https://vvtesh.sarahah.com/

Information Retrieval

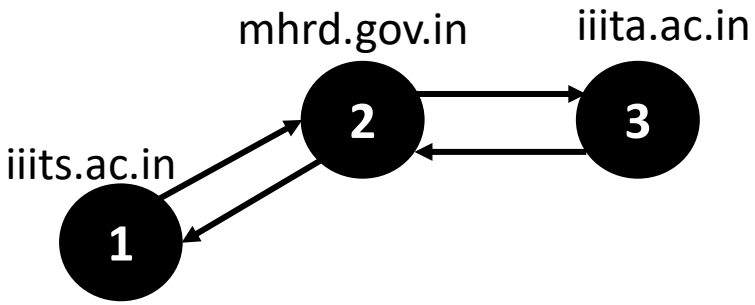
Venkatesh Vinayakarao

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

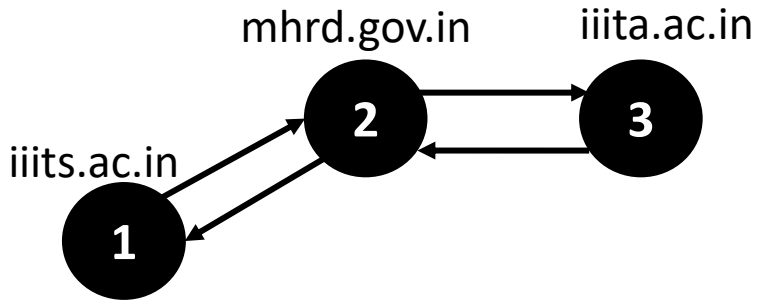
- A random surfer may start from any node with 1/3 probability



- Can you represent this graph using Adjacency Matrix?

A Random Surfer

- May start from any node with 1/3 probability

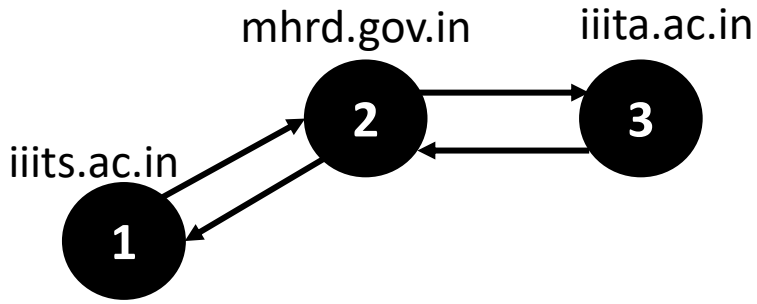


- Can you represent this graph using Adjacency Matrix?

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

A Random Surfer

- May start from any node with 1/3 probability



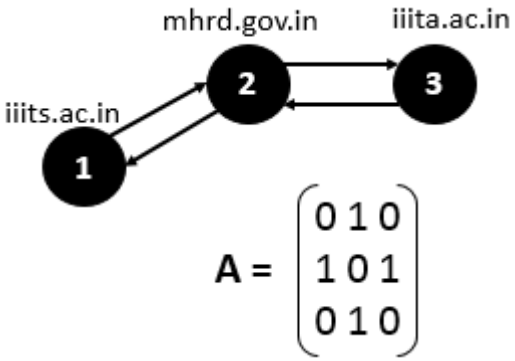
- May also teleport to any node with α probability

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

How can you compute the transition probabilities?

Transition Probabilities

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

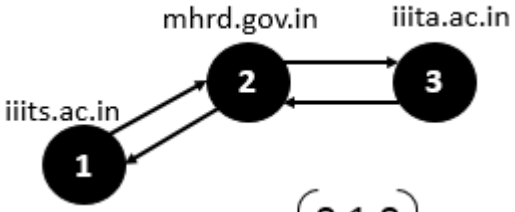


- If the random surfer is at 1,
  - and he did not teleport
    - Probability = 1- α
  - and he teleports
    - he may reach state3 with probability α/3
    - and may reach state 2 with probability α/3

- Transition Probability from 1 is (α/3, (1- α) + α/3, α/3)

Quiz

- If the teleportation probability, α = 0.5, Calculate the transition probability matrix for this network.

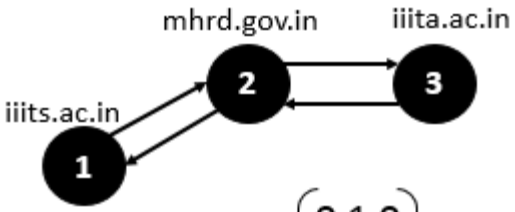


$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

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

Quiz

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



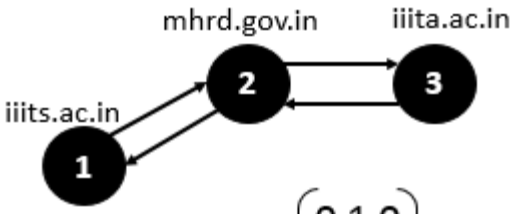
$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

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

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

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 choses to continue his walk?

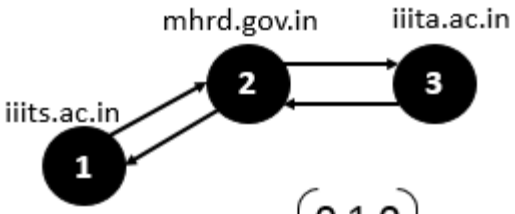


$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

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

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 choses to continue his walk?



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

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

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

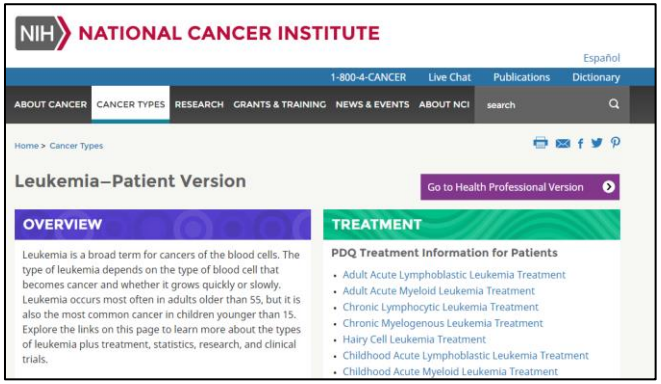
Hubs and Authorities

Topic specific page rank can be useful!

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

Hub Page

https://blog.feedspot.com/leukemia_blogs/

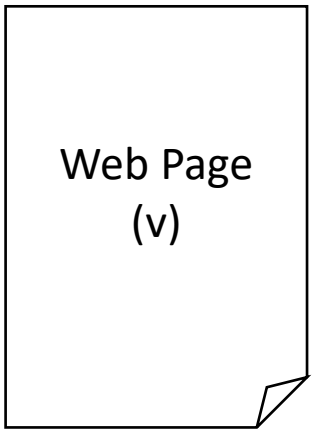


Authority Page

www.cancer.gov

Hubs and Authorities Score

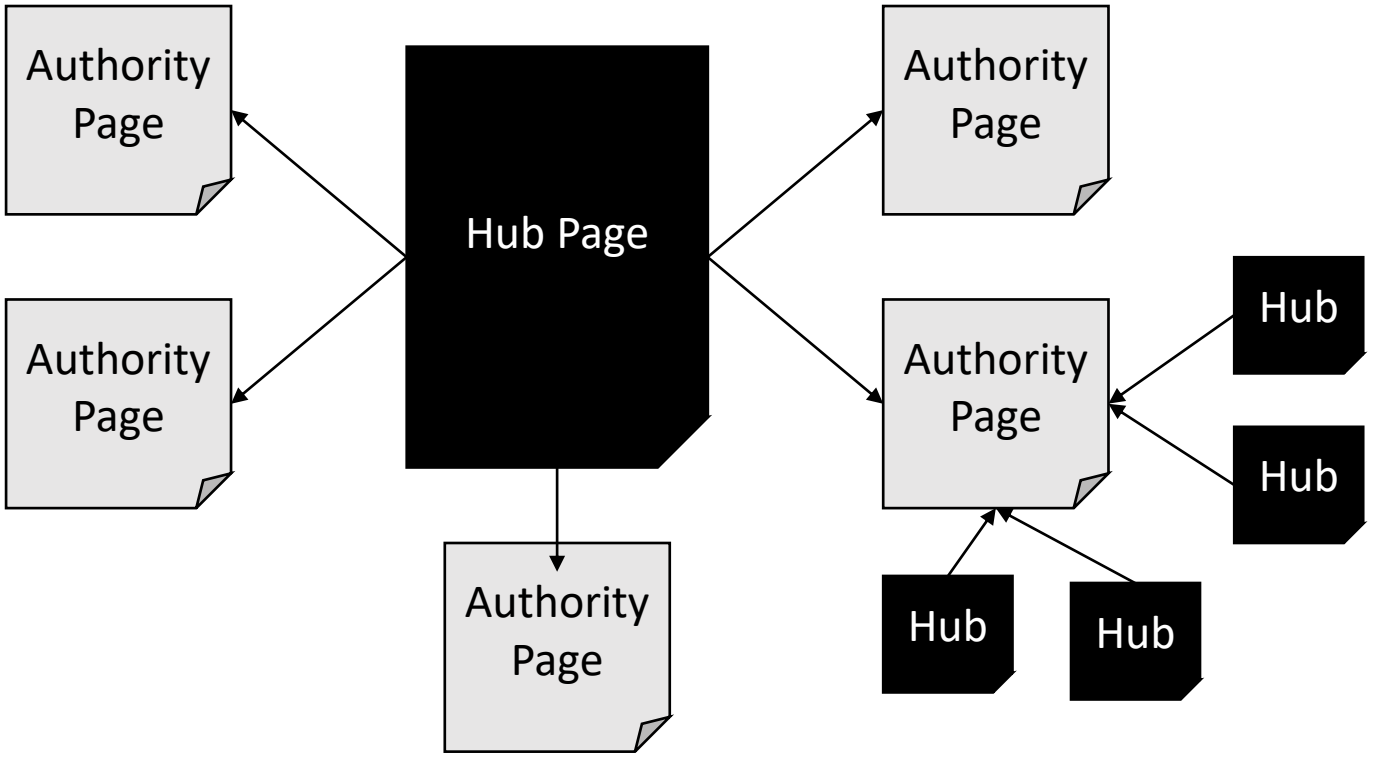
- **Objective**: Assign a hub score and an authority score for each page.



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 → y denotes a link from v to y.
- Iteratively update h(v) and a(v).

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

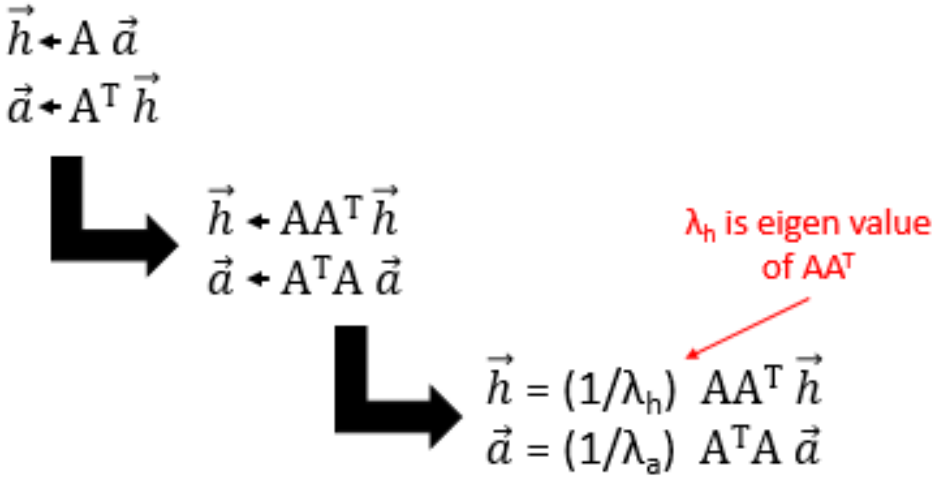
$$\vec{h} = \mathrm{A}\,\vec{a}$$
$$\vec{a} = \mathrm{AT}\,\vec{h}$$

Rewriting in matrix form
A is the adjacency matrix

What happens on repeated updates?

Use your mathematical hat!

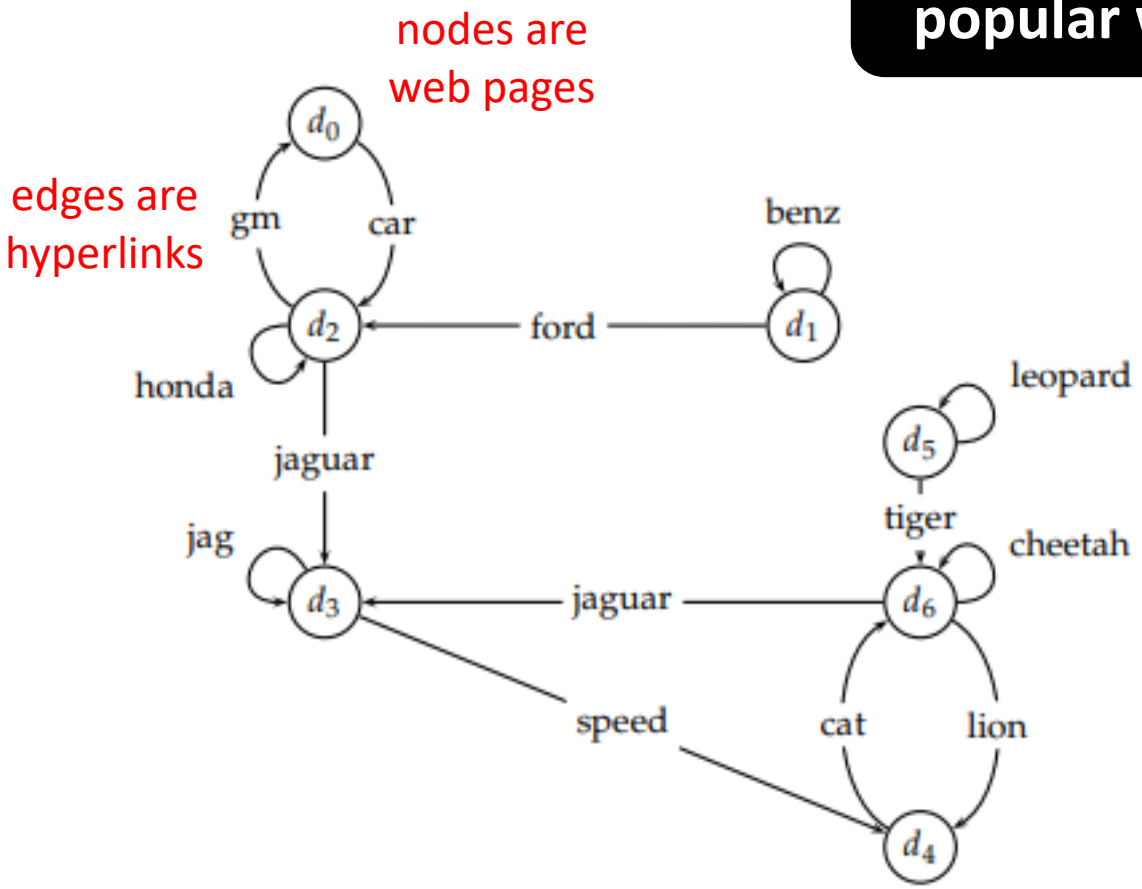
HITS Algorithm



Hyperlink Induced Topic Search

HITS Example

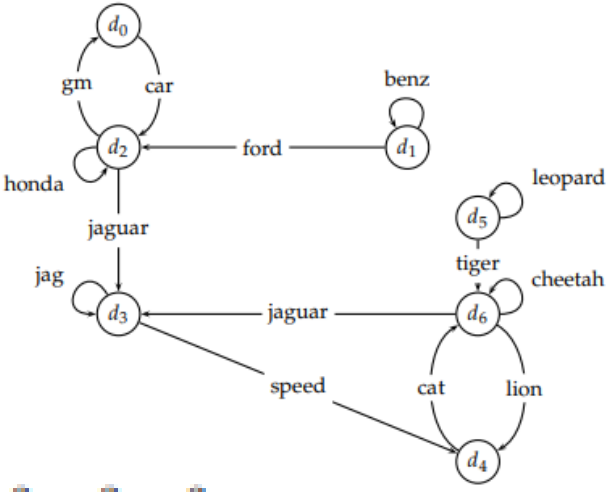
Identify the most popular web page



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

HITS Example

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



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

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

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

Thank You