Department of Computer Science

CS47300: Web Information Search and Management

Graph Structure for IR: PageRank

Prof. Chris Clifton

25 September 2020

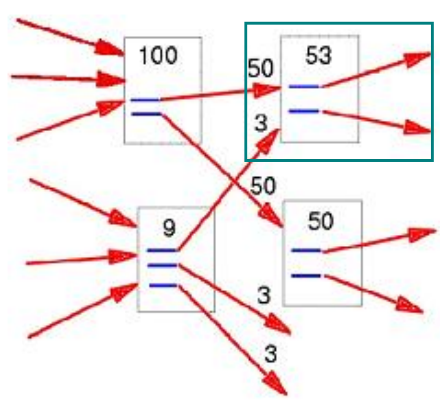
Material adapted from slides created by Dr. Rong Jin (formerly Michigan State, now at Alibaba)

Indiana Center for Database Systems ™

Department of Computer Science

PageRank

- Introduced by Page et al. (1998)
  - The weight is assigned by the rank of parents
- Difference from HITS
  - HITS separates Hubness & Authority weights
  - Page rank is proportional to its parents' rank, but inversely proportional to its parents' outdegree

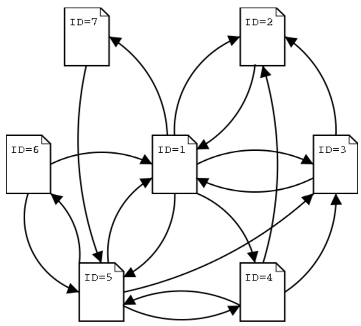


PURDUE UNIVERSITY
Department of Computer Science

Matrix Notation

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{B}_{i,j} = \begin{cases} \dfrac{1}{\sum_j \mathbf{M}_{i,j}} & \sum_j \mathbf{M}_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$



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

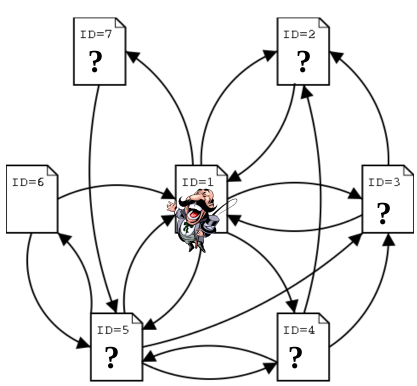
$$\mathbf{B} = \begin{pmatrix} 0&1/5&1/5&1/5&1/5&0&1/5 \\ 1&0&0&0&0&0&0 \\ 1/2&1/2&0&0&0&0&0 \\ 0&1/3&1/3&0&1/3&0&0 \\ 1/4&0&1/4&1/4&0&1/4&0 \\ 1/2&0&0&0&1/2&0&0 \\ 0&0&0&0&1&0&0 \end{pmatrix}$$

29

PURDUE UNIVERSITY
Department of Computer Science

Random Walk Model

- Consider a random walk through the Web graph



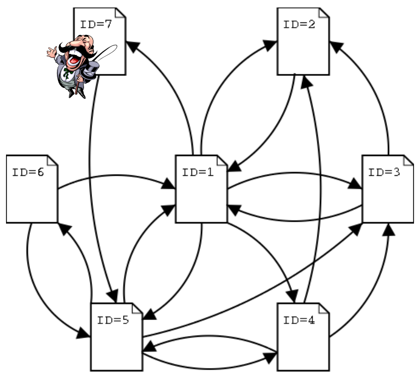
$$\mathbf{B} = \begin{pmatrix} \boxed{0\;\;1/5\;\;1/5\;\;1/5\;\;1/5\;\;0\;\;1/5} \\ 1\;\;0\;\;0\;\;0\;\;0\;\;0\;\;0 \\ 1/2\;\;1/2\;\;0\;\;0\;\;0\;\;0\;\;0 \\ 0\;\;1/3\;\;1/3\;\;0\;\;1/3\;\;0\;\;0 \\ 1/4\;\;0\;\;1/4\;\;1/4\;\;0\;\;1/4\;\;0 \\ 1/2\;\;0\;\;0\;\;0\;\;1/2\;\;0\;\;0 \\ 0\;\;0\;\;0\;\;0\;\;1\;\;0\;\;0 \end{pmatrix}$$

30

Random Walk Model

PURDUE UNIVERSITY
Department of Computer Science

- Consider a random walk through the Web graph



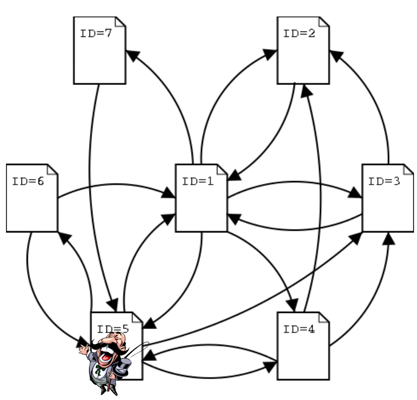
$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

31

Random Walk Model

PURDUE UNIVERSITY
Department of Computer Science

- Consider a random walk through the Web graph



$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

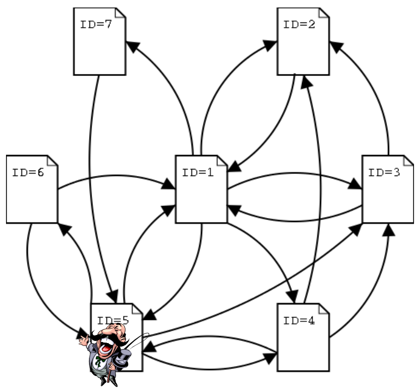
As $T \rightarrow \infty$, what is portion of time that the web surfer will spend at each page?

32

PURDUE UNIVERSITY
Department of Computer Science

Random Walk Model

- Consider a random walk through the Web graph



$$\mathbf{B} = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$p(k)$: percentage of time that the surfer will stay at the i-th site

$$p(k) = \sum_i p(i)\mathbf{B}_{i,k}$$

$$\downarrow$$

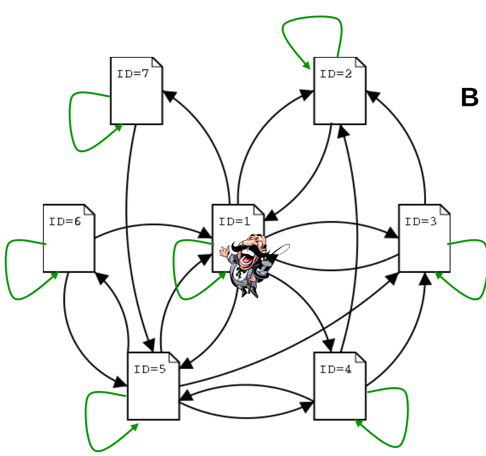
$$\mathbf{p} = \mathbf{B}^T \mathbf{p}$$

33

PURDUE UNIVERSITY
Department of Computer Science

Adding Self Loop

- Allow surfer to decide to stay on the same place



$$\mathbf{B} = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\mathbf{B}' = \alpha \mathbf{B} + (1-\alpha)\mathbf{I}$$

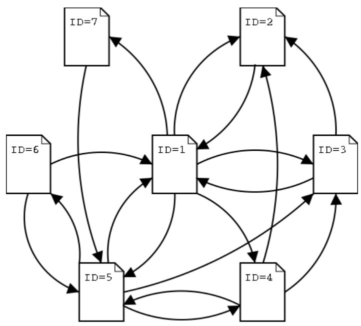
34

PURDUE UNIVERSITY
Department of Computer Science

Matrix Notation

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{the } i\text{th site points to the } j\text{th site} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{B}_{i,j} = \begin{cases} \dfrac{1}{\sum_j \mathbf{M}_{i,j}} & \sum_j \mathbf{M}_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$



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

$$\mathbf{B} = \begin{pmatrix} 0&1/5&1/5&1/5&1/5&0&1/5 \\ 1&0&0&0&0&0&0 \\ 1/2&1/2&0&0&0&0&0 \\ 0&1/3&1/3&0&1/3&0&0 \\ 1/4&0&1/4&1/4&0&1/4&0 \\ 1/2&0&0&0&1/2&0&0 \\ 0&0&0&0&1&0&0 \end{pmatrix}$$

35

PURDUE UNIVERSITY
Department of Computer Science

Matrix Notation

$\mathbf{r} : \mathbf{r}_i$ represents the rank score for the i-th web page

$$r(v) = \alpha \sum_{w \in \mathrm{pa}[v]} \frac{r(w)}{|\mathrm{ch}[w]|'} \longrightarrow$$

r = α **B**^{T} **r**
α : eigenvalue
r : eigenvector of **B**

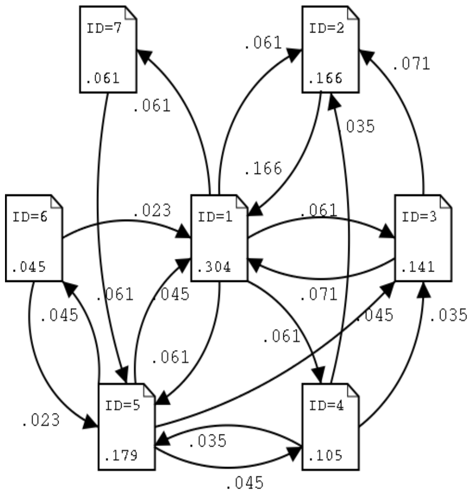
Finding Pagerank

→ find principle eigenvector of B

36

PURDUE UNIVERSITY
Department of Computer Science

Matrix Notation



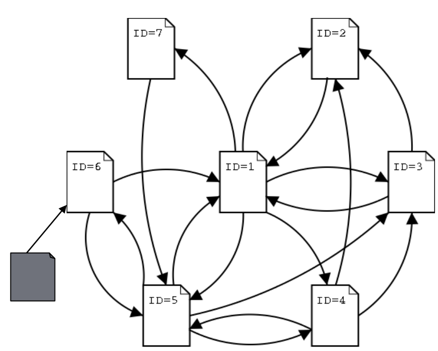
PR	ID	OutLink	InLink
0.304	1	2,3,4,5,7	2,3,5,6
0.179	5	1,3,4,6	1,4,6,7
0.166	2	1	1,3,4
0.141	3	1,2	1,4,5
0.105	4	2,3,5	1,5
0.061	7	5	1
0.045	6	1,5	5

PURDUE UNIVERSITY
Department of Computer Science

Problem

- "Rank Sink" Problem
  - Many Web pages have no inlinks
  - Results in dangling edges in the graph

$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

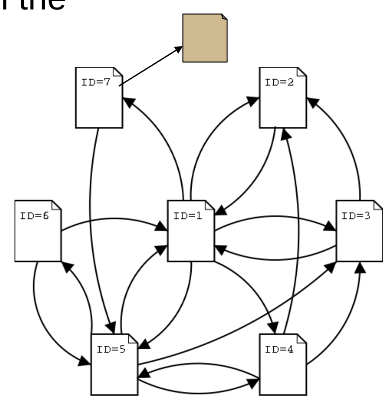


r(new page) = 0

Problem

- "Rank Sink" Problem
  - Many Web pages have no outlinks
  - Results in dangling edges in the graph

$$\mathbf{B} = \left(\begin{array}{cccccccc} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right)$$



r(new page) = 1

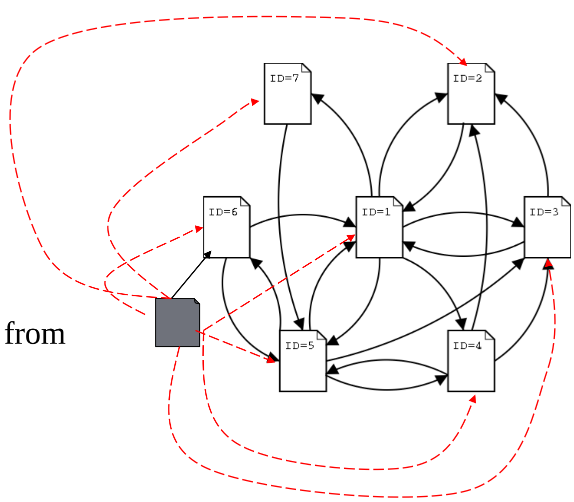
Distribution of the Mixture Model

$$\mathbf{H}_{i,j} = 1/n$$

$$\mathbf{B}' = \varepsilon\mathbf{H} + (1-\varepsilon)\mathbf{B}$$

$$\downarrow$$

$$\mathbf{r} = \mathbf{B}'^{T}\,\mathbf{r}$$



Prevents the page ranks from being 0 or 1

Stability

- Are link analysis algorithms based on eigenvectors stable?
  - Will small changes in graph result in major changes in outcomes?
- What if the connectivity of a portion of the graph is changed arbitrarily?
  - How will this affect the results of algorithms?

41

Stability of HITS

Ng et al (2001)

- A bound on the number of hyperlinks k that can be added or deleted from one page without affecting the authority or hubness weights
- It is possible to perturb a symmetric matrix by a quantity that grows as δ that produces a constant perturbation of the dominant eigenvector

$$k \leq \left(\sqrt{d + \frac{\alpha\delta}{4 + \sqrt{2}\alpha}} - \sqrt{d} \right)^2$$

$$\|\boldsymbol{a} - \tilde{\boldsymbol{a}}\|_2 \leq \alpha$$

δ: eigengap $\lambda_1 - \lambda_2$

d: maximum outdegree of G

42

Stability of PageRank

$$\|\tilde{r} - r\| \leq \frac{2\sum_{j \in \mathrm{V}} r(j)}{\epsilon}$$ Ng et al (2001)

V: the set of vertices touched by the perturbation

- The parameter ε of the mixture model has a stabilization role
- If the set of pages affected by the perturbation have a small rank, the overall change will also be small

43