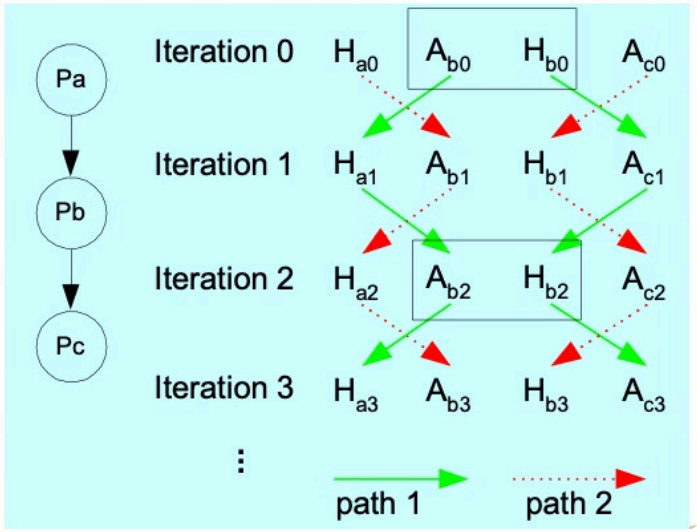
One important observation

- The process of link analysis
  - Convergence of values of hubs and authorities
  - Two (hub, authority) pairs



$$\{(A_{a3}, H_{a3}), (A_{b2}, H_{b2}), (A_{c3}, H_{c3})\}$$

$$\{(A_{a2}, H_{a2}), (A_{b3}, H_{b3}), (A_{c2}, H_{c2})\}$$



PageRank

PAGERANK

- The innovation provided by Google was called "PageRank."
- Evaluating the importance of Web page
- Efficient and accurate Web search engines => Google
- First able to defeat the web spammers

Early Search Engines

- Crawling the Web and listing the terms found in each page in an inverted index
- An inverted index is to find all the places where that term occurs
- Search query was issued, the pages with those terms were extracted and ranked
- Presence in a header is more relevant than in ordinary text
- Large numbers of occurrences

Term Spam

- People who fool search engines=>Spammers
- Add a term ("movie") to your page thousands of times, give it the background color
- Search "movie" => copy the first choice page into your page=> give it the background color
- Term spam: Techniques for fooling search engines into believing your page is about something it is not

Combat term spam

Two innovations:

1. Pages that would have a large number of surfers were considered more "important" than pages that would rarely be visited.
2. The content of a page was judged not only by the terms appearing on that page, but by the terms used in or near the links to that page

Definition of PageRank

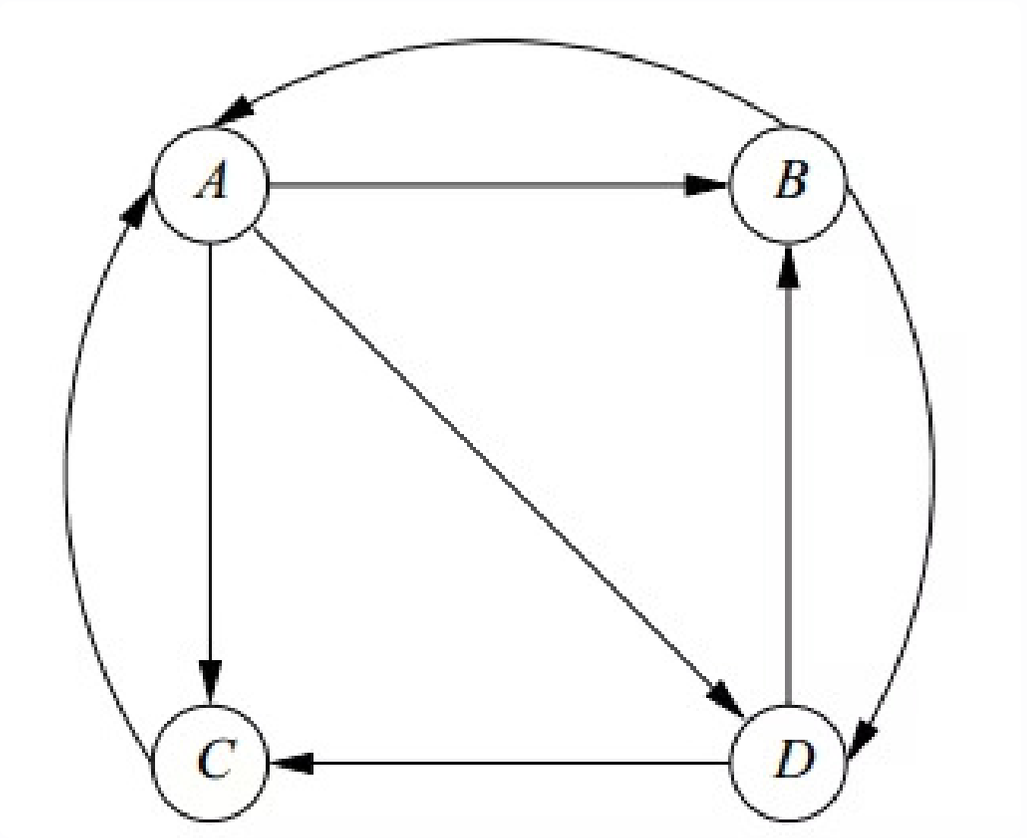
- PageRank assigns a real number to each page
- The higher the PageRank of a page, the more "important" it is
- There is not one fixed PageRank algorithm
- We begin by defining the basic, idealized PageRank

Definition of PageRank

- Web is a directed graph

(1)pages are nodes

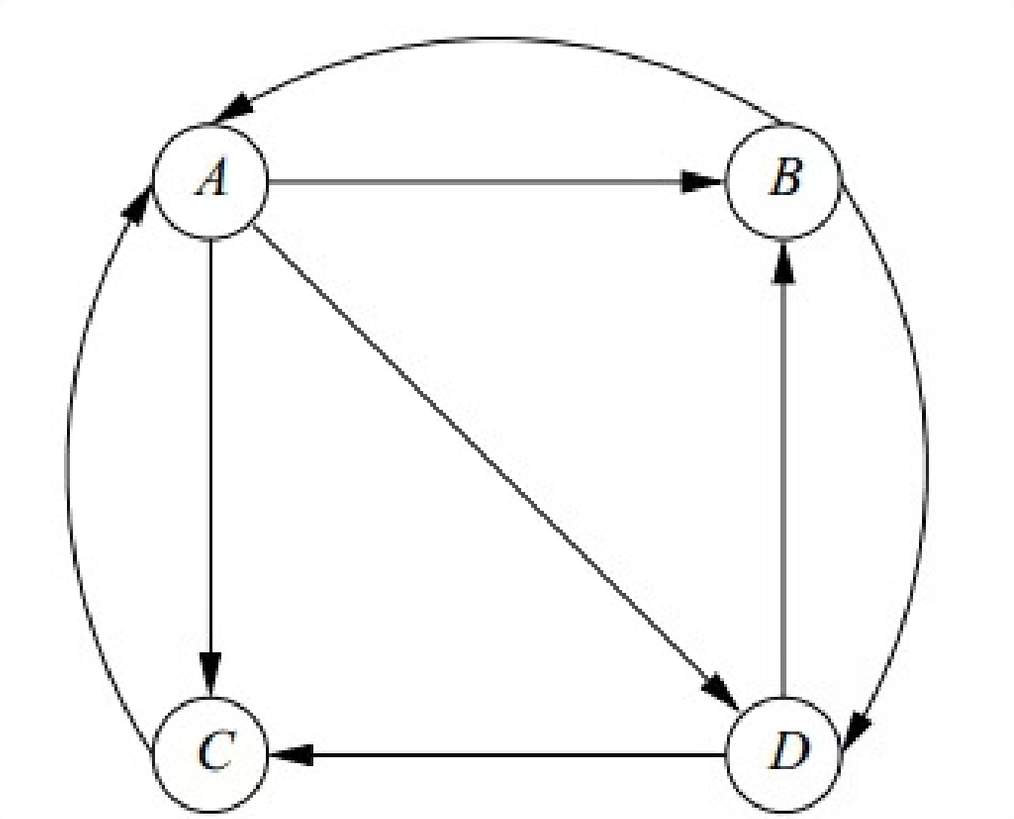
(2)Arc from p1 to p2 =>links from p1 to p2.



Definition of PageRank

TRANSITION MATRIX

Page A has links to B, C, and D with probability 1/3, and has 0 of being at A

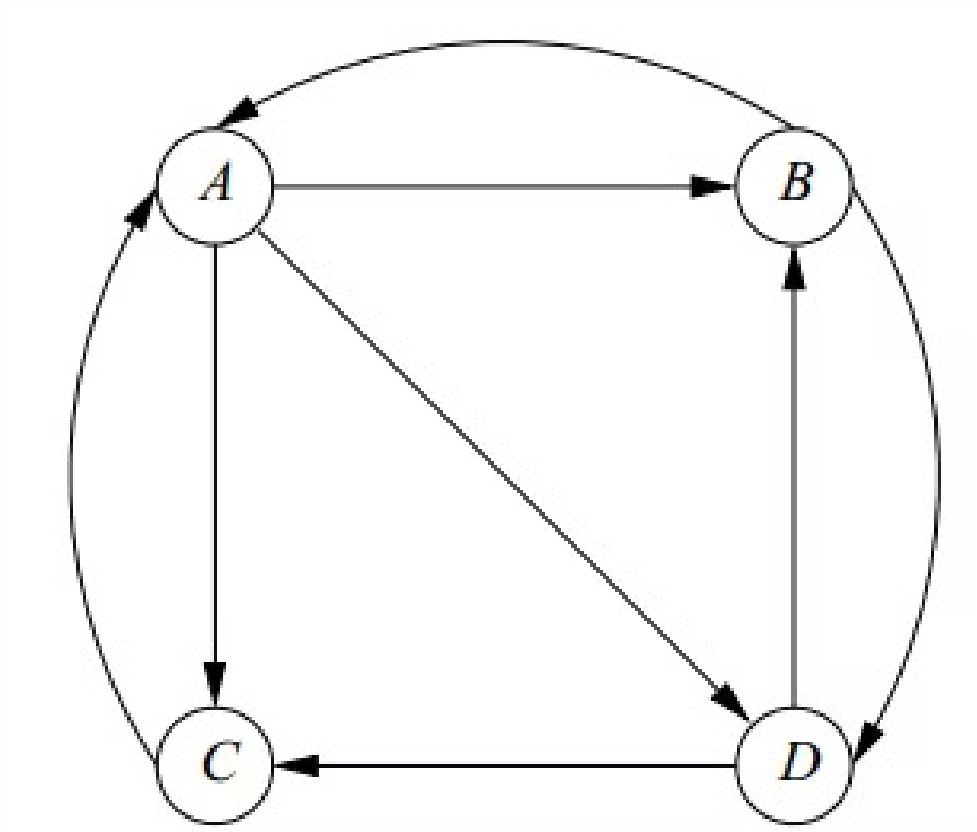


$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$\begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

Definition of PageRank

- Initial vector v0 will have 1/n for each component
- After one step=>Mv0
- After two steps =>M(Mv0) = v0 so on...



$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

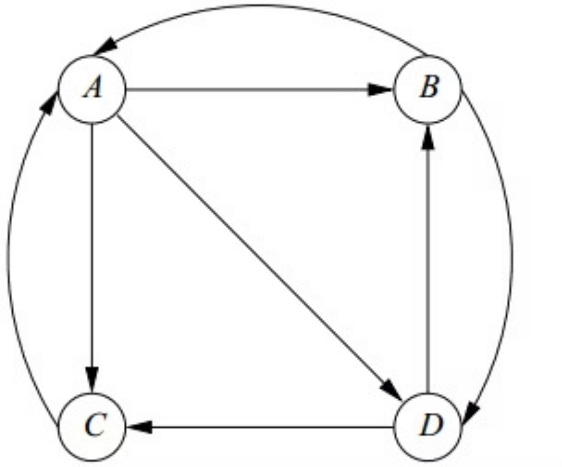
$$V0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{array}{l} A \\ B \\ C \\ D \end{array}$$

Definition of PageRank

- Initial vector v0 will have 1/n for each component
- After one step=>Mv0
- After two steps =>M(Mv0) = v0 so on...

$$V1 = \underset{M}{\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}} \underset{V0}{\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}} = \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}$$

$$V2 = \underset{M}{\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}} \underset{V1}{\begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}} = \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}$$



Definition of PageRank

- **Condition**:

(1) 矩陣每一個元素都是大於或等於 0 的實數

(2) 每一行的和都是 1

=> 馬可夫的原理

- **EX**：

某市及其近郊人口遷移狀況為：每年住在城裡的人有 90%留在城裡，有 10%流向郊區；而郊區的人有 80%留在郊區，有 20%搬到城裡。

城裡　郊區

$$A=\begin{bmatrix} 90\% & 20\% \\ 10\% & 80\% \end{bmatrix}\begin{matrix} \text{城裡} \\ \text{郊區} \end{matrix}$$

如果最初的狀態是城裡人口佔$\frac{3}{4}$，郊區人口佔$\frac{1}{4}$，

則一年後的人口分布情形為：

$$\begin{bmatrix} 90\% & 20\% \\ 10\% & 80\% \end{bmatrix}\begin{bmatrix} 3/4 \\ 1/4 \end{bmatrix}=\begin{bmatrix} 90\%\times 3/4+20\%\times 1/4 \\ 10\%\times 3/4+80\%\times 1/4 \end{bmatrix}=\begin{bmatrix} 29/40 \\ 11/40 \end{bmatrix}$$

Definition of PageRank

- **馬可夫的原理**:

城裡的人口與郊區的人口經過長時間的改變後，會趨近於一個穩定狀態，即 AX=X

$$A=\begin{bmatrix} 90\% & 20\% \\ 10\% & 80\% \end{bmatrix}$$

城裡 / 郊區

設 $X=\begin{bmatrix} x \\ y \end{bmatrix}$

$$\begin{bmatrix} 90\% & 20\% \\ 10\% & 80\% \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 90\%\times x+20\%\times y \\ 10\%\times x+80\%\times y \end{bmatrix}=\begin{bmatrix} x \\ y \end{bmatrix}$$

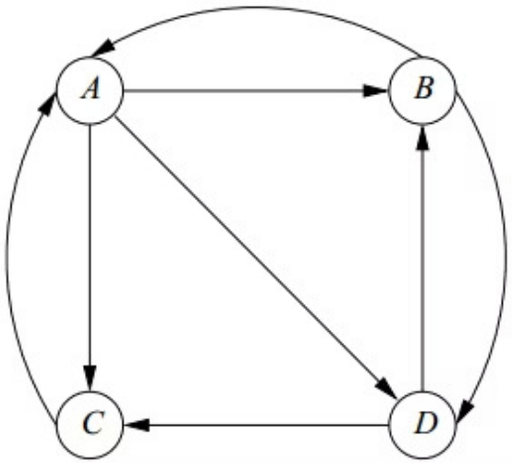
$$\begin{cases} 0.9x+0.2y=x \\ 0.1x+0.8y=y \end{cases}\Rightarrow\begin{cases} 0.1x=0.2y \\ 0.1x=0.2y \end{cases}$$ 即 $x:y=2:1$。

長期而言城市人口佔 $\frac{2}{3}$，郊區人口佔 $\frac{1}{3}$，所以並不會有「空城」的情形發生。

Definition of PageRank

Example 5.2: Suppose we apply the process described above to the matrix M from Example 5.1. The initial vector v0 has four components, each 1/4. The sequence of approximations to the limit:

$$\begin{matrix} V0 & V1 & V2 & & Vn \end{matrix}$$
$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix} \cdots \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$



$$M = \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$Vn = \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \end{bmatrix}$$

Definition of PageRank

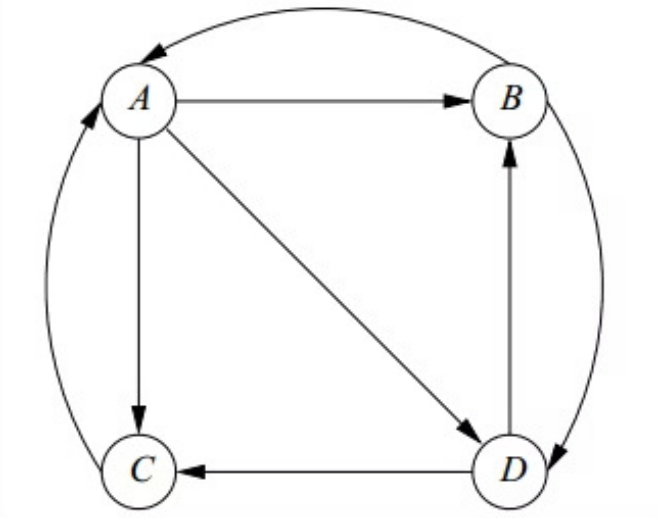
$$\underset{V0}{\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}} \underset{V1}{\begin{bmatrix} 9/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix}} \underset{V2}{\begin{bmatrix} 15/48 \\ 11/48 \\ 11/48 \\ 11/48 \end{bmatrix}} \cdots \underset{Vn}{\begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}}$$

$$\underset{M}{\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}} \underset{Vn}{\begin{bmatrix} V1 \\ V2 \\ V3 \\ V4 \end{bmatrix}} = \underset{Vn}{\begin{bmatrix} V1 \\ V2 \\ V3 \\ V4 \end{bmatrix}}$$

=> v1:v2:v3:v4=3:2:2:2 =>

$$Vn = \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

Definition of PageRank



$$\underset{M}{\begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}} \underset{Vn}{\begin{bmatrix} V1 \\ V2 \\ V3 \\ V4 \end{bmatrix}} = \underset{Vn}{\begin{bmatrix} V1 \\ V2 \\ V3 \\ V4 \end{bmatrix}}$$

=> v1:v2:v3:v4=3:2:2:2 =>

$$Vn = \begin{bmatrix} 3/9 \\ 2/9 \\ 2/9 \\ 2/9 \end{bmatrix}$$

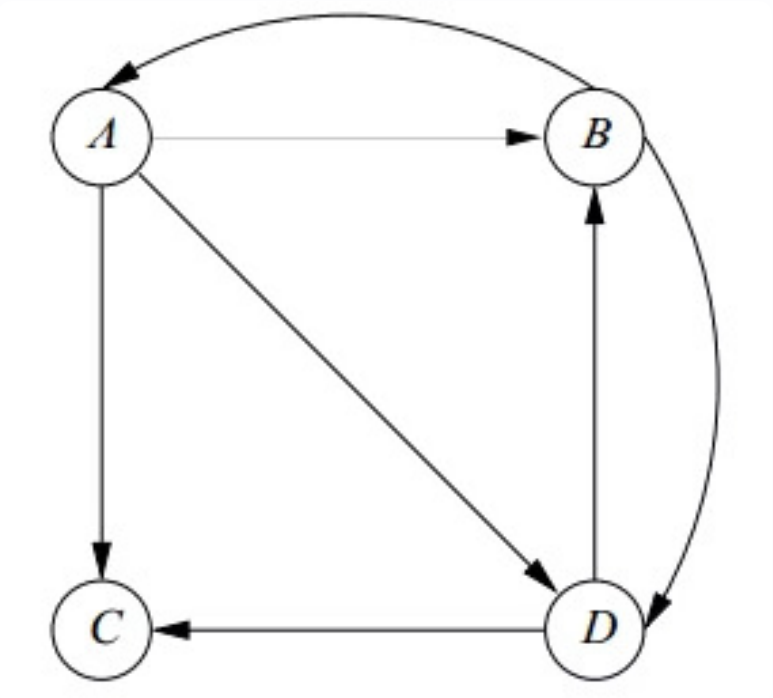
Avoiding Dead Ends

Dead end: a page with no link out

no longer stochastic (matrix whose column sums are 1)

Consequence: importance "drains out"

=>get nothing about the relative importance of pages



$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

Avoiding Dead Ends

Example 5.3:

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 3/24 \\ 5/24 \\ 5/24 \\ 5/24 \end{bmatrix} \begin{bmatrix} 5/48 \\ 7/48 \\ 7/48 \\ 7/48 \end{bmatrix} \begin{bmatrix} 21/288 \\ 31/288 \\ 31/288 \\ 31/288 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \qquad Vn = \begin{bmatrix} v1 \\ v2 \\ v3 \\ v4 \end{bmatrix}$$

=> v1,v2,v3,v4=0

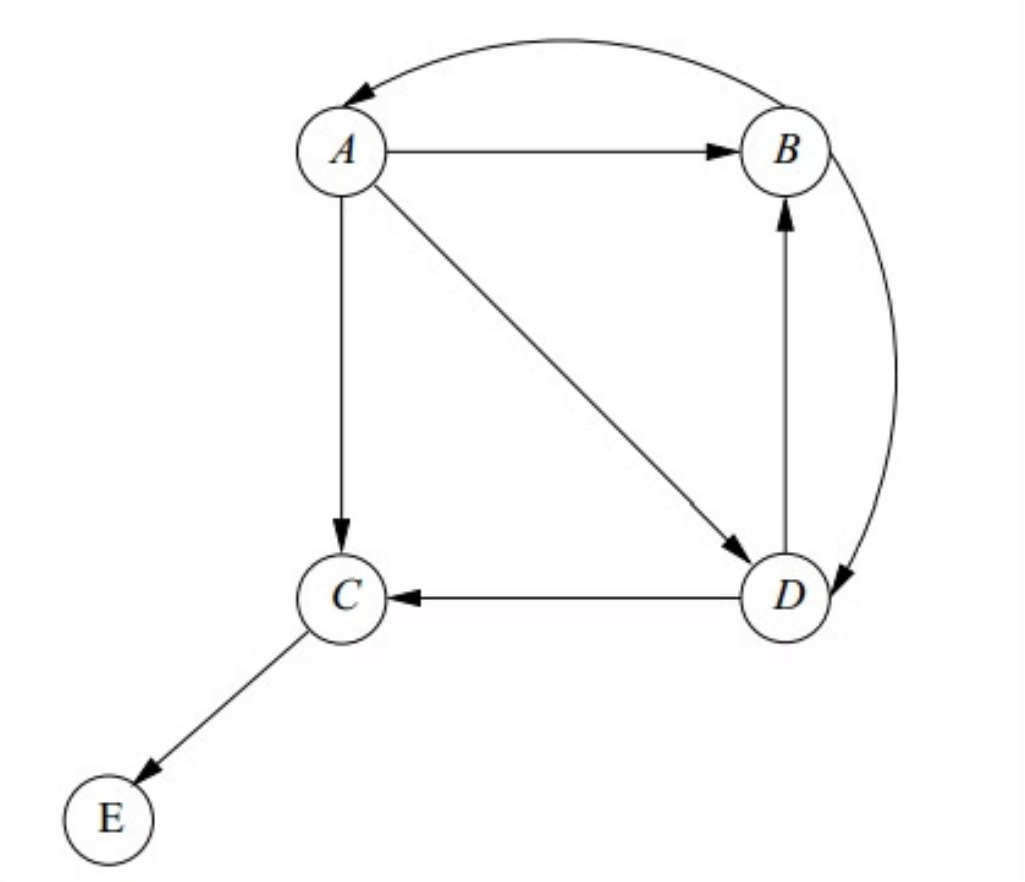
Avoiding Dead Ends

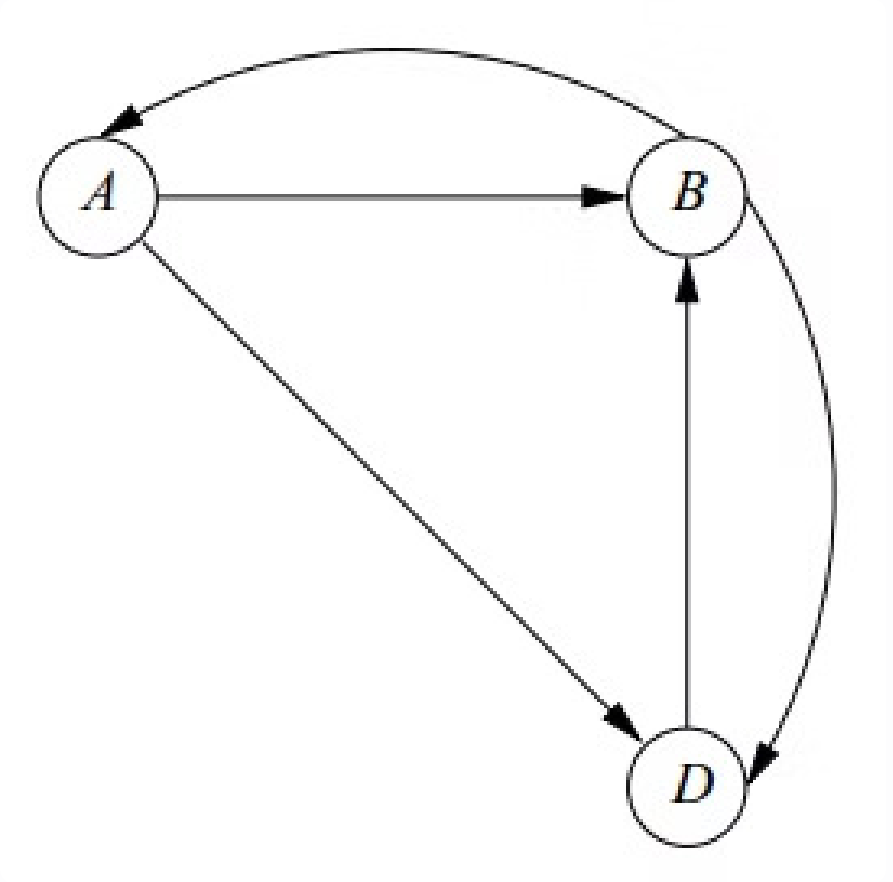
Two methods:

1. Drop the dead ends recursively
2. taxation: Modify the process by which random surfers are assumed to move

Avoiding Dead Ends

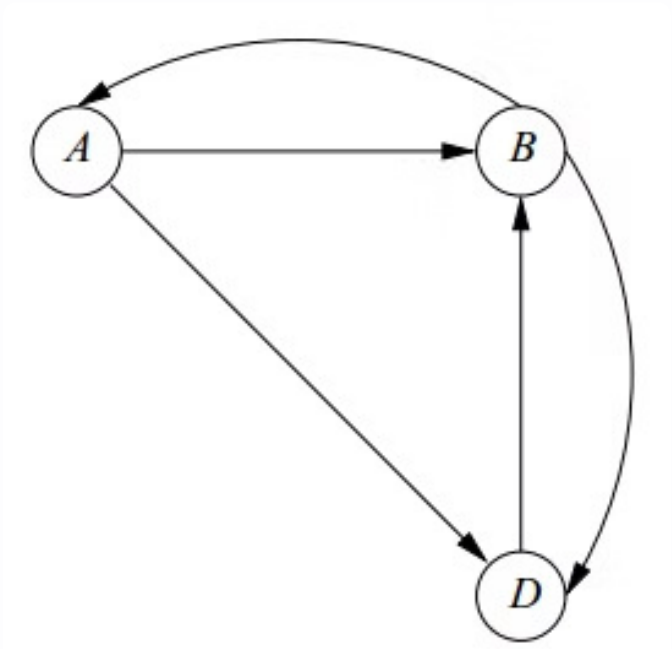
- Drop the dead ends recursively
- Example 5.4





Avoiding Dead Ends

- Drop the dead ends recursively
- Example 5.4

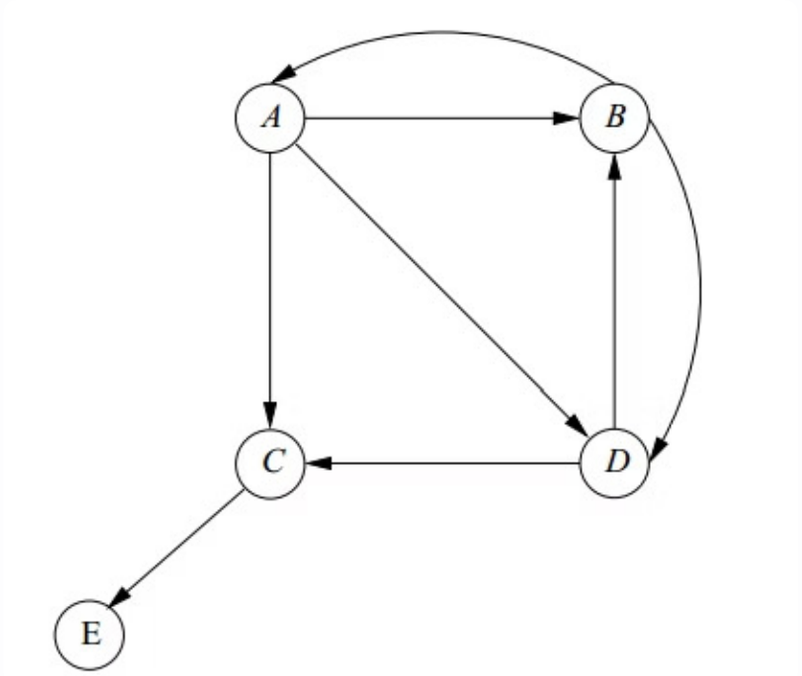


$$M = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \begin{bmatrix} 1/6 \\ 3/6 \\ 2/6 \end{bmatrix} \begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix} \begin{bmatrix} 5/24 \\ 11/24 \\ 8/24 \end{bmatrix} \cdots \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

Avoiding Dead Ends

- Drop the dead ends recursively
- Example 5.4



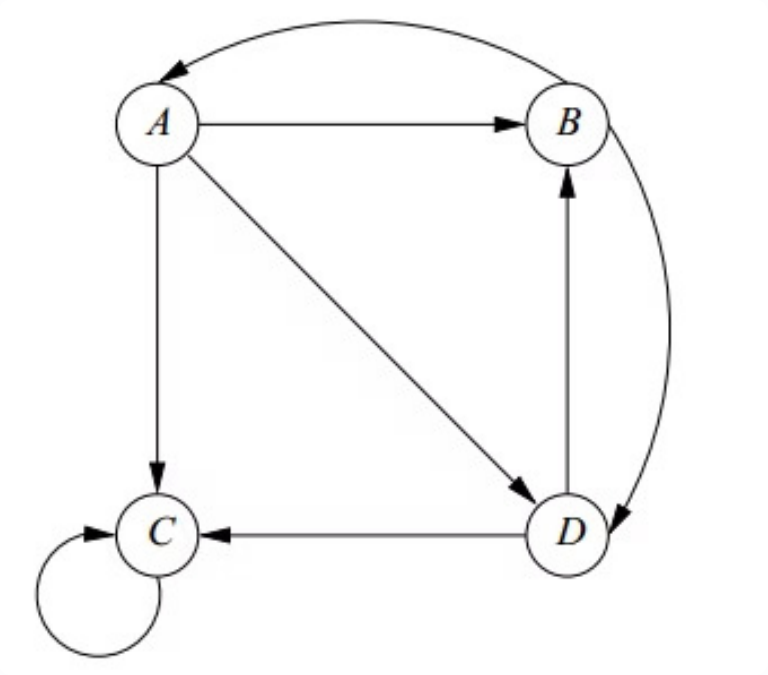
$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \begin{bmatrix} 1/6 \\ 3/6 \\ 2/6 \end{bmatrix} \begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix} \begin{bmatrix} 5/24 \\ 11/24 \\ 8/24 \end{bmatrix} \cdots \begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$$

PageRank for C: 1/3×2/9 + 1/2×3/9 = 13/54

PageRank for E:1×13/54 = 13/54

Spider Traps and Taxation

- **Spider Traps :** a spider trap is a set of nodes with no dead ends but no arcs out

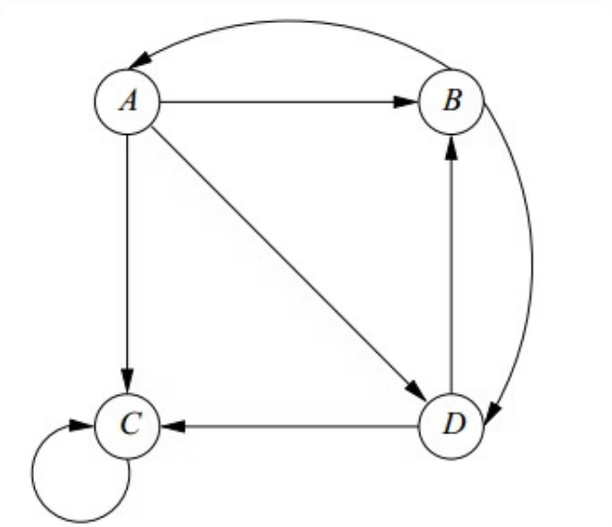


$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- **Consequence** : the PageRank calculation to place all the PageRank within the spider traps

Spider Traps and Taxation

Example 5.5



$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 3/24 \\ 5/24 \\ 11/24 \\ 5/24 \end{bmatrix} \begin{bmatrix} 5/48 \\ 7/48 \\ 29/48 \\ 7/48 \end{bmatrix} \begin{bmatrix} 21/288 \\ 31/288 \\ 205/288 \\ 31/288 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Spider Traps and Taxation

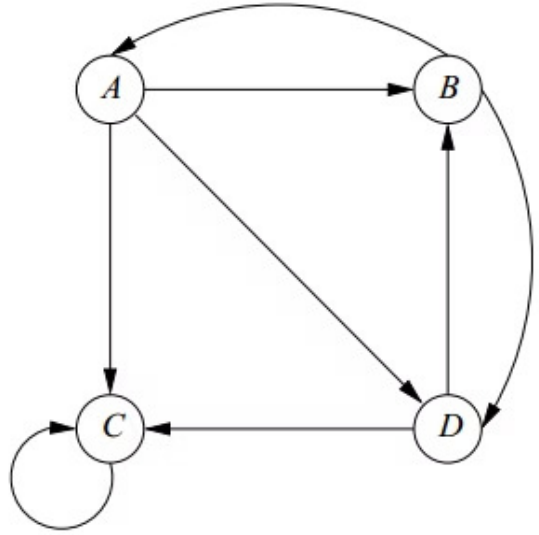
- **Taxation** : modify the calculation of PageRank by sharing a small probability to a random page

v'=MV

=>

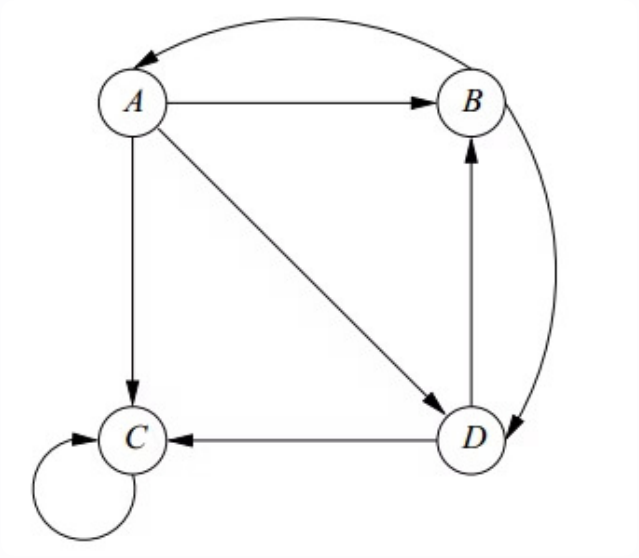
$$\mathbf{v}' = \beta M\mathbf{v} + (1-\beta)\mathbf{e}/n$$

where β is a chosen constant (0.8 ~0.9), e is a vector of all 1's



Spider Traps and Taxation

Example 5.6:



$$\mathbf{v}' = \beta M \mathbf{v} + (1 - \beta)\mathbf{e}/n$$

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 1 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

β=0.8=4/5 1−β = 1/5 and n = 4 => (1- β)/n=1/20

Spider Traps and Taxation

- Example 5.6 :

$$\mathbf{v}' = \beta M \mathbf{v} + (1 - \beta)\mathbf{e}/n$$

β=0.8=4/5 1−β = 1/5 and n = 4 => (1- β)/n=1/20

$$\mathbf{v}' = \begin{bmatrix} 0 & 2/5 & 0 & 0 \\ 4/15 & 0 & 0 & 2/5 \\ 4/15 & 0 & 4/5 & 2/5 \\ 4/15 & 2/5 & 0 & 0 \end{bmatrix} \mathbf{v} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 9/60 \\ 13/60 \\ 25/60 \\ 13/60 \end{bmatrix} \begin{bmatrix} 41/300 \\ 53/300 \\ 153/300 \\ 53/300 \end{bmatrix} \begin{bmatrix} 543/4500 \\ 707/4500 \\ 2543/4500 \\ 707/4500 \end{bmatrix} \cdots \begin{bmatrix} 15/148 \\ 19/148 \\ 95/148 \\ 19/148 \end{bmatrix}$$

Using PageRank in a Search Engine

- Each search engine has a secret formula
- Google is said to use over 250 different properties of pages
- Among the qualified pages:
  - Important component: the PageRank
  - Other components: presence of search terms such as headers or the links to the page itself

Search Engine Optimization (SEO)

Techniques to improve website ranking in search results, directly influenced by link analysis algorithms like HITS

On-Page Factors

Content quality, keyword placement, and HTML structure signal relevance to search engines.

Off-Page Signals

Backlink profile quality determines authority, similar to HITS algorithm mechanics.

Technical SEO

Site architecture and performance metrics affect crawlability and user experience signals.

Q&A

Recap words from Dr. Hsieh

1. 「優秀是一種習慣」- 堅毅不見得會帶來成功，但從成功的人身上，你會看到的是他們都會保持凡事追求優秀的態度，自然而然地，當成一種習慣。
2. 「印象比事實更有殺傷力」- 即使別人認知的不是事實，但真正的人才，都會坦然接受表面印象帶來的的不公平。不必抱怨現實的不完美，你更要找到方法，讓人對你有正確的、好的印象。
3. 「因為信任，所以簡單」- 最終的機會，都來自於被信任。值得信任的人都是那種reliable的人，在充滿風險的創新機會中，最終機會都會擠向那些少少被信任的人。不要以為他們為什麼那麼簡單就獲得機會，你要想想為什麼他能被信任。在AI時代下，永遠有一種人不會被淘汰，就是「解決問題的人」 - 職場工作者必須成為「為最終成果負責」的角色。