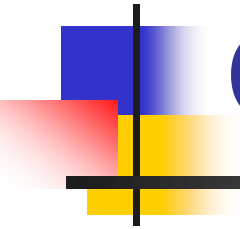
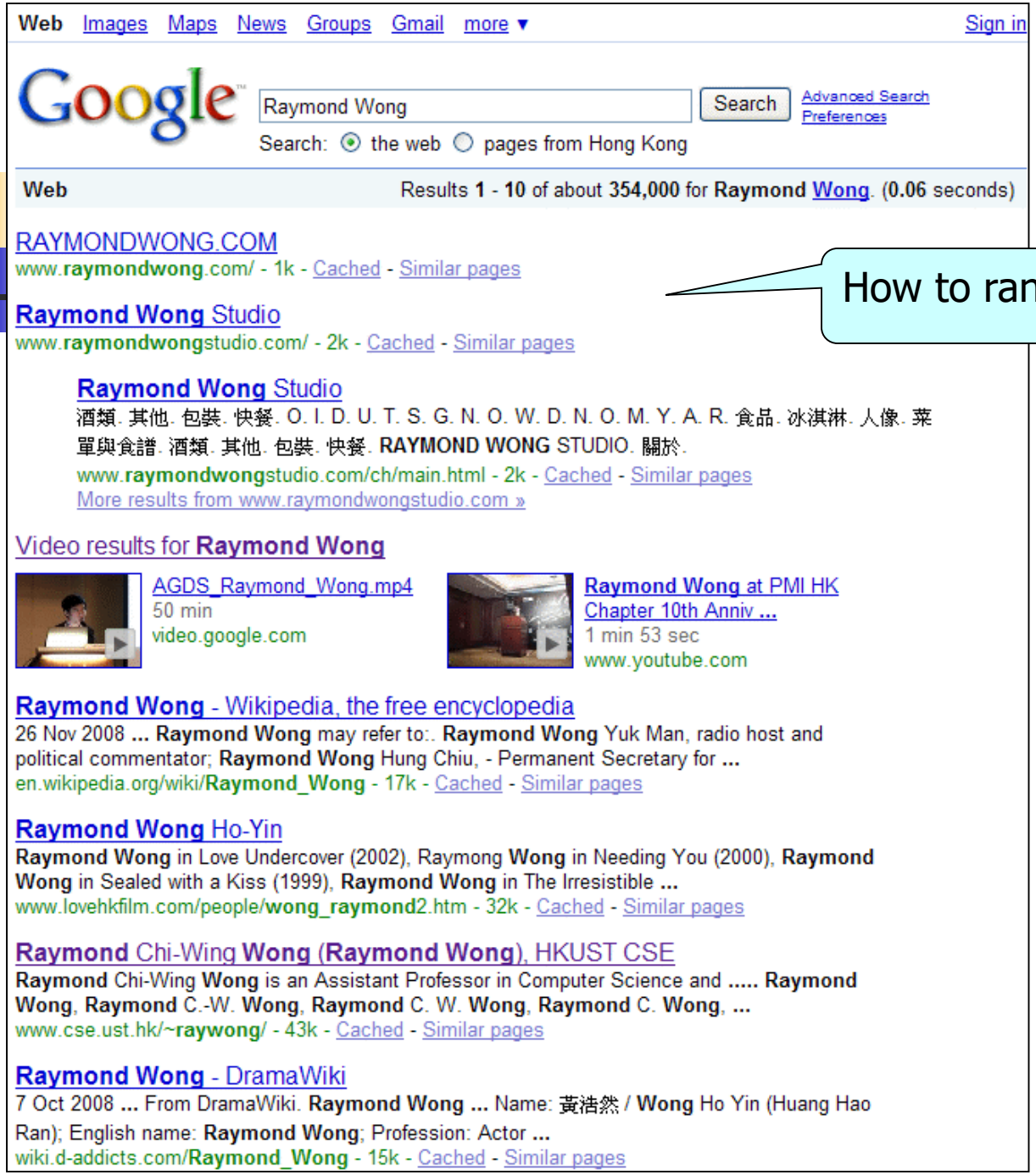


COMP5331

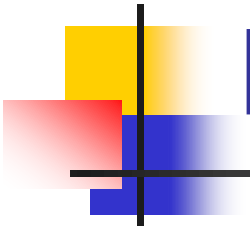
Web databases

Prepared by Raymond Wong
Presented by Raymond Wong
raywong@cse

Web Databases

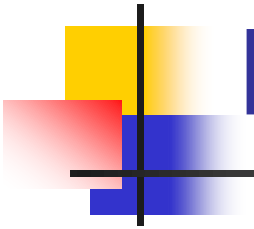


How to rank the webpages?



Ranking Methods

- HITS Algorithm
- PageRank Algorithm

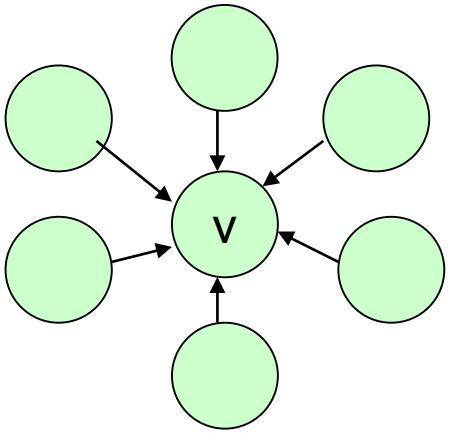


HITS Algorithm

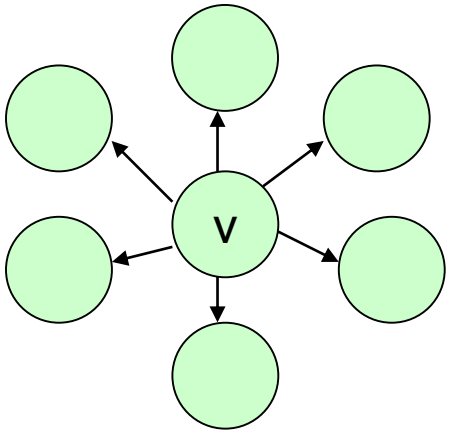
- HITS is a ranking algorithm which ranks “hubs” and “authorities”.

HITS Algorithm

- Authority



- Hub



Each page has
two weights
1. Authority weight a(v)
2. Hub weight h(v)

HITS Algorithm

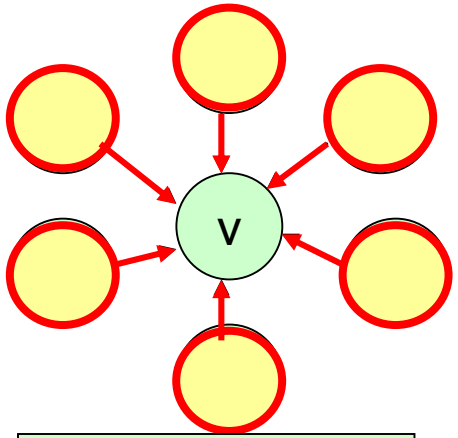
- Each vertex has two weights
  - Authority weight
  - Hub weight

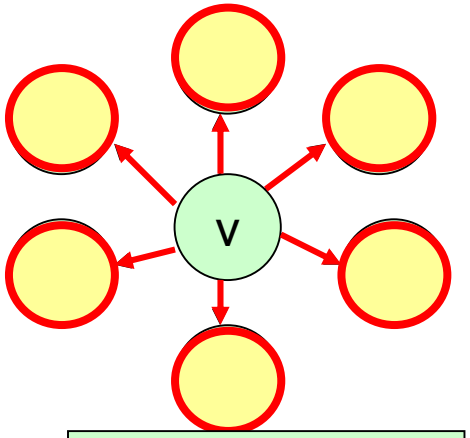
A good hub has many outgoing edges to good authorities

A good authority has many edges from good hubs

Authority Weight

Hub Weight



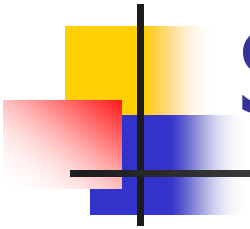


$a(v) = \sum_{u \to v} h(u)$

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

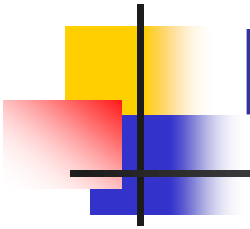
HITS Algorithm

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



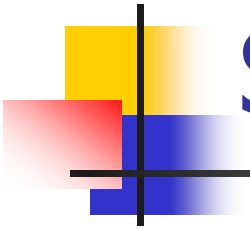
Step 1 – Sampling Step

- 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

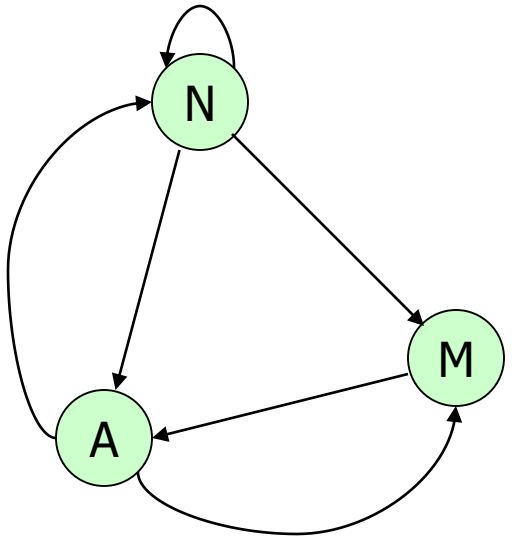
- **Goal:** to find the base pages that are good hubs and good authorities

Step 2 – Iteration St

Adjacency matrix M

$$= \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

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)

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

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

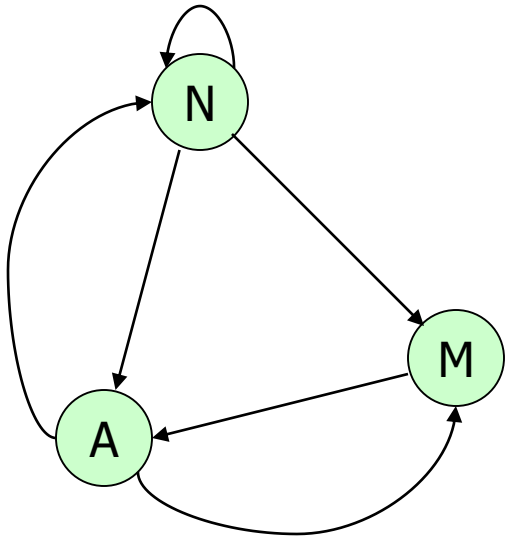
$$\vec{h} = \begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} \quad \vec{a} = \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix}$$

Step 2 – Iteration St

Adjacency matrix M

$$= \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

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

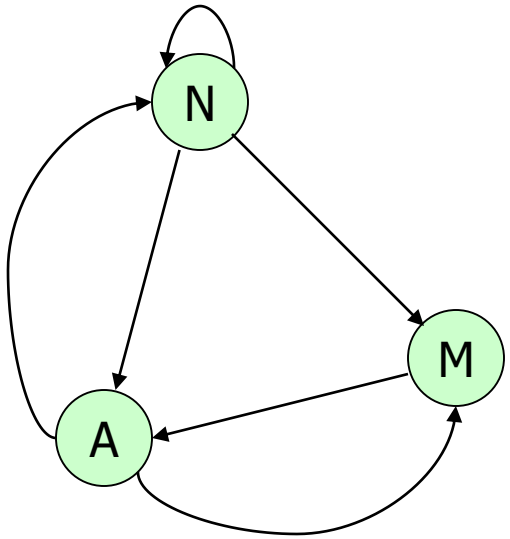
$$\vec{h} = MM^T\vec{h}$$

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

Step 2 – Iteration Step

$$M = \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{array}$$

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



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

$$M^TM = \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \end{array}$$

Hub

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

Step 2 – Iteration Step

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

$MM^T =$

	N	MS	A
N	3	1	2
MS	1	1	0
A	2	0	2

Hub (non-normalized)

Iteration No.	1	2	3	4	5	6	7
N	1	6	7	7.071	7.091	7.096	7.098
MS	1	2	2	1.929	1.909	1.904	1.902
A	1	4	5	5.143	5.182	5.192	5.195

Hub (normalized)

The sum of all elements in the vector = 3

Iteration No.	1	2	3	4	5	6	7
N	1	1.5	1.5	1.5	1.5	1.5	1.5
MS	1	0.5	0.429	0.409	0.404	0.402	0.402
A	1	1	1.071	1.091	1.096	1.098	1.098

Step 2 – Iteration Step

Hub

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

Authority

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

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

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

Authority (non-normalized)

Iteration No.	1	2	3	4	5	6	7
N	1	5	5.143	5.182	5.192	5.195	5.196
MS	1	5	5.143	5.182	5.192	5.195	5.196
A	1	4	3.857	3.818	3.808	3.805	3.804

Authority (normalized)

The sum of all elements in the vector = 3

Iteration No.	1	2	3	4	5	6	7
N	1	1.071	1.091	1.096	1.098	1.098	1.098
MS	1	1.071	1.091	1.096	1.098	1.098	1.098
A	1	0.857	0.818	0.808	0.805	0.804	0.804

How to Rank

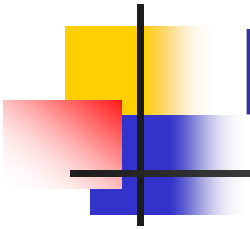
Hub

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

Authority

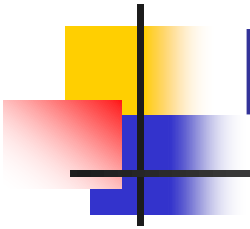
$$\begin{pmatrix} N \\ MS \\ A \end{pmatrix} = \begin{pmatrix} 1.098 \\ 1.098 \\ 0.804 \end{pmatrix}$$

- 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)



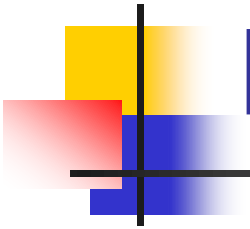
Ranking Methods

- 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)

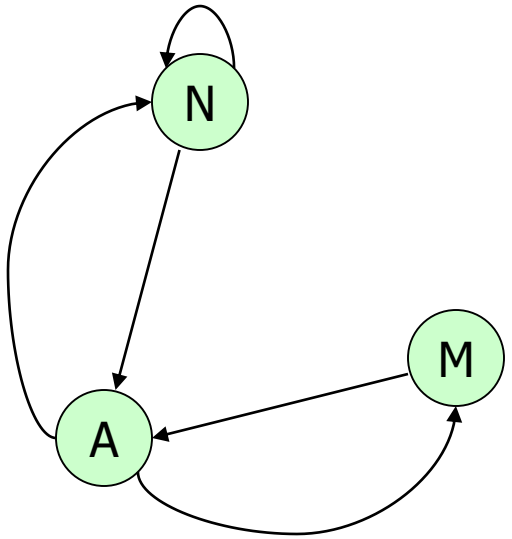
- PageRank Algorithm makes use of **Stochastic approach** to rank the pages

PageRank Algorithm (Google)

N: Netscape
MS: Microsoft
A: Amazon.com

Stochastic matrix M

$$= \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0.5 & 1 & 0 \end{pmatrix} \end{array}$$



PageRank Algorithm (Google)

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

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

Page Rank

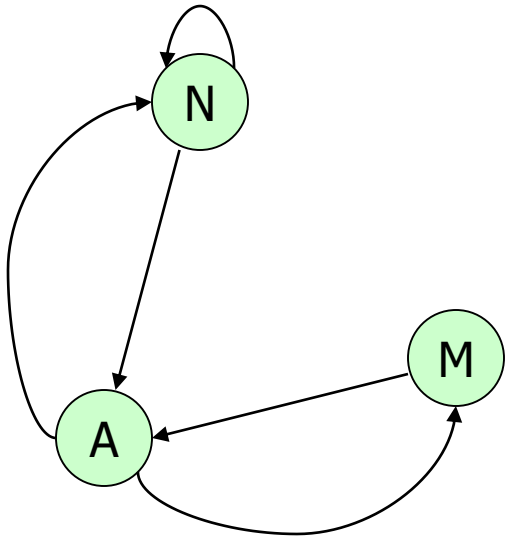
Iteration No.	1	2	3	4	5	..	33
N	1	1	1.25	1.125	1.156	…	1.20
MS	1	0.5	0.75	0.5	0.531	…	0.60
A	1	1.5	1	1.375	1.313	…	1.20

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

PageRank Algorithm (Google)

Stochastic matrix M

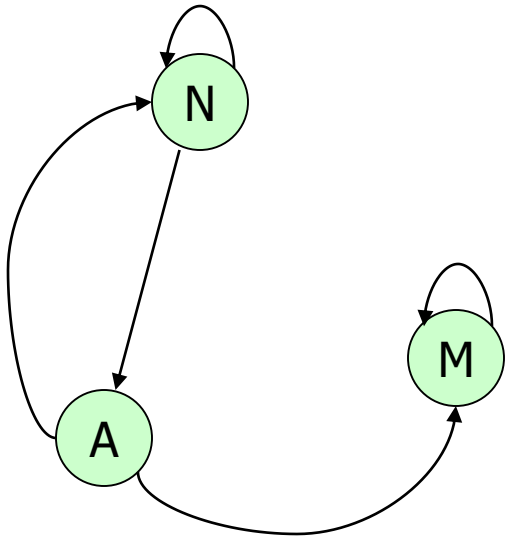
$$= \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 0 & 0.5 \\ 0.5 & 1 & 0 \end{pmatrix} \end{array}$$



PageRank Algorithm (Google)

Stochastic matrix M

$$= \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{array}$$



PageRank Algorithm (Google)

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

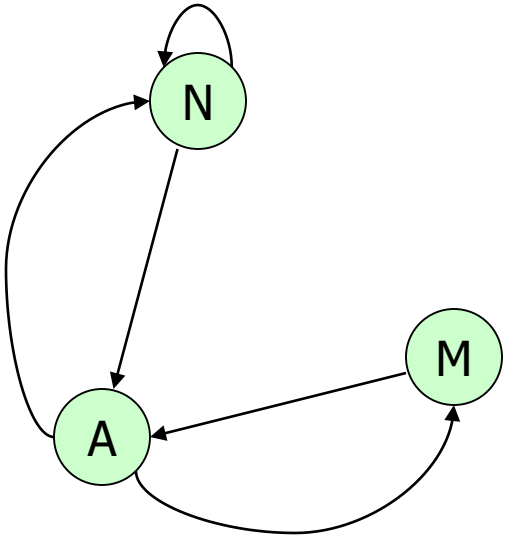
$$M = \begin{array}{c} \\ N \\ MS \\ A \end{array} \begin{array}{c} \begin{array}{ccc} N & MS & A \end{array} \\ \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0 & 0 \end{pmatrix} \end{array}$$

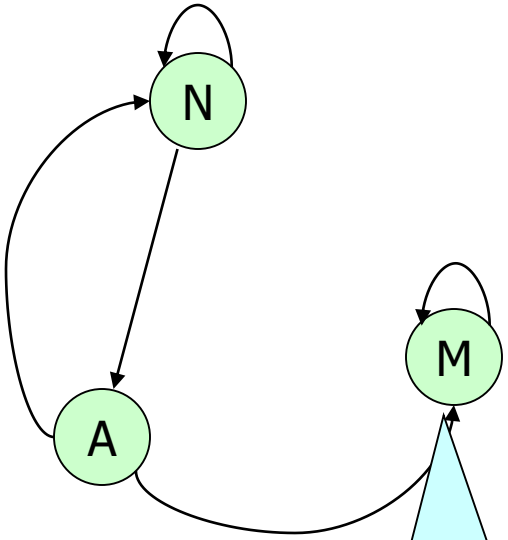
Page Rank

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)

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

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

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

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

PageRank Algorithm (Google)

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

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

Page Rank

Iteration No.	1	2	3	4	5	..	20
N	1	1	0.84	0.776	0.725	…	0.636
MS	1	1.4	1.56	1.688	1.765	…	1.909
A	1	0.6	0.6	0.536	0.510	…	0.455

We have a more reasonable distribution of importance than before.