

# Information Retrieval

Venkatesh Vinayakarao

Term: Aug – Sep, 2019  
Chennai Mathematical Institute

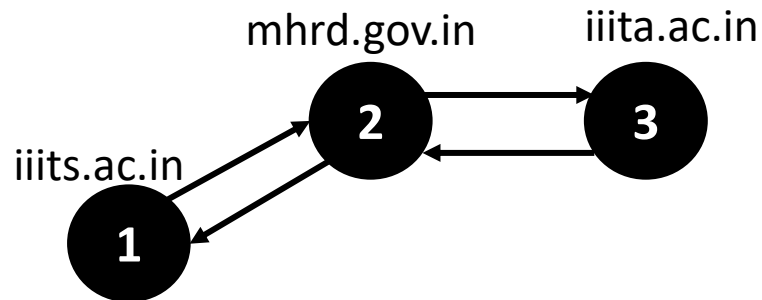


So much of life, it seems to me, is determined by pure randomness.  
– **Sidney Poitier.**



# A Random Surfer Model

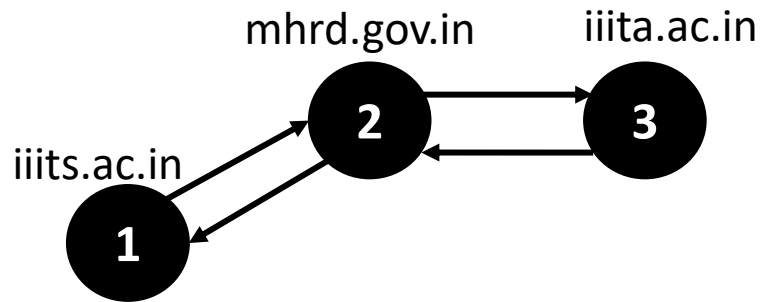
- A random surfer may start from any node with  $1/3$  probability



- Can you represent this graph using Adjacency Matrix?

# A Random Surfer

- May start from any node with 1/3 probability

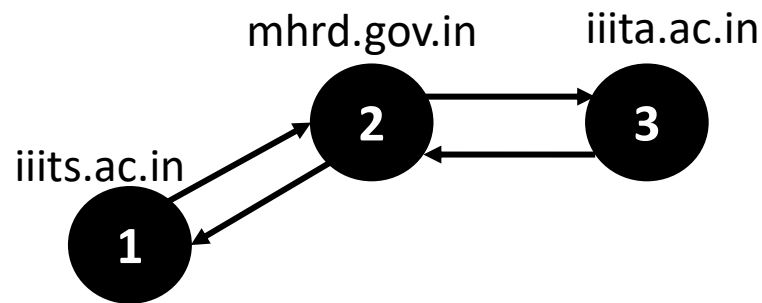


- Can you represent this graph using Adjacency Matrix?

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

# A Random Surfer

- May start from any node with  $1/3$  probability



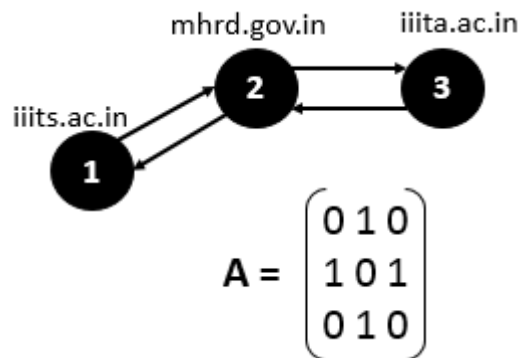
- May also teleport to any node with  $\alpha$  probability

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

How can you compute the transition probabilities?

# Transition Probabilities

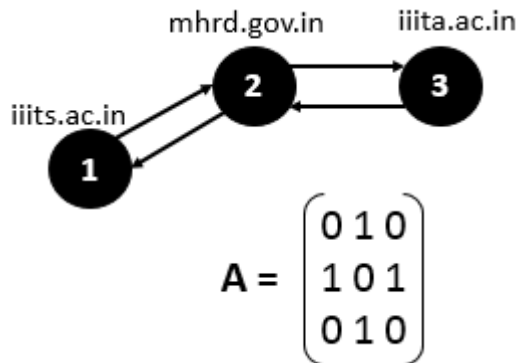
- We can convert the Adjacency Matrix (A) to Transition Probability Matrix (P)



- If the random surfer is at 1,
  - and he did not teleport
    - Probability =  $1 - \alpha$
  - and he teleports
    - he may reach state 3 with probability  $\alpha/3$
    - and may reach state 2 with probability  $\alpha/3$
- Transition Probability from 1 is  $(\alpha/3, (1 - \alpha) + \alpha/3, \alpha/3)$

# Quiz

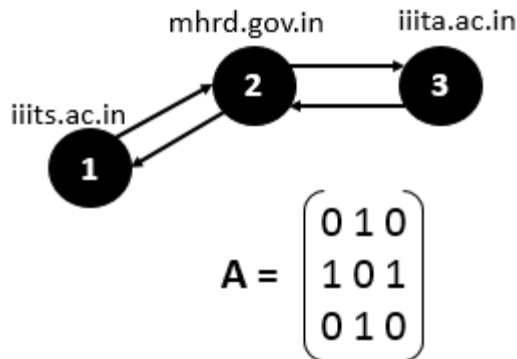
- If the teleportation probability,  $\alpha = 0.5$ , Calculate the transition probability matrix for this network.



$$P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ ?? & ?? & ?? \\ ?? & ?? & ?? \end{pmatrix}$$

# Quiz

- If the teleportation probability,  $\alpha = 0.5$ , Calculate the transition probability matrix (P) for this network.

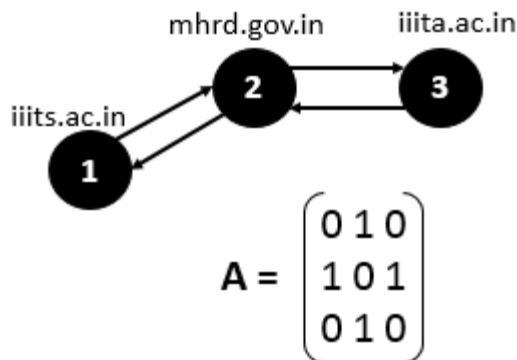


$$P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ ?? & ?? & ?? \\ ?? & ?? & ?? \end{pmatrix}$$

$$P = \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix}$$

# Which page is more popular?

- If a random surfer at 1 can reach (1,2,3) with probabilities (1/6, 2/3, 1/6), where will he end up in the next time slot if chooses to continue his walk?



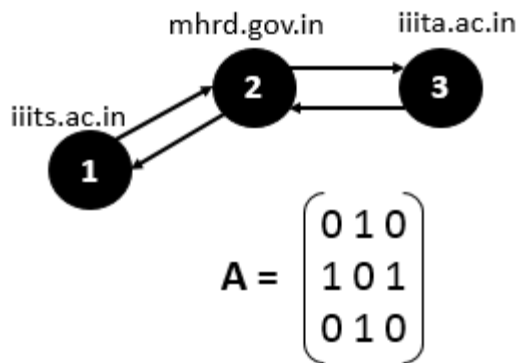
$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

$$P = (1/6, 2/3, 1/6) \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = (??, ??, ??)$$



# Which page is more popular?

- If a random surfer at 1 can reach (1,2,3) with probabilities (1/6, 2/3, 1/6), where will he end up in the next time slot if chooses to continue his walk?



$$P = (1/6, 2/3, 1/6) \begin{pmatrix} 1/6 & 2/3 & 1/6 \\ 5/12 & 1/6 & 5/12 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} = (1/3, 1/3, 1/3)$$

# Steady State Probability

- If the random surfer keeps walking, the probabilities tend to converge!
  - Since we have an **Ergodic Markov Chain!**
    - A markov chain is **ergodic** if every state is reachable from every other state (not necessarily in a single jump).
- In our case, we should get  $(5/18, 8/18, 5/18)$
- So, page 2 gets the highest rank.

# HITS

*Hyperlink-Induced Topic Search.*

Another approach to finding popular pages

# Hubs and Authorities

Topic specific page rank can be useful!



Query: I wish to learn about **leukemia**

**1. Nature.com - Leukemia News**

📍 London  
**About Blog** Latest news and research from Nature.com on the topic of Leukaemia  
**Frequency** about 4 posts per week  
**Blog** [nature.com/subjects/leukaemia](https://nature.com/subjects/leukaemia)  
Facebook fans 907,612. Twitter followers 1,574,259.

View Latest Posts ▾

Subscribe newsletter

Enter email   OR

**2. St. Baldrick's Foundation - Childhood Cancer Research Foundation**

📍 Global  
**About Blog** St. Baldrick's is a childhood cancer charity funding the most promising cures for kids with cancer. Read St. Baldrick's blog to learn more about childhood cancer, specifically on Leukemia.

Hub Page

[https://blog.feedspot.com/leukemia\\_blogs/](https://blog.feedspot.com/leukemia_blogs/)

**NIH NATIONAL CANCER INSTITUTE**

1-800-4-CANCER Live Chat Publications Dictionary Español

ABOUT CANCER **CANCER TYPES** RESEARCH GRANTS & TRAINING NEWS & EVENTS ABOUT NCI search

Home > Cancer Types

**Leukemia—Patient Version**

**OVERVIEW**

Leukemia is a broad term for cancers of the blood cells. The type of leukemia depends on the type of blood cell that becomes cancer and whether it grows quickly or slowly. Leukemia occurs most often in adults older than 55, but it is also the most common cancer in children younger than 15. Explore the links on this page to learn more about the types of leukemia plus treatment, statistics, research, and clinical trials.

**TREATMENT**

**PDQ Treatment Information for Patients**

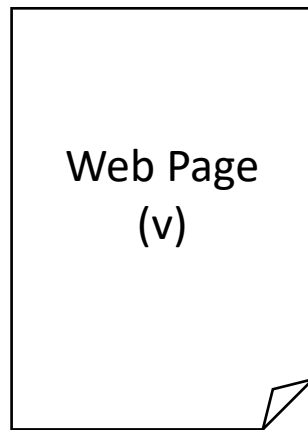
- Adult Acute Lymphoblastic Leukemia Treatment
- Adult Acute Myeloid Leukemia Treatment
- Chronic Lymphocytic Leukemia Treatment
- Chronic Myelogenous Leukemia Treatment
- Hairy Cell Leukemia Treatment
- Childhood Acute Lymphoblastic Leukemia Treatment
- Childhood Acute Myeloid Leukemia Treatment

Authority Page

[www.cancer.gov](http://www.cancer.gov)

# Hubs and Authorities Score

- **Objective:** Assign a hub score and an authority score for each page.

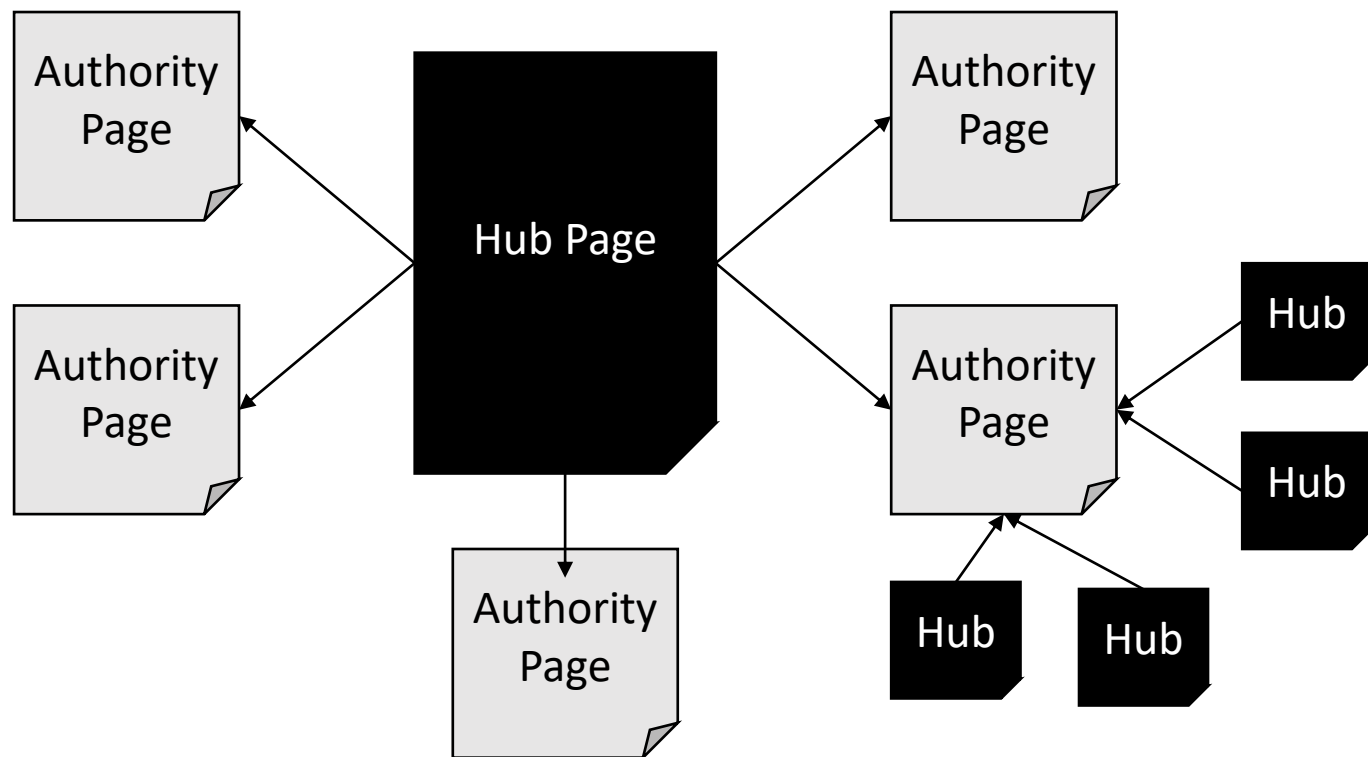


Hub Score (say,  $h(v) = 0.002$ )

Authority Score (say,  $a(v) = 0.17$ )

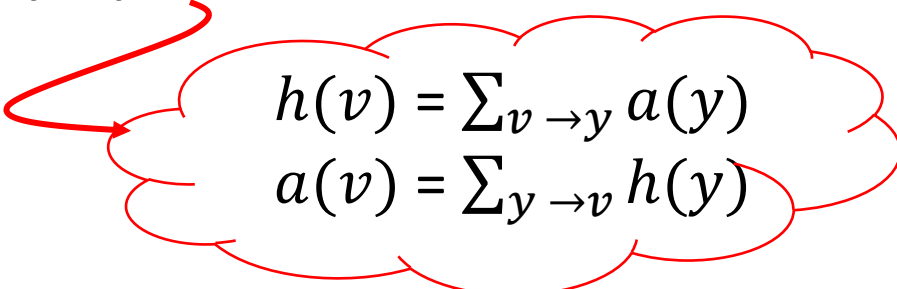
# Hub Page

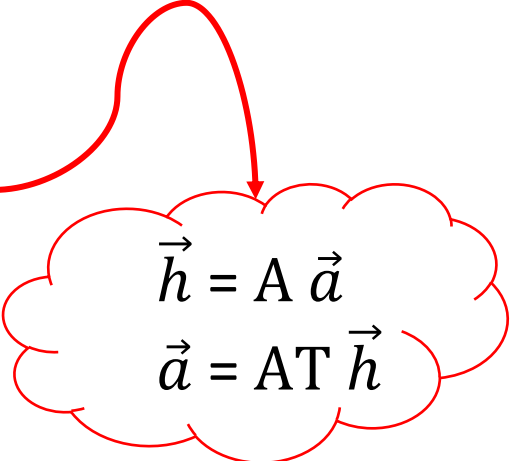
- A good hub page points to many good authorities
- A good authority page is pointed-to by many hubs



# HITS Algorithm

- For all web pages, initialize hub score ( $h(v)$ ) and authority score ( $a(v)$ ) to 1. Here,  $v$  is a web page.
- $v \rightarrow y$  denotes a link from  $v$  to  $y$ .
- Iteratively update  $h(v)$  and  $a(v)$ .


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


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

**What happens on repeated updates?**

Rewriting in matrix form  
 $A$  is the adjacency matrix

Use your mathematical hat!

# HITS Algorithm

$$\begin{array}{l} \vec{h} \leftarrow A \vec{a} \\ \vec{a} \leftarrow A^T \vec{h} \end{array}$$

**L**  $\rightarrow$

$$\begin{array}{l} \vec{h} \leftarrow AA^T \vec{h} \\ \vec{a} \leftarrow A^T A \vec{a} \end{array}$$

**L**  $\rightarrow$

$$\begin{array}{l} \vec{h} = (1/\lambda_h) AA^T \vec{h} \\ \vec{a} = (1/\lambda_a) A^T A \vec{a} \end{array}$$

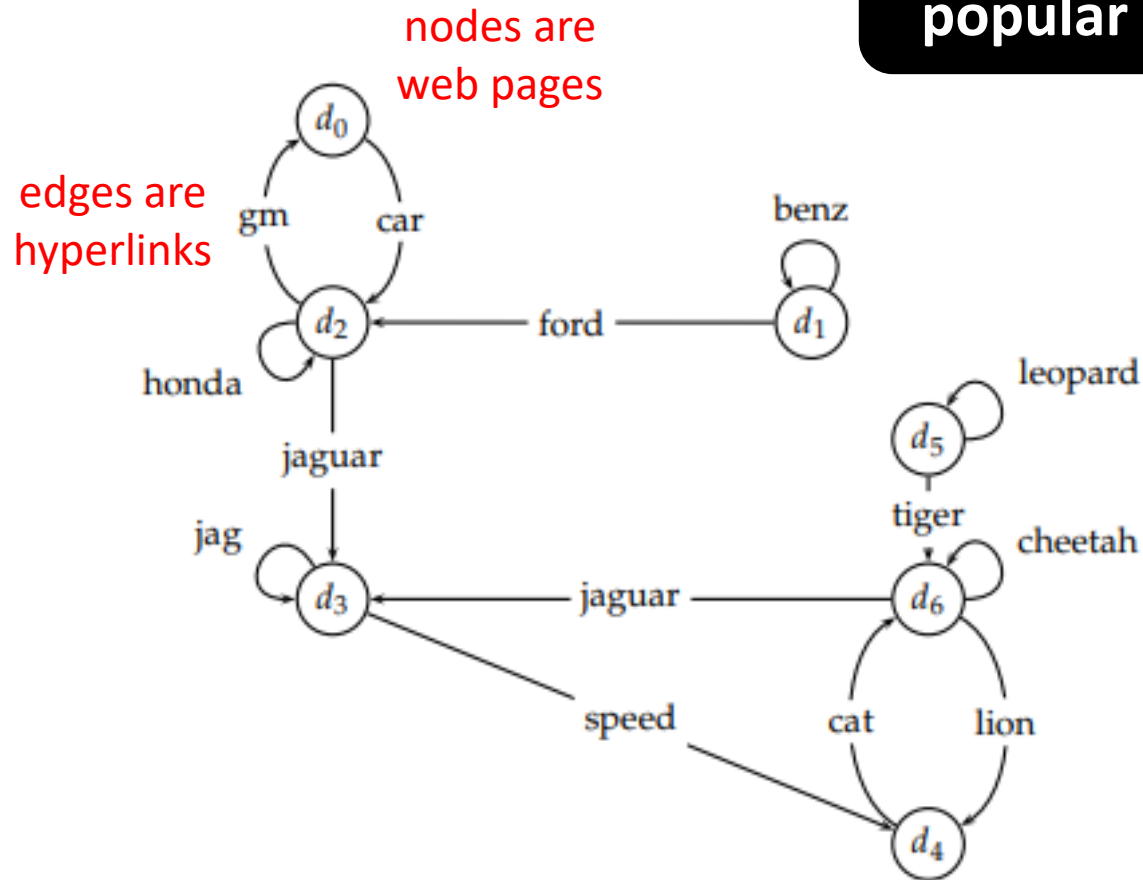
$\lambda_h$  is eigen value of  $AA^T$

**Hyperlink Induced Topic Search**



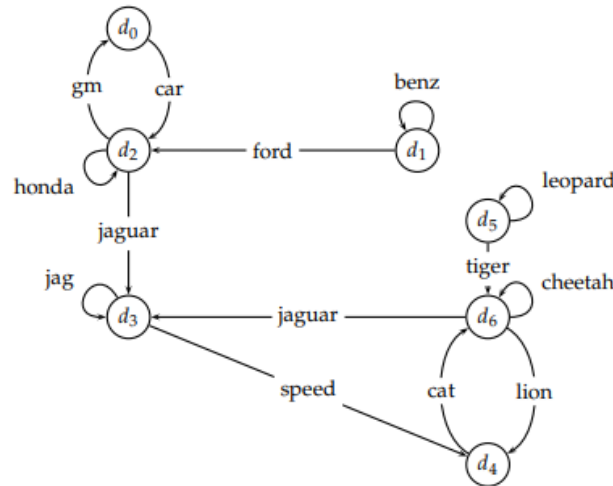
# HITS Example

Identify the most popular web page



\*Words in this graph are the anchor texts on the links

# HITS Example



Query = **jaguar** is double-weighted.

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 \end{pmatrix}$$

$$\vec{h} = (0.03 \quad 0.04 \quad 0.33 \quad 0.18 \quad 0.04 \quad 0.04 \quad 0.35)$$

$$\vec{a} = (0.10 \quad 0.01 \quad 0.12 \quad 0.47 \quad 0.16 \quad 0.01 \quad 0.13)$$

Thank You