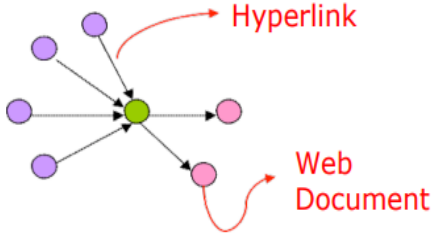
Web Structure Mining

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages



Web Graph Structure

Web Structure Mining is the process of discovering structure information from the Web

- This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level
- The research at the hyperlink level is also called *Hyperlink Analysis*

PageRank algorithm

What is the oiginal problem? We want to rank websites in their search engine results

There are two popular algorithms to rank web pages by popularity

1.) **HITS** – Hypertext Induced Topic Search

2.) **PageRank** algorithm

Page Rank Algorithm

- Page Rank Algorithm

HITS Algorithm

- Hyperlink Induced Topic Search (HITS) is an algorithm used in link analysis.
- It could discover and rank the webpages relevant for a particular search.
- The idea of this algorithm originated from the fact that an ideal website should link to other relevant sites and also being linked by other important sites.

HITS Algorithm

- Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg.
- This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.
- HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

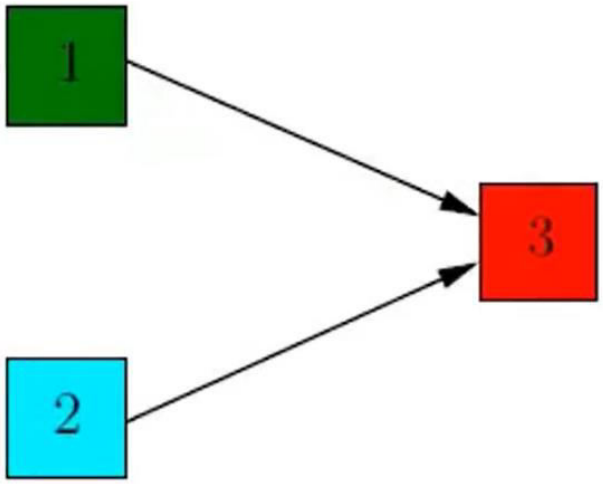
- HITS uses hubs and authorities to define a recursive relationship between webpages.
  - Authority: A node is high-quality if many high-quality nodes link to it
  - Hub: A node is high-quality if it links to many high-quality nodes
- Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
- Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

HITS ALGORITHM

- Algorithm Steps
  - Initialize the hub and authority of each node with a value of 1
  - For each iteration, update the hub and authority of every node in the graph
  - The new authority is the sum of the hub of its parents
  - The new hub is the sum of the authority of its children
  - Normalize the new authority and hub

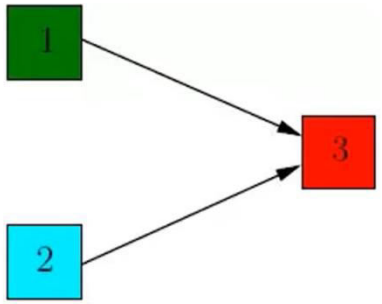
HITS Algorithm

- Compute the Hub and Authority weights for the following graph.



HITS Algorithm

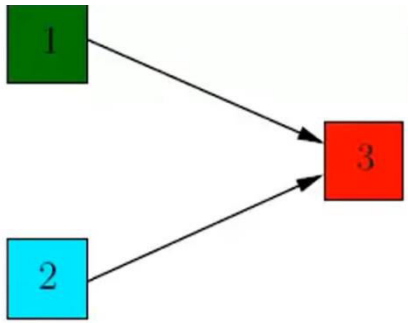
Step 1: Find the adjacency matrix of the graph



	1	2	3
1	0		
2			
3			

Because there's no link between 1 and 2

Step 1: Find the adjacency matrix of the graph

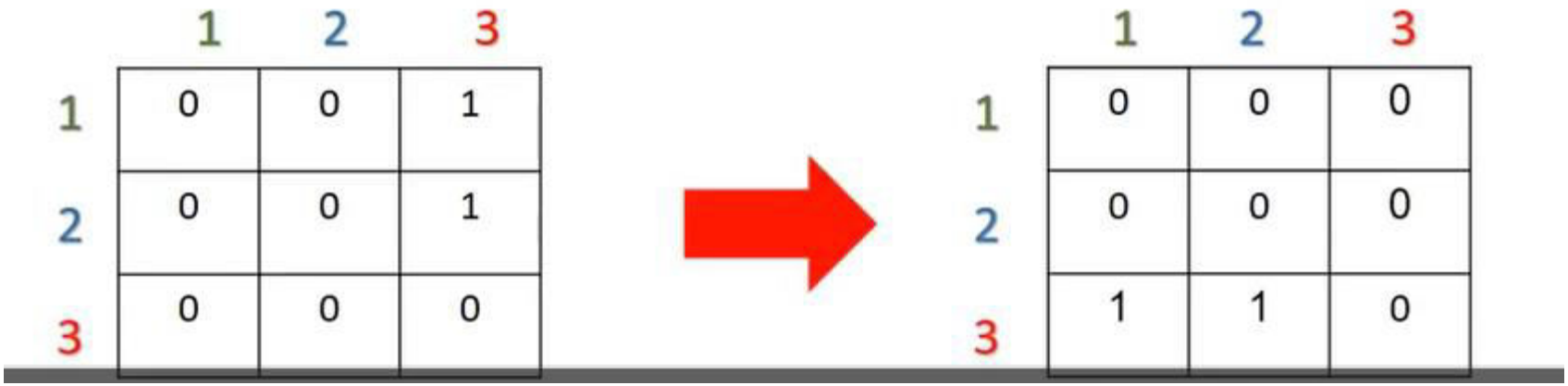


	1	2	3
1	0	0	1
2	0	0	
3	0	0	

Because there's a link between 1 and 3

HITS Algorithm

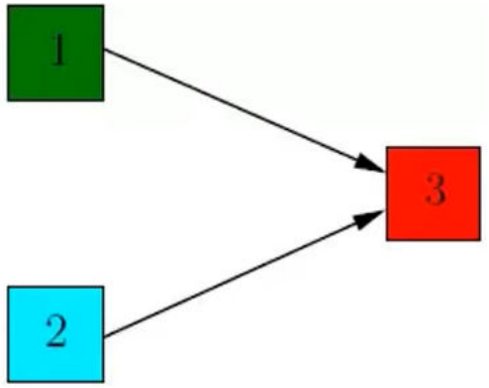
- Find transpose of the matrix



HITS Algorithm

$A^t =$

	1	2	3
1	0	0	0
2	0	0	0
3	1	1	0

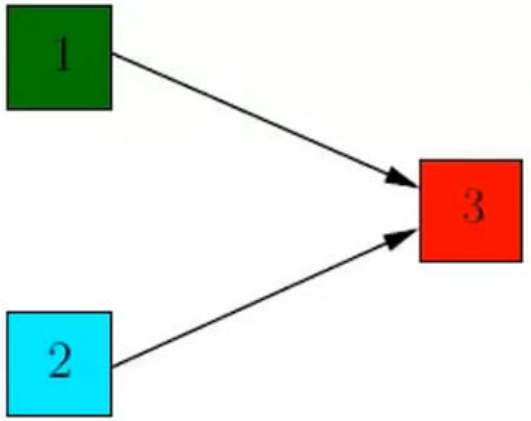


Assume the initial hub weight vector is: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

HITS Algoirthm

$A^t =$

	1	2	3
1	0	0	0
2	0	0	0
3	1	1	0



Assume the initial hub weight vector is: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

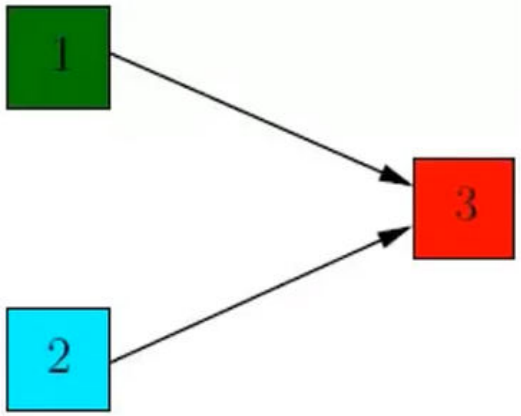
compute the authority weight vector

$$v = A^t \cdot u = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

HITS Algorithm

$A^t =$

	1	2	3
1	0	0	0
2	0	0	0
3	1	1	0



Assume the initial hub weight vector is: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

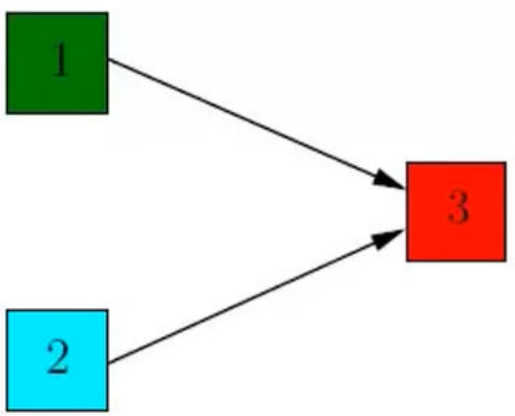
compute the authority weight vector

$$v = A^t \cdot u = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

HITS Algorithm

$A^t =$

	1	2	3
1	0	0	0
2	0	0	0
3	1	1	0

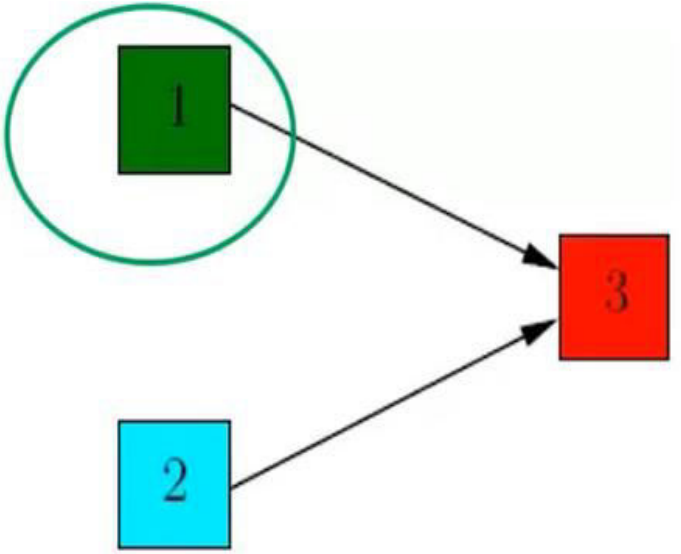


Assume the initial hub weight vector is: $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Then update hub weight:

$$u = A \cdot v = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

HITS Algorithm



Results

hub weights

$$\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

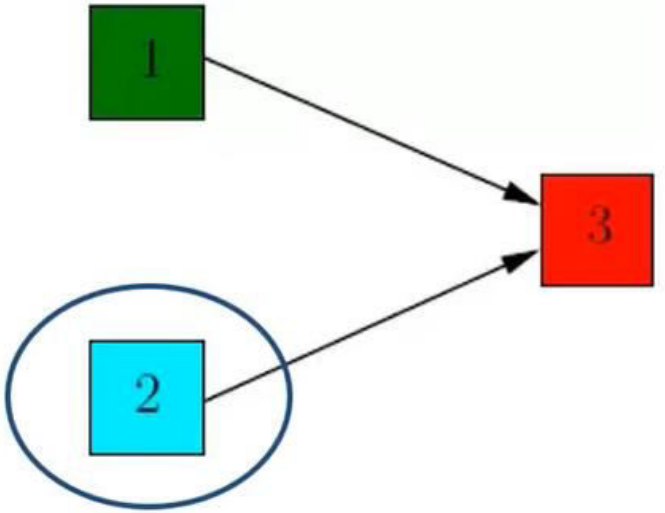
authority weights

$$\begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

nodes 1 is a hub since $2 > 0$

HITS Algorithm

Results



hub weights

$$\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

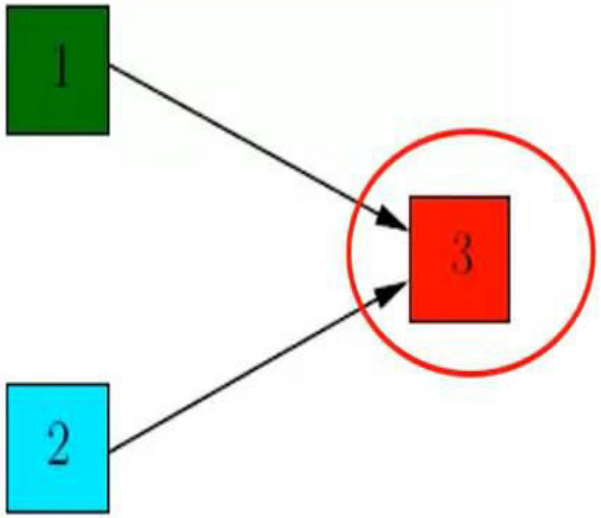
authority weights

$$\begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

nodes 2 is a hub since $2 > 0$

HITS Algorithm

Results



hub weights

$$\begin{bmatrix} 2 \\ 2 \\ 0 \end{bmatrix}$$

authority weights

$$\begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$$

node 3 is the most authoritative since $0 < 2$

Example to solve

To identify the best hub and authority for the given adjacency matrix. Calculate the hubs and authority score using hits algorithm for k = 3. The adjacency matrix is

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Example to solve

0	1	1	1
0	0	1	1
1	0	0	1
0	0	0	1

Adjacency matrix with nodes

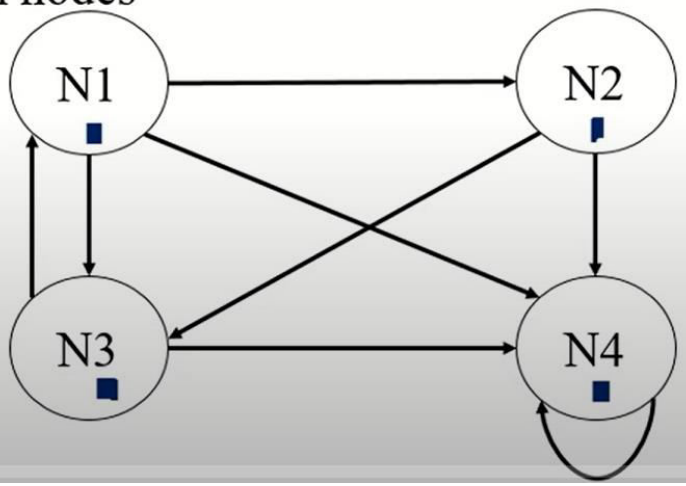
	N1	N2	N3	N4
N1	0	1	1	1
N2	0	0	1	1
N3	1	0	0	1
N4	0	0	0	1

Example to solve

Adjacency matrix with nodes

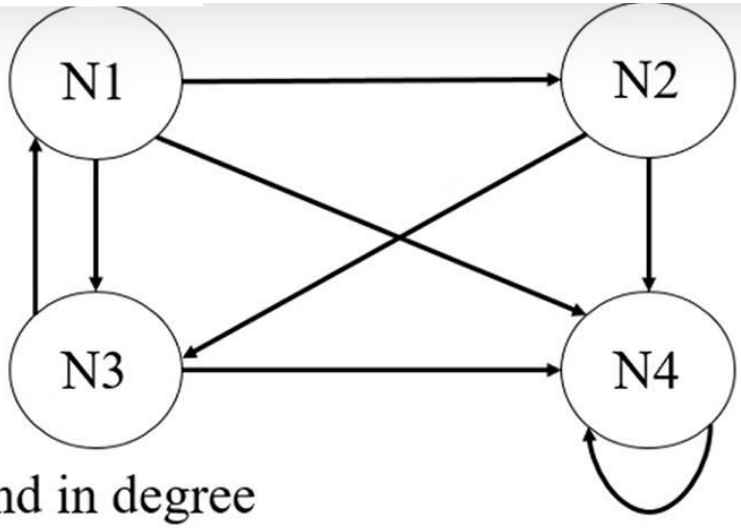
	N1	N2	N3	N4
N1	0	1	1	1
N2	0	0	1	1
N3	1	0	0	1
N4	0	0	0	1

Graph with nodes



Example to solve

Graph with nodes



Ranks using out degree and in degree

Nodes	Out-degree(Hub)	In-degree(Authority)
N1	3	1
N2	2	1
N3	2	2
N4	1	4

Example to solve

Ranks using out degree and in degree

Nodes	Out-degree(Hub)	In-degree(Authority)
N1	3	1
N2	2	1
N3	2	2
N4	1	4

HUB: N1, N2, N3 {TIE}, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

Example to solve

Adjacency matrix , A with nodes

	N1	N2	N3	N4
N1	0	1	1	1
N2	0	0	1	1
N3	1	0	0	1
N4	0	0	0	1

Transpose of Matrix, A^T

	N1	N2	N3	N4
N1	0	0	1	0
N2	1	0	0	0
N3	1	1	0	0
N4	1	1	1	1

Assuming initial hub weight vector, u as $\underline{1}$

Example to solve

Authority weight vector , $v = A^T * u$

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} * \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Authority, v $\begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \end{pmatrix}$

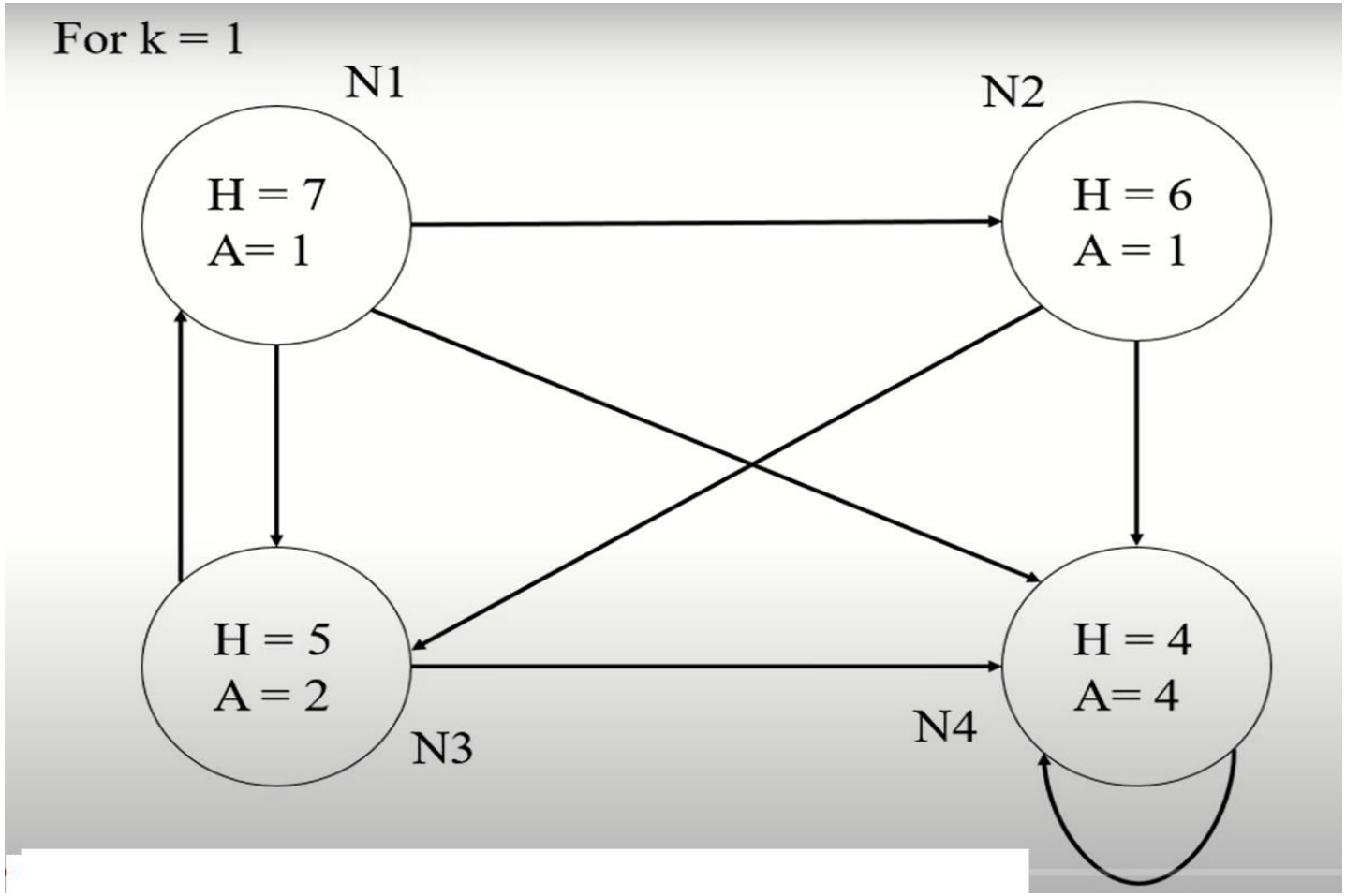
Example to solve

Updated Hub weight vector , u = A * v

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} * \begin{pmatrix} 1 \\ 1 \\ 2 \\ 4 \end{pmatrix}$$

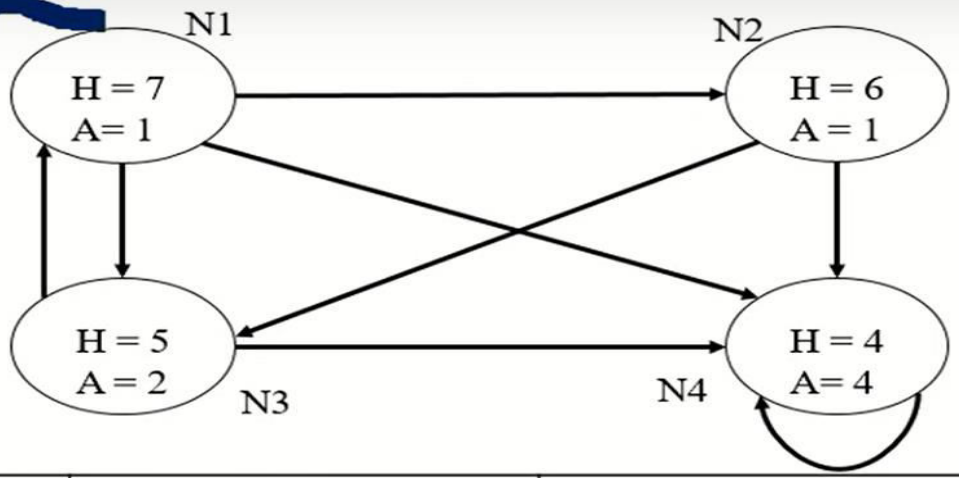
Hub, u $\begin{pmatrix} 7 \\ 6 \\ 5 \\ 4 \end{pmatrix}$

Example to solve



Example to solve

For k = 1



Nodes	Hub	Authority
N1	7	1
N2	6	1
N3	5	2
N4	4	4

Example to solve

For k = 1

Nodes	Hub	Authority
N1	7	1
N2	6	1
N3	5	2
N4	4	4

HUB: N1, N2, N3, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

Example to Solve

For k = 1

Nodes	Hub	Authority
N1	7	1
N2	6	1
N3	5	2
N4	4	4

HUB: N1, N2, N3, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

Calculate new Authority from K = 1

$$v_1 = 1^2 + 1^2 + 2^2 + 4^2 = 22$$

$$= \frac{1}{\sqrt{22}}, \frac{1}{\sqrt{22}}, \frac{2}{\sqrt{22}}, \frac{4}{\sqrt{22}}$$

$$v_1 = 0.213, 0.213, 0.426, 0.853$$

N1 N2 N3 N4

Example to Solve

For k = 1

Nodes	Hub	Authority
N1	7	1
N2	6	1
N3	5	2
N4	4	4

HUB: N1, N2, N3, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

Calculate new hub from K = 1

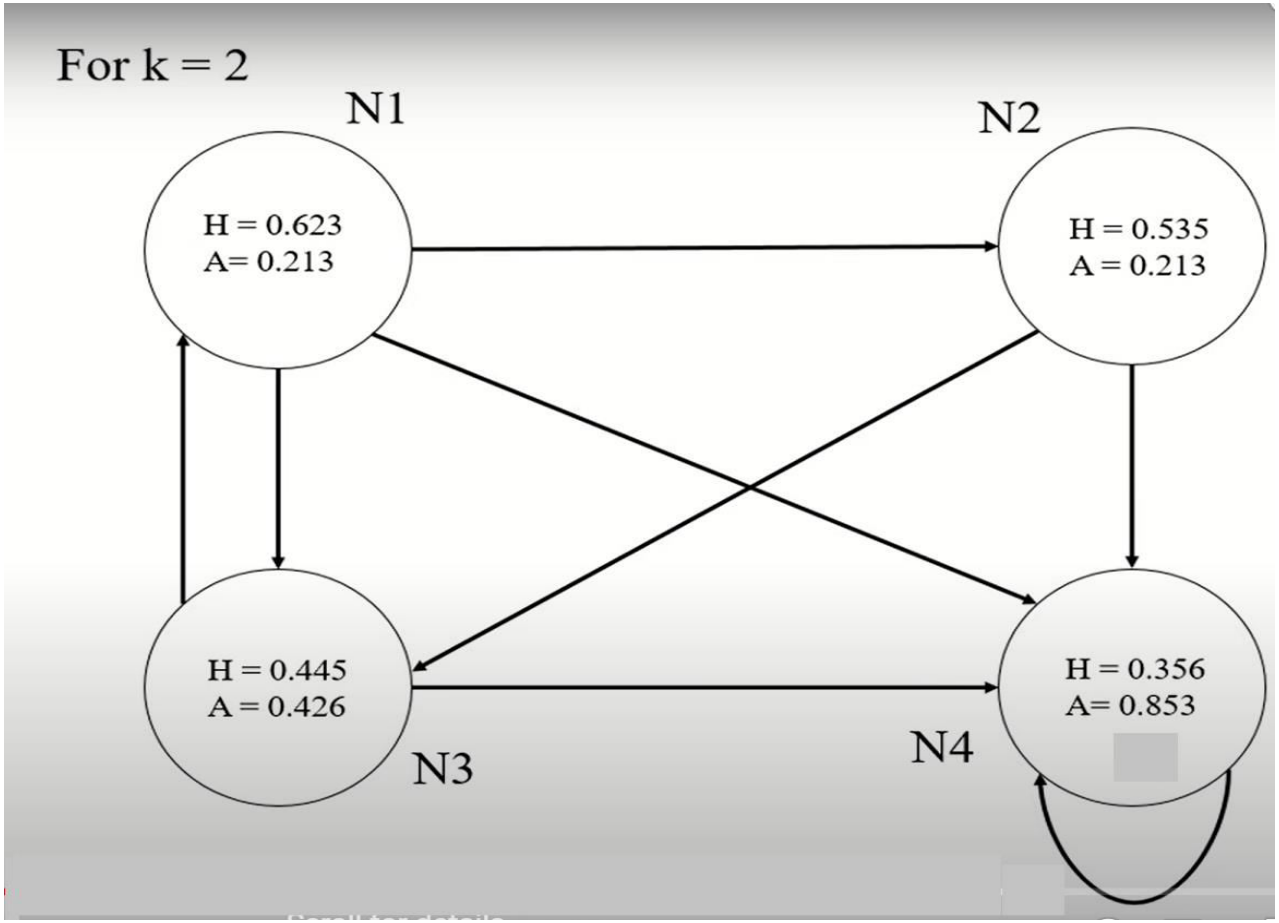
$$u_1 = 7^2 + 6^2 + 5^2 + 4^2 = 126$$

$$= \frac{7}{\sqrt{126}}, \frac{6}{\sqrt{126}}, \frac{5}{\sqrt{126}}, \frac{4}{\sqrt{126}}$$

$$u_1 = 0.623, 0.535, 0.445, 0.356$$

N1 N2 N3 N4

Example to Solve



Example to Solve

For k = 2

Nodes	Hub	Authority
N1	0.623	0.213
N2	0.535	0.213
N3	0.445	0.426
N4	0.356	0.853

HUB: N1, N2, N3, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

Calculate new Authority from K = 2

$$v_1 = 0.213^2 + 0.213^2 + 0.426^2 + 0.853^2 = 0.999$$

$$= \frac{0.213}{\sqrt{0.999}}, \frac{0.213}{\sqrt{0.999}}, \frac{0.426}{\sqrt{0.999}}, \frac{0.853}{\sqrt{0.999}}$$

$$v_1 = 0.213, 0.213, 0.426, 0.853$$

N1 N2 N3 N4

Example to Solve

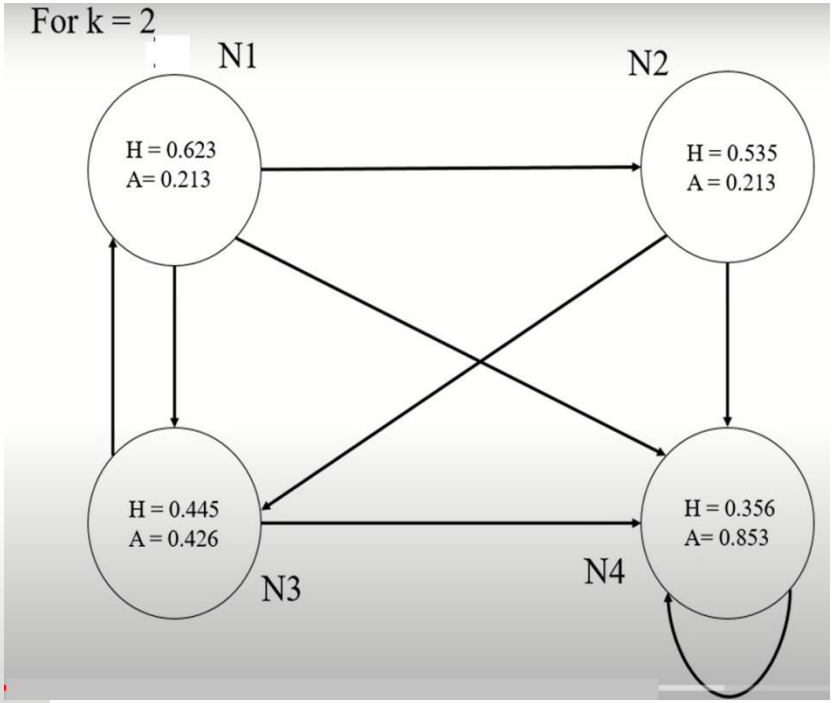
Calculate new hub from K = 2

$$u_1 = 0.623^2 + 0.535^2 + 0.445^2 + 0.356^2 = 0.999$$

$$= \frac{0.623}{\sqrt{0.999}}, \frac{0.535}{\sqrt{0.999}}, \frac{0.445}{\sqrt{0.999}}, \frac{0.356}{\sqrt{0.999}}$$

$$u_1 = 0.623, 0.535, 0.445, 0.356$$

N1 N2 N3 N4



Example to Solve

For k = 2 ≈3

Nodes	Hub	Authority
N1	0.623	0.213
N2	0.535	0.213
N3	0.445	0.426
N4	0.356	0.853

HUB: N1, N2, N3, N4

AUTHORITY: N4, N3, N2, N1 {TIE}

WHERE Hub and authority scores have come to a consistent value FOR K2 AND K3 , ALGORITHM WILL STOP.

Example to Solve