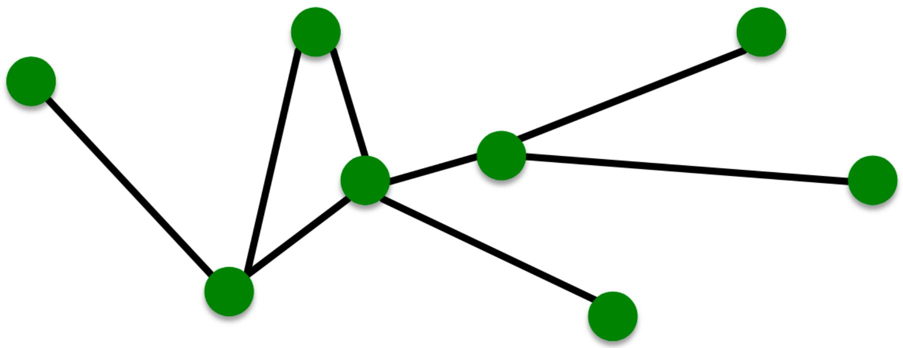
PageRank

Hongchang Gao

Spring 2024

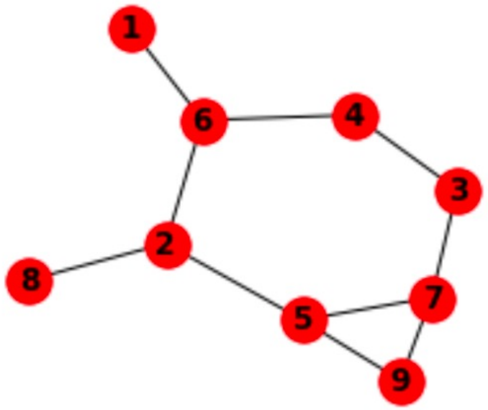
Graph Terminology

- Components of a Graph
  - Nodes/vertices
  - Edges/links
  - Graph/Network

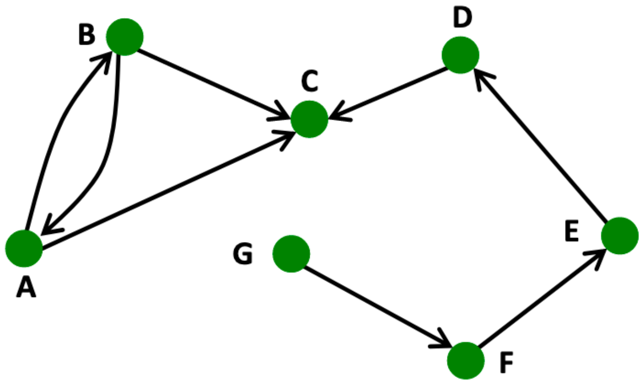


Graph Terminology

- Types of Graphs
  - Undirected Graph: links are undirected
    - Friendship on Facebook
  - Directed Graph: links are directed
    - Following on Twitter



Undirected graph

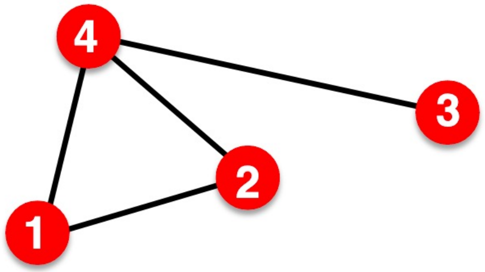


Directed graph

Graph Terminology

- Adjacency matrix

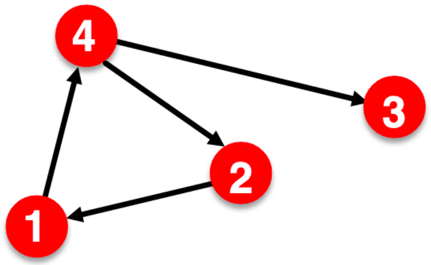
Undirected



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

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

Directed

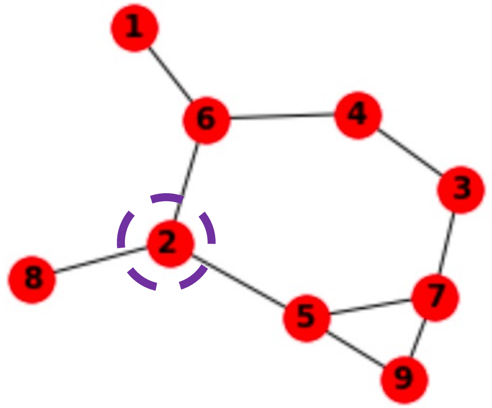


$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

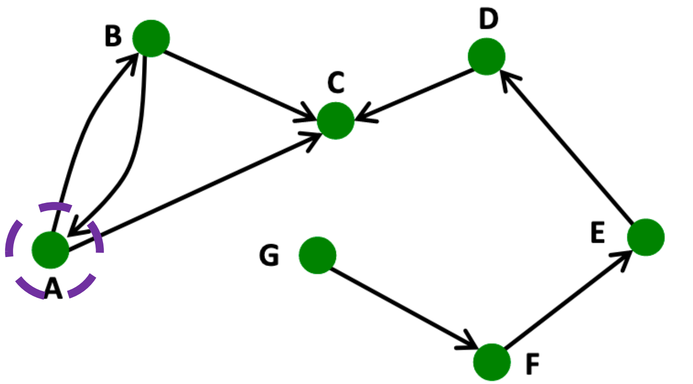
Graph Terminology

- Node degrees of undirected graph
  - The number of edges adjacent to a node



Node 2: d = 3

- Node degrees of directed graph
  - In-degree: the number of head ends adjacent to a node
  - Out-degree: the number of tail ends adjacent to a node

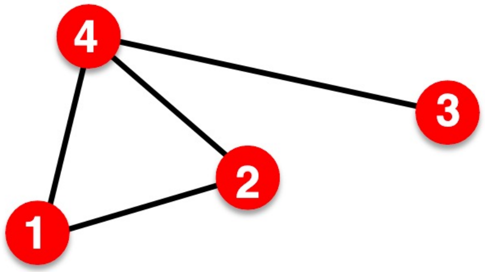


Node A: $d_{in}=1$, $d_{out}=2$

Graph Terminology

- Node degrees

Undirected

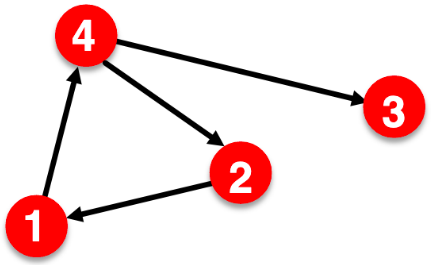


$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

Node 2: 1+0+0+1=2

$$A_{ij} = A_{ji}$$
$$A_{ii} = 0$$

Directed



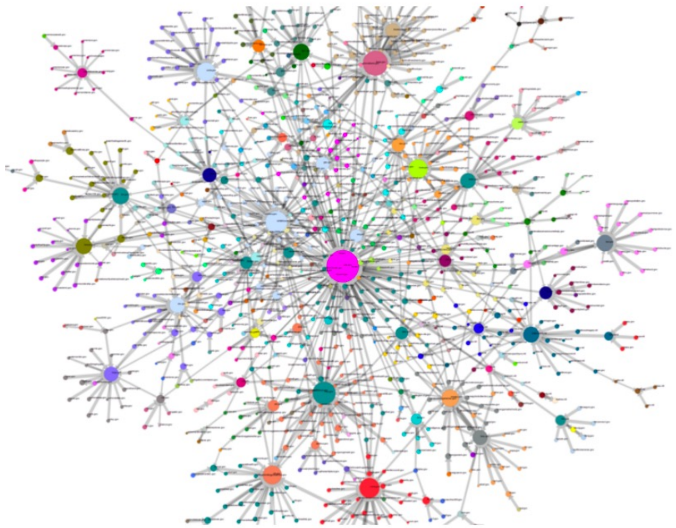
$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

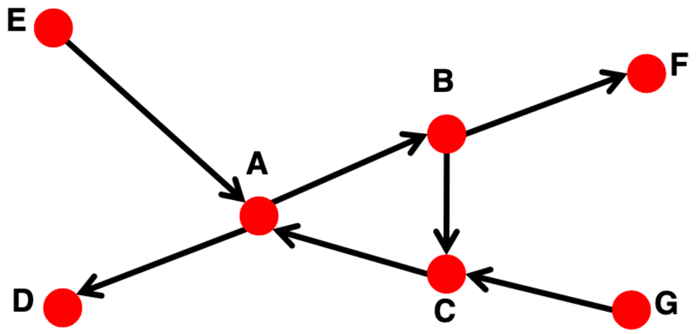
Node 4:
Indegree (column sum): 1+0+0+0=1
Outdegree (row sum): 0+1+1+0=2

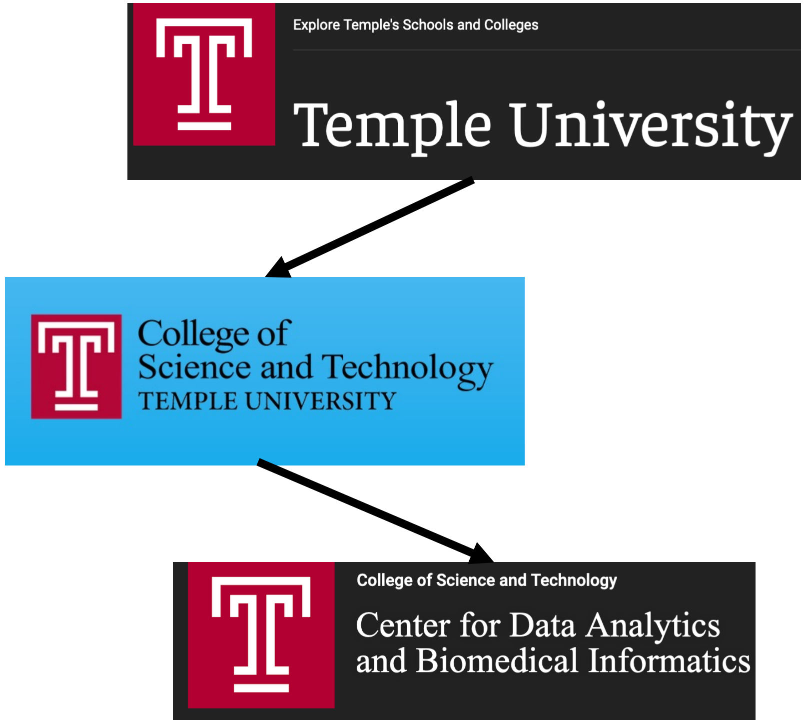
$$A_{ij} \neq A_{ji}$$
$$A_{ii} = 0$$

PageRank

- The web is a graph
  - Nodes: web pages
  - Edges: hyperlinks

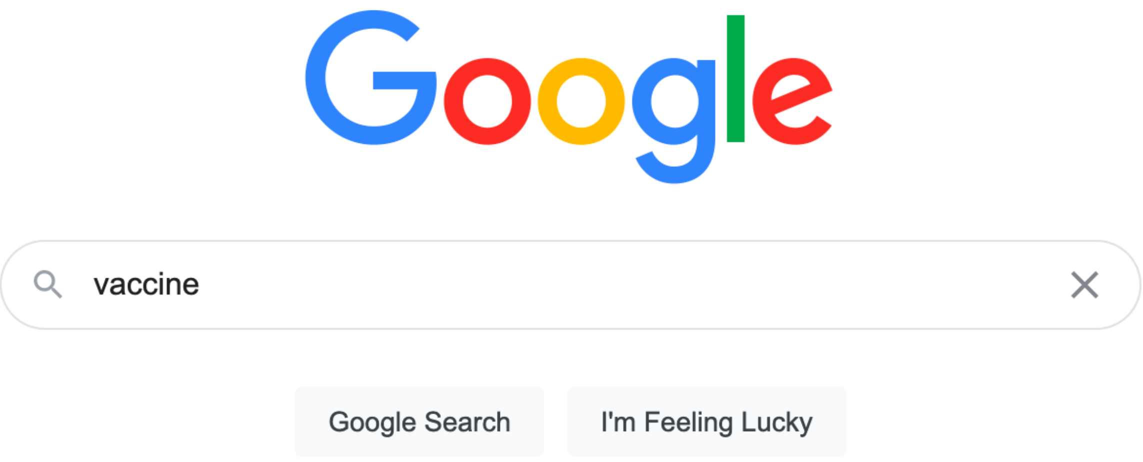


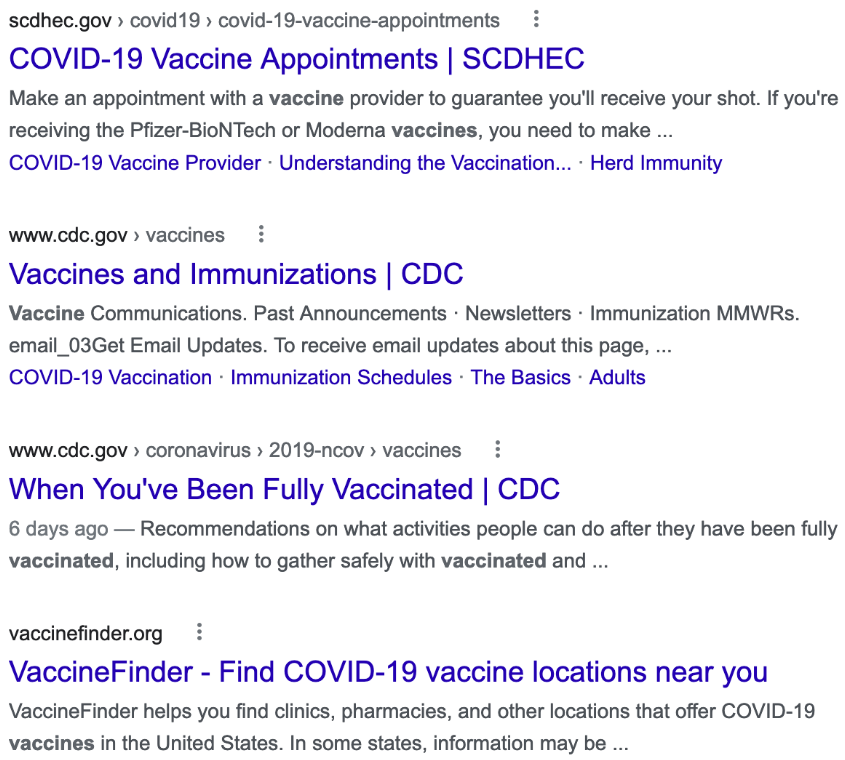




PageRank

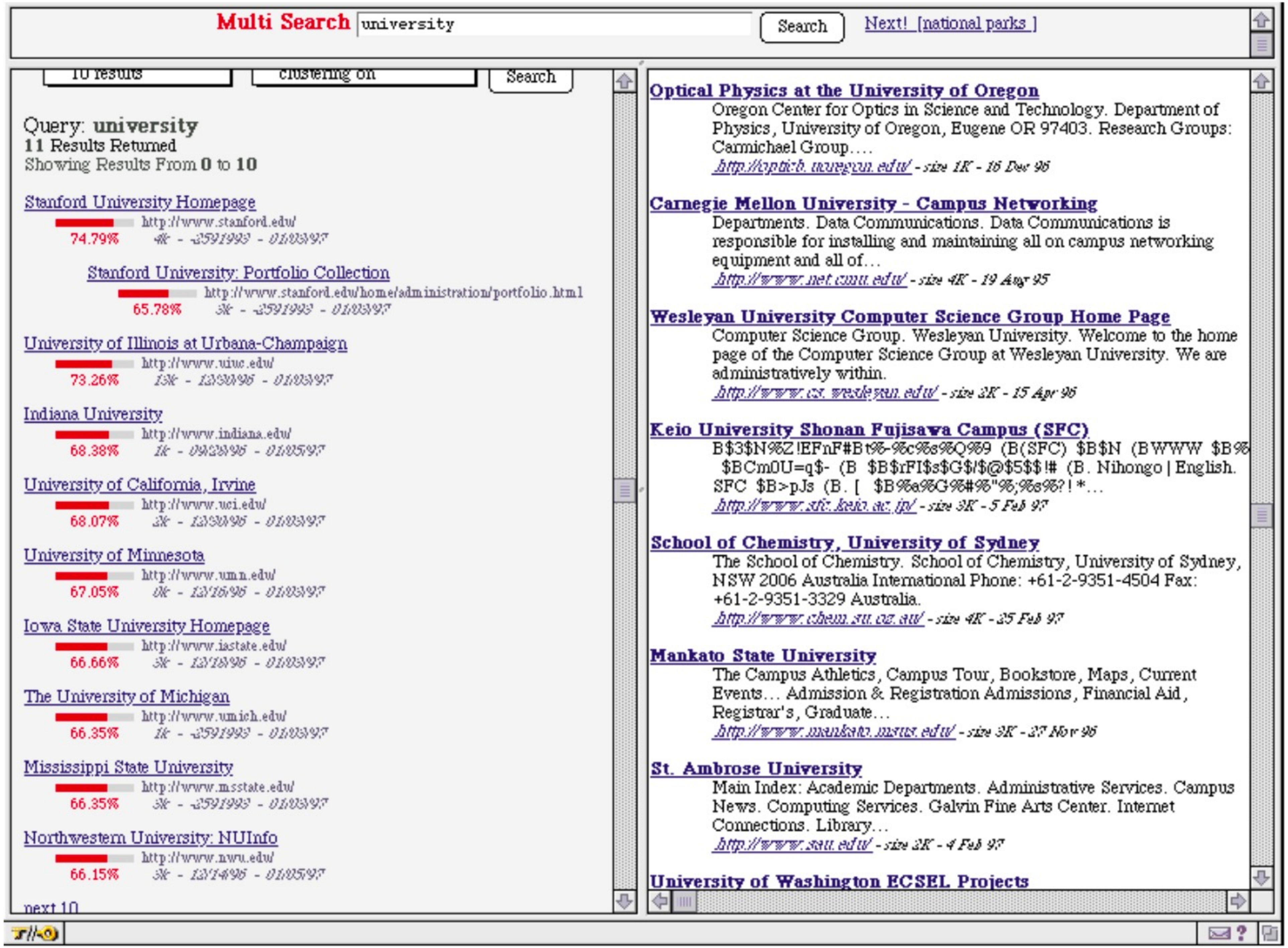
- All web pages are not equally “important”
  - Some webpages should be assigned more priority than others, for being more important
  - Which node (webpage) is important?





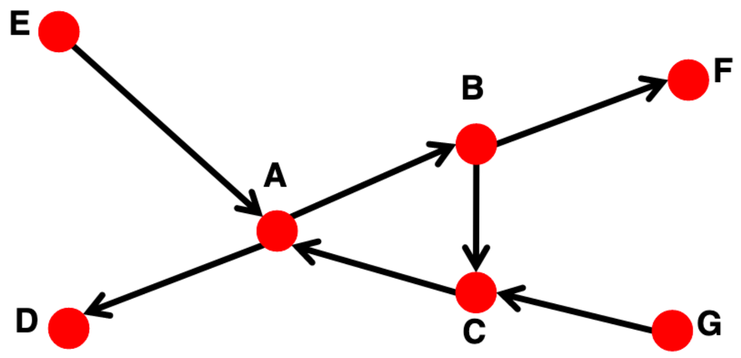
Pa

- Ra
  -
- Pa
  -
- Se
  - ery
  -



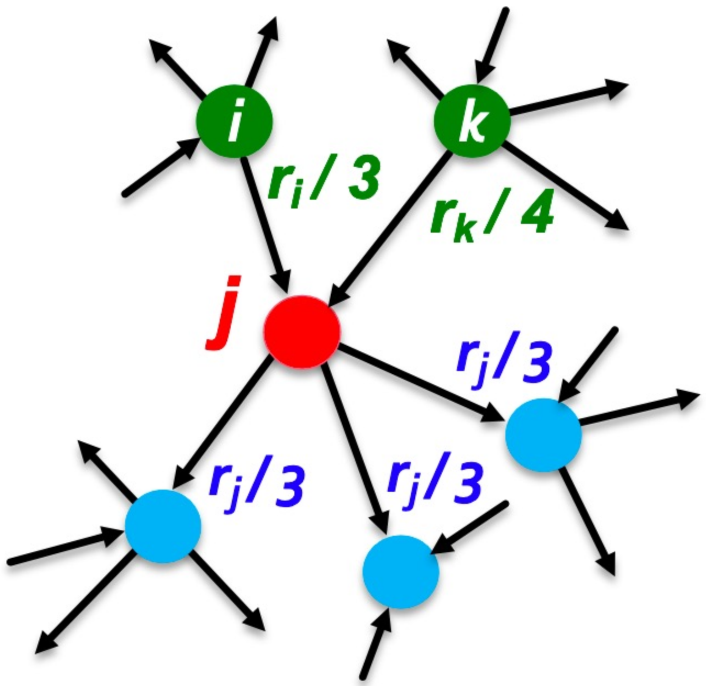
PageRank

- Compute the importance of nodes in a graph
  - Idea: Links as votes
  - Page is more important if it has more links
- Use in-links as votes
  - How to use votes to compute the importance score???
  - Q: E and C may be different. C may be more important.
  - How to differentiate their importance when they vote for A?



PageRank

- A vote from an important page is worth more:
  - Each link's vote is proportional to the **importance** of its source page
  - If page $\boldsymbol{i}$ with importance $\boldsymbol{r_i}$ has $\boldsymbol{d_i}$ out-links, each link gets $\boldsymbol{r_i / d_i}$ votes
  - Page $\boldsymbol{j}$'s own importance $\boldsymbol{r_j}$ is the sum of the votes on its in-links



$$r_j = r_i/3 + r_k/4$$

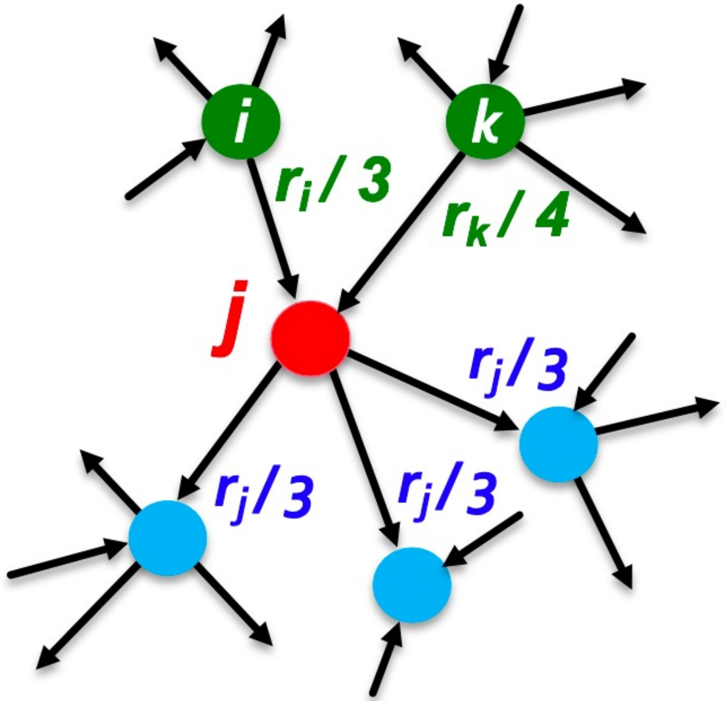
PageRank

- Formally, the importance score of each node is

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … **out-degree of node i**

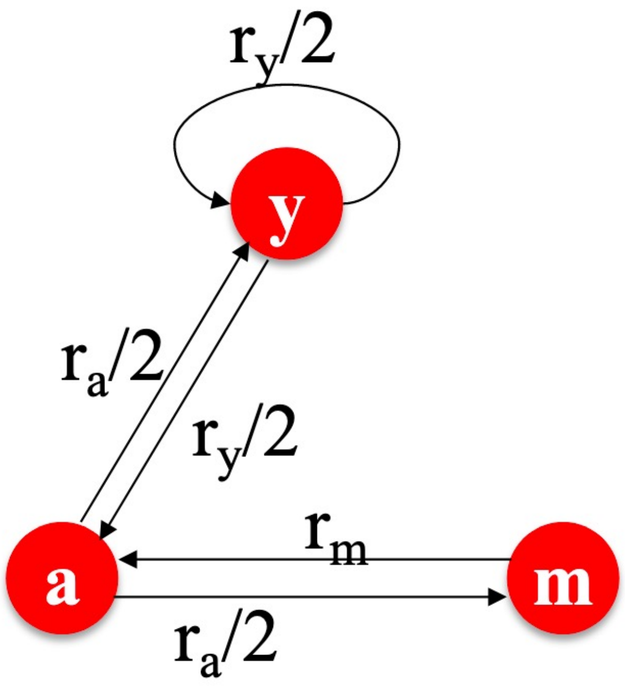
A page is important if it is pointed by other important pages



$r_j = r_i/3 + r_k/4$

PageRank

- Example



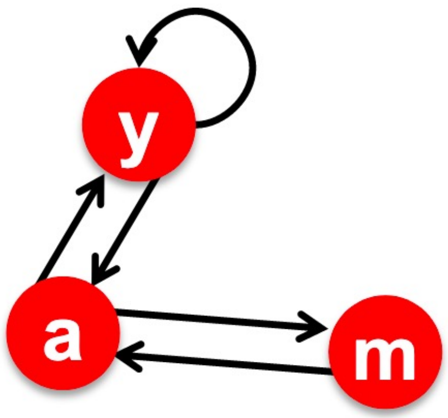
$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

PageRank

- Matrix Form



Graph

Out-degree 2, 2, 1

	y	a	m
y	1	1	0
a	1	0	1
m	0	1	0

Adjacency matrix

	r_y	r_a	r_m
r_y	½	½	0
r_a	½	0	1
r_m	0	½	0

Transition matrix M

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i **… out-degree of node** i

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

$$\underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{r} = \underbrace{\begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{r}$$

PageRank

- Property of the transition matrix M
  - Column sum is 1

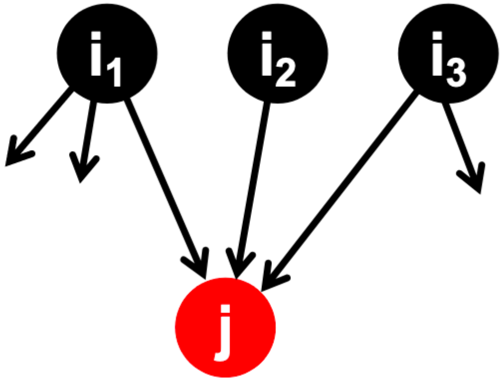
- Property of the rank vector r
  - r_i is the importance score of page i

$$\sum_i \boldsymbol{r_i} \;= \mathbf{1}$$

$$\underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}} = \underset{\boldsymbol{M}}{\boxed{\begin{matrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{matrix}}} \; \underset{\boldsymbol{r}}{\boxed{\begin{matrix} r_y \\ r_a \\ r_m \end{matrix}}}$$

PageRank

- Interpretation
  - At time t, the user is on page i
  - At time t+1, the user follows an out-link from i uniformly at random
  - Ends up on some page j linked from i



$$r_j = \sum_{i \to j} \frac{r_i}{\mathrm{d}_{\mathrm{out}}(\mathrm{i})}$$

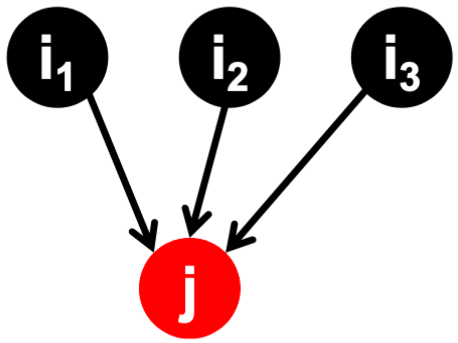
$$\underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{\boldsymbol{r}} = \underbrace{\begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{bmatrix}}_{\boldsymbol{M}} \underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{\boldsymbol{r}}$$

PageRank

- Interpretation (continue):
  - Define $\boldsymbol{p(t)}$ is a probability distribution over pages
  - At time t+1, we have

$$\boldsymbol{p(t+1) = M \cdot p(t)}$$

Random walk

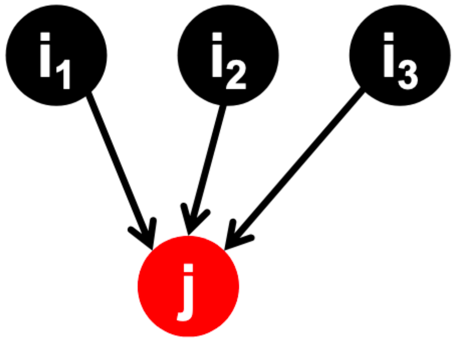


PageRank

- Solution of the importance score r:

$\boldsymbol{p}(t)$ is **stationary distribution** of a random walk

$$\boldsymbol{p}(\boldsymbol{t}+1) = \boldsymbol{M} \cdot \boldsymbol{p}(\boldsymbol{t})$$



Random walk

Stationary distribution

$$\boldsymbol{r} = \boldsymbol{M} \cdot \boldsymbol{r}$$

r is the stationary distribution of the random walk

r is the eigenvector of the transition matrix M (with eigenvalue 1)

PageRank

- Solution of the importance score r:
  - Compute the eigenvector of the transition matrix M with eigenvalue 1
  - Use power iteration to compute the eigenvector efficiently

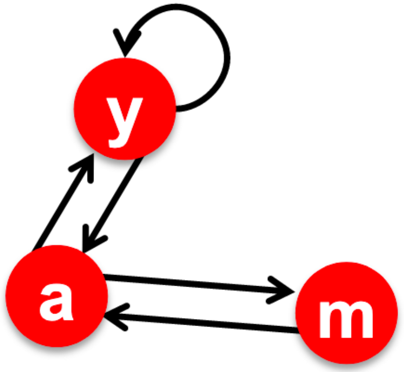
- Assign each node an initial page rank
- Repeat until convergence ($\sum_i |r_i^{t+1} - r_i^t| < \epsilon$)
  - Calculate the page rank of each node

$$r_j^{(t+1)} = \sum_{i \to j} \frac{r_i^{(t)}}{d_i}$$

- Initialize: $\boldsymbol{r}^0 = [1/N, \dots., 1/N]^T$
- Iterate: $\boldsymbol{r}^{(t+1)} = \boldsymbol{M} \cdot \boldsymbol{r}^t$
- Stop when $|\boldsymbol{r}^{(t+1)} - \boldsymbol{r}^t|_1 < \varepsilon$

PageRank

- Example



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$\underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{\boldsymbol{r}} = \underbrace{\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}}_{\boldsymbol{M}} \underbrace{\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}}_{\boldsymbol{r}}$$

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \begin{pmatrix} 1/3 \\ 3/6 \\ 1/6 \end{pmatrix} \begin{pmatrix} 5/12 \\ 1/3 \\ 3/12 \end{pmatrix} \begin{pmatrix} 9/24 \\ 11/24 \\ 1/6 \end{pmatrix} \ldots \begin{pmatrix} 6/15 \\ 6/15 \\ 3/15 \end{pmatrix}$$

Iteration 0, 1, 2, …