



COMP5331

Web databases

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse

Web Databases



Search bar with a text input field.

[Advanced Search](#)
[Preferences](#)
[Language Tools](#)

Search: ☒ the web ☐ pages in Hong Kong

[Classic Home](#) | [iGoogle: Hong Kong Home](#)

Google.com.hk offered in: [中文 \(繁體\)](#)

Raymond Wong

[Advertising Programs](#) - [About Google](#) - [Go to Google.com](#)

©2008 - [Privacy](#)



Raymond Wong

Search

[Advanced Search](#)
[Preferences](#)

Search: ☒ the web ☐ pages from Hong Kong

Web

Results 1 - 10 of about 354,000 for **Raymond Wong**. (0.06 seconds)

[RAYMONDWONG.COM](#)

[www.raymondwong.com/](#) - 1k - [Cached](#) - [Similar pages](#)

[Raymond Wong Studio](#)

[www.raymondwongstudio.com/](#) - 2k - [Cached](#) - [Similar pages](#)

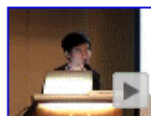
[Raymond Wong Studio](#)

酒類. 其他. 包裝. 快餐. O. I. D. U. T. S. G. N. O. W. D. N. O. M. Y. A. R. 食品. 冰淇淋. 人像. 菜單與食譜. 酒類. 其他. 包裝. 快餐. RAYMOND WONG STUDIO. 關於.

[www.raymondwongstudio.com/ch/main.html](#) - 2k - [Cached](#) - [Similar pages](#)

[More results from www.raymondwongstudio.com »](#)

[Video results for Raymond Wong](#)



[AGDS_Raymond_Wong.mp4](#)

50 min

[video.google.com](#)



[Raymond Wong at PMI HK](#)

[Chapter 10th Anniv ...](#)

1 min 53 sec

[www.youtube.com](#)

[Raymond Wong - Wikipedia, the free encyclopedia](#)

26 Nov 2008 ... **Raymond Wong** may refer to: **Raymond Wong** Yuk Man, radio host and political commentator; **Raymond Wong** Hung Chiu, - Permanent Secretary for ...

[en.wikipedia.org/wiki/Raymond_Wong](#) - 17k - [Cached](#) - [Similar pages](#)

[Raymond Wong Ho-Yin](#)

Raymond Wong in Love Undercover (2002), **Raymond Wong** in Needing You (2000), **Raymond Wong** in Sealed with a Kiss (1999), **Raymond Wong** in The Irresistible ...

[www.lovehkfilm.com/people/wong_raymond2.htm](#) - 32k - [Cached](#) - [Similar pages](#)

[Raymond Chi-Wing Wong \(Raymond Wong\), HKUST CSE](#)

Raymond Chi-Wing Wong is an Assistant Professor in Computer Science and **Raymond Wong**, **Raymond C.-W. Wong**, **Raymond C. W. Wong**, **Raymond C. Wong**, ...

[www.cse.ust.hk/~raywong/](#) - 43k - [Cached](#) - [Similar pages](#)

[Raymond Wong - DramaWiki](#)

7 Oct 2008 ... From DramaWiki. **Raymond Wong** ... Name: 黃浩然 / Wong Ho Yin (Huang Hao Ran); English name: **Raymond Wong**; Profession: Actor ...

[wiki.d-addicts.com/Raymond_Wong](#) - 15k - [Cached](#) - [Similar pages](#)

How to rank the webpages?



Ranking Methods

- HITS Algorithm
- PageRank Algorithm

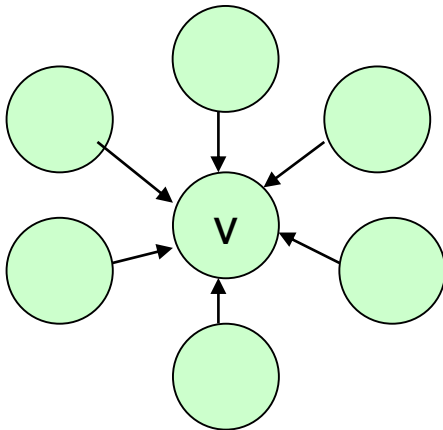


HITS Algorithm

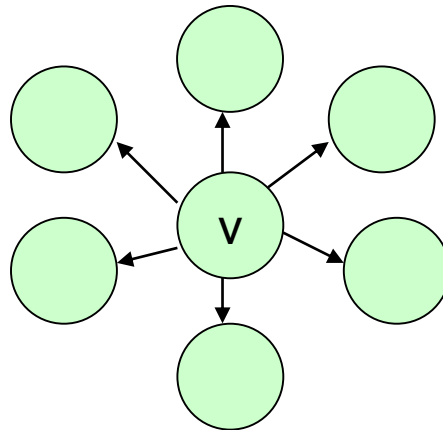
- HITS is a ranking algorithm which ranks “hubs” and “authorities”.

HITS Algorithm

■ Authority



■ Hub



Each page has two weights

1. Authority weight $a(v)$
2. Hub weight $h(v)$

HITS Algorithm

A good hub has many outgoing edges to good authorities

- Each vertex has two weights

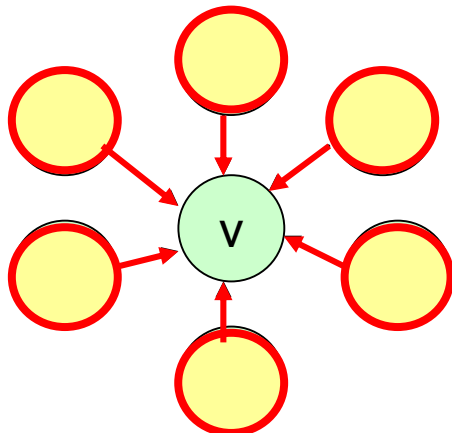
- Authority weight

- Hub weight

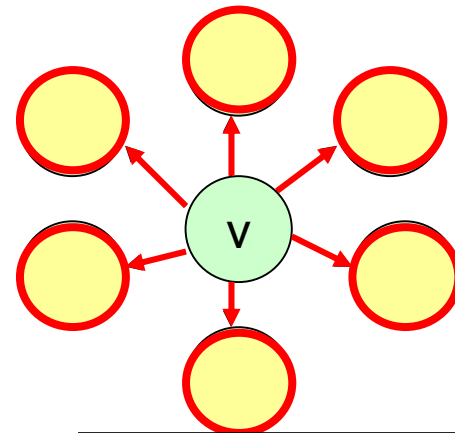
A good authority has many edges from good hubs

Authority Weight

Hub Weight



$$a(v) = \sum_{u \rightarrow v} h(u)$$



$$h(v) = \sum_{v \rightarrow u} a(u)$$



HITS Algorithm

- HITS involves two major steps.
 - Step 1: Sampling Step
 - Step 2: Iteration Step



Step 1 – Sampling Step

- Given a user query with several terms
 - Collect a set of pages that are very relevant – called the **base set**
- How to find **base set**?
 - We retrieve all webpages that contain the query terms. The set of webpages is called the **root set**.
 - Next, find the **link pages**, which are either pages with a hyperlink to some page in the root set or some page in the root set has hyperlink to these pages
 - All pages found form the **base set**.



HITS Algorithm

- HITS involves two major steps.
 - Step 1: Sampling Step
 - Step 2: Iteration Step



Step 2 – Iteration Step

- **Goal:** to find the base pages that are good hubs and good authorities

Step 2 – Iteration St

Adjacency matrix M

$$= \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix}$$

N: Netscape
MS: Microsoft
A: Amazon.com

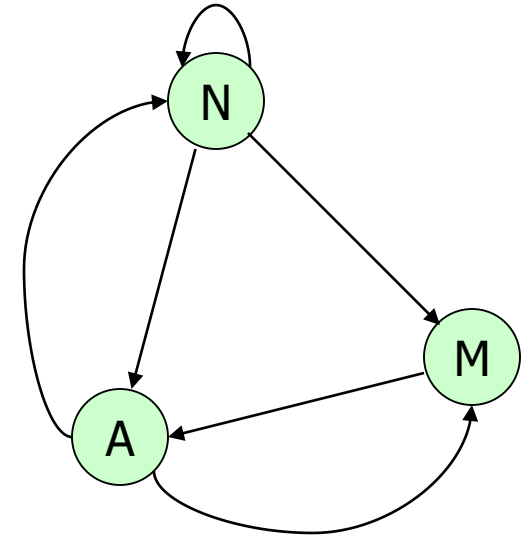
$$h(N) = a(N) + a(MS) + a(A)$$

$$h(MS) = a(A)$$

$$h(A) = a(N) + a(MS)$$

$$\begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix}$$

$$\vec{h} = M\vec{a}$$



$$\vec{h} = \begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} \quad \vec{a} = \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix} \quad 12$$

Step 2 – Iteration St

$$\text{Adjacency matrix } M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix}$$

N: Netscape
MS: Microsoft
A: Amazon.com

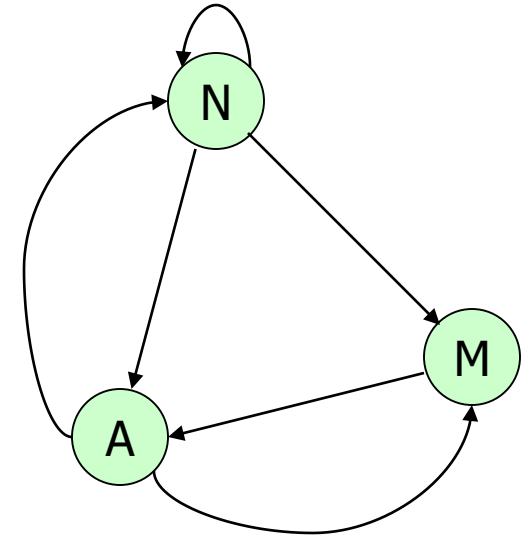
$$a(N) = h(N) + h(A)$$

$$a(MS) = h(N) + h(A)$$

$$a(A) = h(N) + h(MS)$$

$$\begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix}$$

$$\vec{a} = M^T \vec{h}$$



$$\vec{h} = \begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} \quad \vec{a} = \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix}$$



Step 2 – Iteration Step

We have

$$\vec{h} = M\vec{a}$$

$$\vec{a} = M^T \vec{h}$$

We derive

$$\vec{h} = MM^T \vec{h}$$

$$\vec{a} = M^T M \vec{a}$$

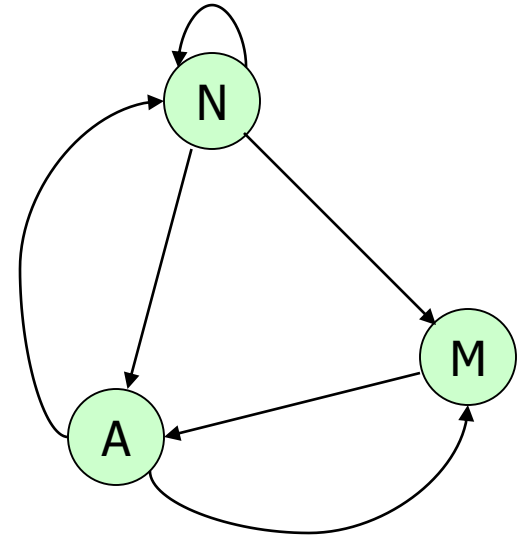
Step 2 – Iteration Step

$$M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$M^T = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix}$$

$$MM^T = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{pmatrix} \end{matrix}$$

$$M^T M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \end{matrix}$$



Hub

$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.5 \\ 0.402 \\ 1.098 \end{pmatrix}$$

Step 2 – Iteration Step

$$\vec{h} = MM^T \vec{h}$$

$$MM^T = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{pmatrix} \end{matrix}$$

Hub (non-normalized)

Iteration No.	1	2	3	4	5	6	7
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	6 2 4	7 2 5	7.071 1.929 5.143	7.091 1.909 5.182	7.096 1.904 5.192	7.098 1.902 5.195

Hub (normalized)

The sum of all elements in the vector = 3

Iteration No.	1	2	3	4	5	6	7
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	1.5 0.5 1	1.5 0.429 1.071	1.5 0.409 1.091	1.5 0.404 1.096	1.5 0.402 1.098	1.5 0.402 1.098

Step 2 – Iteration Step

$$\vec{a} = M^T M \vec{a}$$

$$M^T M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \end{matrix}$$

Hub

$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.5 \\ 0.402 \\ 1.098 \end{pmatrix}$$

Authority

$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.098 \\ 1.098 \\ 0.804 \end{pmatrix}$$

Authority (non-normalized)

Iteration No.	1	2	3	4	5	6	7
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	5 5 4	5.143 5.143 3.857	5.182 5.182 3.818	5.192 5.192 3.808	5.195 5.195 3.805	5.196 5.196 3.804

Authority (normalized)

The sum of all elements in the vector = 3

Iteration No.	1	2	3	4	5	6	7
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	1.071 1.071 0.857	1.091 1.091 0.818	1.096 1.096 0.808	1.098 1.098 0.805	1.098 1.098 0.804	1.098 1.098 0.804



How to Rank

Hub

$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.5 \\ 0.402 \\ 1.098 \end{pmatrix}$$

Authority

$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.098 \\ 1.098 \\ 0.804 \end{pmatrix}$$

- Many ways
 - Rank in descending order of hub only
 - Rank in descending order of authority only
 - Rank in descending order of the value computed from both hub and authority (e.g., the sum of the hub value and the authority value)



Ranking Methods

- HITS Algorithm
- PageRank Algorithm



PageRank Algorithm (Google)

- Disadvantage of HITS:
 - Since there are two concepts, namely hubs and authorities, we do not know which concept is more important for ranking.
- Advantage of PageRank:
 - PageRank involves only one concept for ranking



PageRank Algorithm (Google)

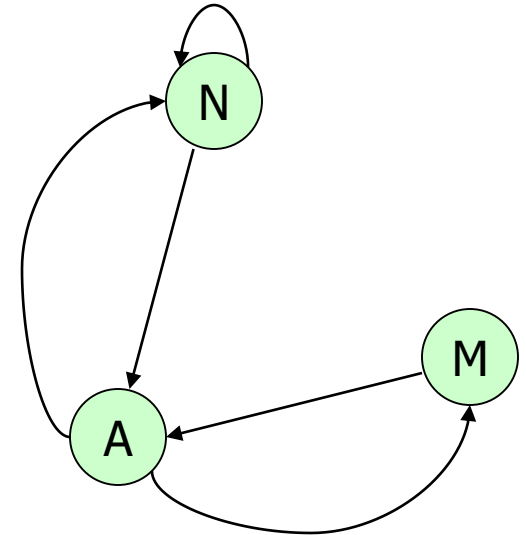
- PageRank Algorithm makes use of **Stochastic approach** to rank the pages

PageRank Algorithm (Google)

Stochastic matrix M

$$= \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0.5 & 1 & 0 \end{pmatrix} \end{matrix}$$

N: Netscape
MS: Microsoft
A: Amazon.com



PageRank Algorithm (Google)

$$\vec{r} = M\vec{r}$$

$$M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0.5 & 1 & 0 \end{pmatrix} \end{matrix}$$

Page Rank

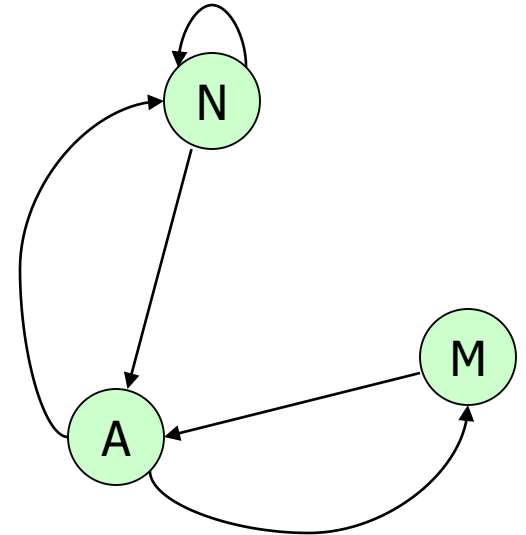
Iteration No.	1	2	3	4	5	..	33
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	1 0.5 1.5	1.25 0.75 1	1.125 0.5 1.375	1.156 0.531 1.313	...	1.20 0.60 1.20

Microsoft (MS) is quite upset with this result.
Microsoft decides to link only to itself from now on.

PageRank Algorithm (Google)

Stochastic matrix M

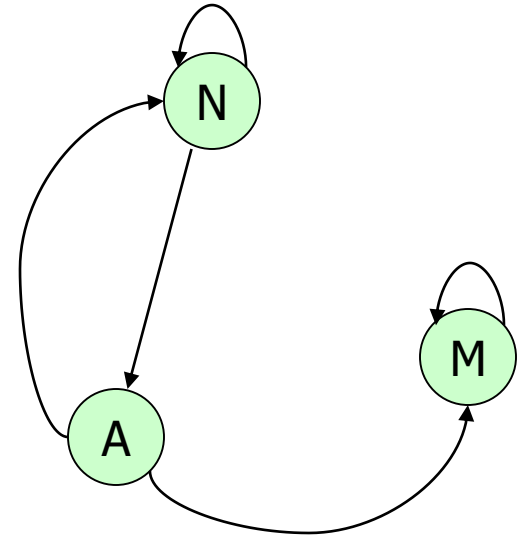
$$= \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0.5 & 1 & 0 \end{pmatrix} \end{matrix}$$



PageRank Algorithm (Google)

Stochastic matrix M

$$= \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{matrix}$$



PageRank Algorithm (Google)

$$\vec{r} = M\vec{r}$$

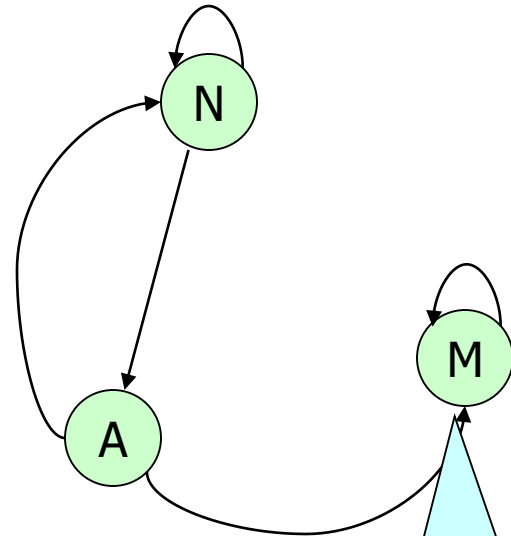
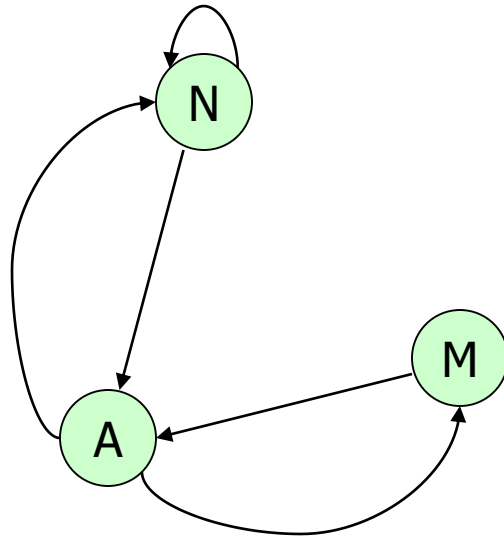
$$M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{matrix}$$

Page Rank

Iteration No.	1	2	3	4	5	..	40
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	1 1.5 0.5	0.75 1.75 0.5	0.625 2 0.375	0.5 2.188 0.313	...	0 3 0

Microsoft (MS) is happy. It is the most important now.
Others is not happy.

PageRank Algorithm (Google)



Spider trap: a group of one or more pages that have no links out of the group will eventually accumulate all the importance of the web.

Microsoft has become a spider trap.

How to solve it?



PageRank Algorithm (Google)

$$\vec{r} = M\vec{r}$$

$$M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$\vec{r} = 0.8 \times M\vec{r} + \vec{c}$$

$$\vec{c} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$$

PageRank Algorithm (Google)

$$\vec{r} = 0.8 \times M\vec{r} + \vec{c}$$

$$M = \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{matrix}$$

Page Rank

Iteration No.	1	2	3	4	5	..	20
$\begin{pmatrix} N \\ MS \\ A \end{pmatrix}$	1 1 1	1 1.4 0.6	0.84 1.56 0.6	0.776 1.688 0.536	0.725 1.765 0.510	...	0.636 1.909 0.455

We have a more reasonable distribution of importance than before.