PURDUE
UNIVERSITY

CS54701: Information Retrieval

Link Analysis

16 February 2016

Prof. Jennifer Neville

Indiana Center for Database Systems ™

Ad-Hoc Retrieval: Beyond the Words

- Web is a graph
  - Each web site correspond to a node
  - A link from one site to another site forms a directed edge

2

Citation Analysis

3

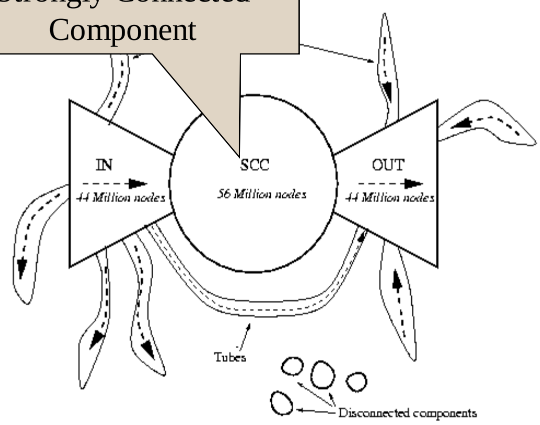
Ad-Hoc Retrieval: Beyond the Words

- Web is a graph
  - Each web site correspond to a node
  - A link from one site to another site forms a directed edge
- What does it look like?
  - Web is small world
  - The diameter of the web is 19
    - e.g. the average number of clicks from one web site to another is 19

4

Bowtie Structure

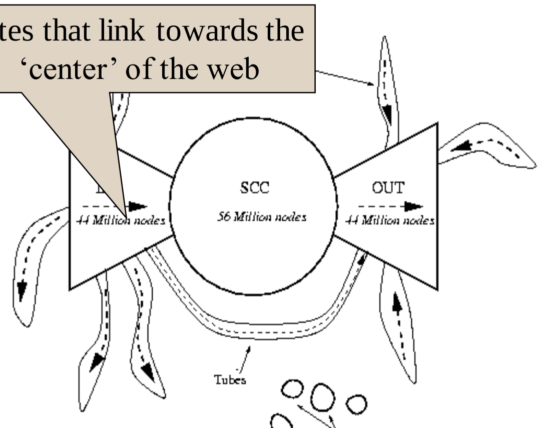
Strongly Connected Component



Broder et al., 2001

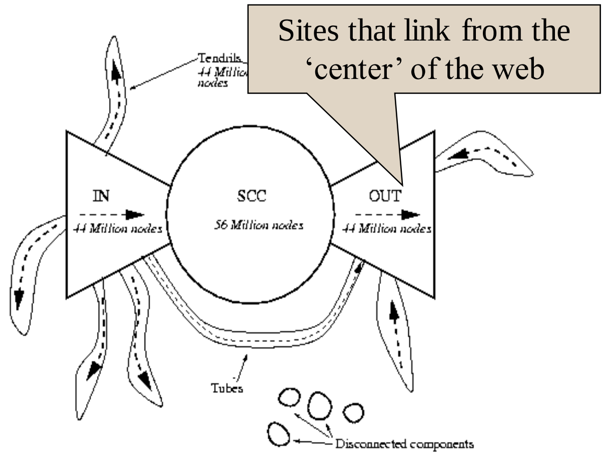
Bowtie Structure

Sites that link towards the 'center' of the web



Broder et al., 2001

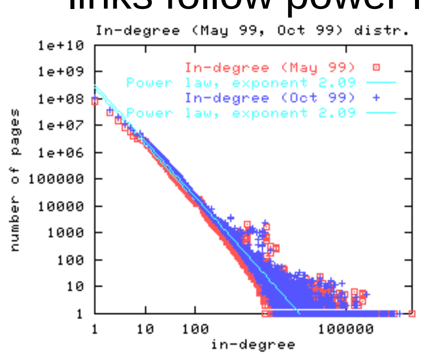
Bowtie Structure

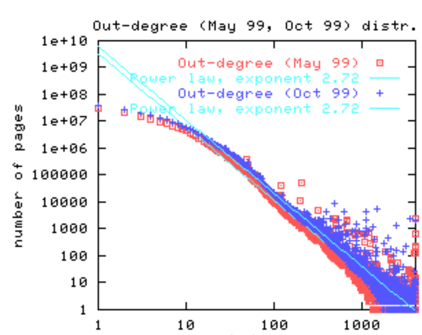


Broder et al., 2001

Inlinks and Outlinks

- Both degrees of incoming and outgoing links follow power law





Broder et al., 2001

Early Approaches

Basic Assumptions

- Hyperlinks contain information about the human judgment of a site
- The more incoming links to a site, the more it is judged important

Bray 1996

- The visibility of a site is measured by the number of other sites pointing to it
- The luminosity of a site is measured by the number of other sites to which it points

➢ Limitation: failure to capture the relative importance of different parents (children) sites

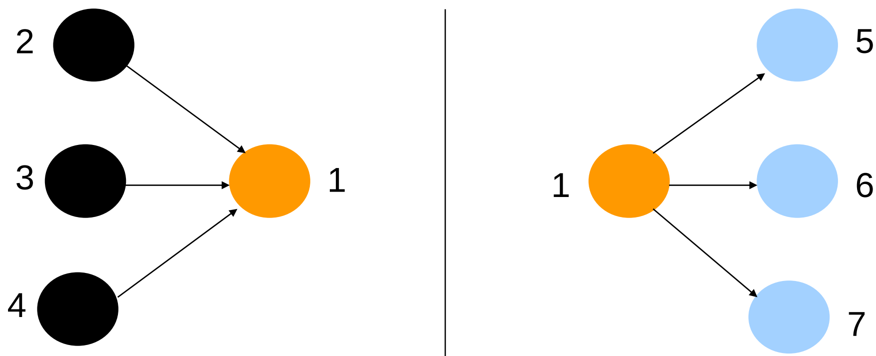
10

HITS - Kleinberg's Algorithm

- HITS – Hypertext Induced Topic Selection
- For each vertex v ϵ V in a subgraph of interest:
  - a(v) - the authority of v
  - h(v) - the hubness of v
- A site is very authoritative if it receives many citations.
  - Citation from important sites weight more than citations from less-important sites
- Hubness shows the importance of a site.
  - A good hub is a site that links to many authoritative sites

11

Authority and Hubness



$a(1) = h(2) + h(3) + h(4)$

$h(1) = a(5) + a(6) + a(7)$

12

Authority and Hubness: Version 1

Recursive dependency

$$a(v) = \sum_{w \in pa[v]} h(w)$$

$$h(v) = \sum_{w \in ch[v]} a(w)$$

HubsAuthorities(G)

$\mathbf{1} \leftarrow [1,\ldots,1] \in R^{|V|}$

$a_0 \leftarrow h_0 \leftarrow \mathbf{1}$

$t \leftarrow 1$

repeat

for each v in V

do $a_t(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$

$h_t(v) \leftarrow \sum_{w \in pa[v]} a_{t-1}(w)$

$t \leftarrow t + 1$

until $\|a_t - a_{t-1}\| + \|h_t - h_{t-1}\| < \varepsilon$

return (a_t, h_t)

Problems ?

13

Authority and Hubness: Version 2

Recursive dependency

$$a(v) = \sum_{w \in pa[v]} h(w)$$

$$h(v) = \sum_{w \in ch[v]} a(w)$$

+ Normalization

$$a(v) = \frac{a(v)}{\sum_w a(w)}$$

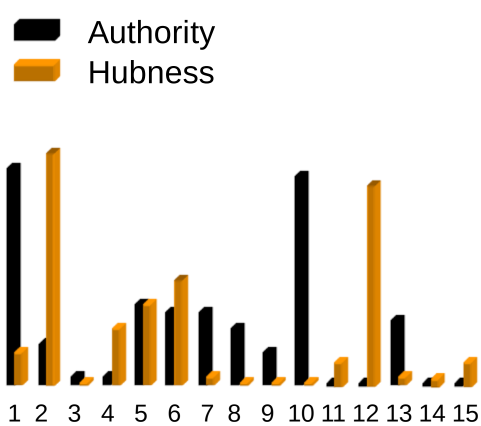
$$h(v) = \frac{h(v)}{\sum_w h(w)}$$

HubsAuthorities(G)

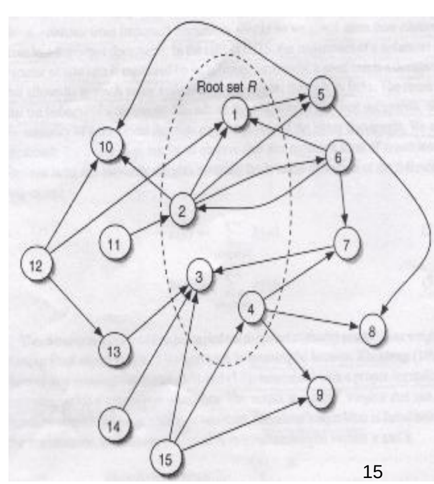
```
1 ← [1,…,1] Є R^|V|
a_0 ← h_0 ← 1
t ← 1
repeat
    for each v in V
        do  a_t (v) ← Σ_{w Є pa[v]} h_{t-1} (w)
            h_t (v) ← Σ_{w Є pa[v]} a_{t-1} (w)
        a_t ← a_t / || a ||
        h_t ← h_t / || h ||
        t ← t + 1
until || a_t – a_{t-1} || + || h_t – h_{t-1} || < ε
return (a_t , h_t)
```

14

HITS Example Results



Authority and hubness weights



15

Authority and Hubness

- Authority score
  - Not only depends on the number of incoming links
  - But also the ‘quality’ (e.g., hubness) of the incoming links
- Hubness score
  - Not only depends on the number of outgoing links
  - But also the ‘quality’ (e.g., hubness) of the outgoing links

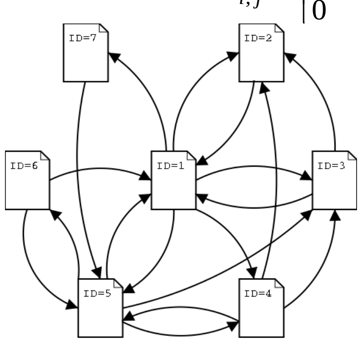
16

Authority and Hub

- Column vector ***a***: a_i is the authority score for the i-th site
- Column vector ***h***: h_i is the hub score for the i-th site

- Matrix **M**:

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



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

17

Authority and Hub

- Vector ***a***: a_i is the authority score for the i-th site
- Vector ***h***: h_i is the hub score for the i-th site
- Matrix **M**:

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

- Recursive dependency:

$$a(v) \leftarrow \Sigma_{w \in pa[v]} h(w)$$

$$h(v) \leftarrow \Sigma_{w \in ch[v]} a(w)$$

18

Authority and Hub

•Column vector ***a***: a_i is the authority score for the i-th site
•Column vector ***h***: h_i is the hub score for the i-th site

•Matrix **M**:

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

- Recursive dependency:

$$a(v) \leftarrow \Sigma_{w \in pa[v]} h(w)$$

$$h(v) \leftarrow \Sigma_{w \in ch[v]} a(w)$$

$$\Rightarrow \quad \mathbf{a} = \mathbf{M}^T\mathbf{h}, \quad \mathbf{h} = \mathbf{M}\mathbf{a}$$

19

Authority and Hub

- Column vector ***a***: a_i is the authority score for the i-th site
- Column vector ***h***: h_i is the hub score for the i-th site
- Matrix **M**:

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

- Recursive dependency:

a(v) ← Σ w Є pa[v] h(w)

h(v) ← Σ w Є ch[v] a(w)

Normalization Procedure

$$\mathbf{a}_t = \boxed{\alpha_t} \mathbf{M}^T \mathbf{h}_t$$
$$\mathbf{h}_t = \boxed{\beta_t} \mathbf{M}\mathbf{a}_t$$

20

Authority and Hub

$$\left.\begin{aligned} \mathbf{a}_t &= \alpha_t \mathbf{M}^T \mathbf{h}_t \\ \mathbf{h}_t &= \beta_t \mathbf{M}\mathbf{a}_t \end{aligned}\right\} \rightarrow \begin{aligned} \mathbf{a}_t &= \alpha_t \beta_t \mathbf{M}^T \mathbf{M}\mathbf{a}_t \\ \mathbf{h}_t &= \alpha_t \beta_t \mathbf{M}\mathbf{M}^T \mathbf{h}_t \end{aligned}$$

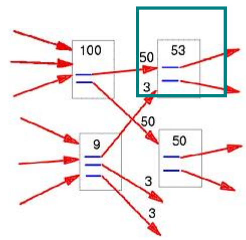
- Apply SVD to matrix **M**

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T = \sum_i \lambda_i \mathbf{u}_i \mathbf{v}_i^T \longrightarrow \mathbf{a} = \mathbf{u}_1, \mathbf{h} = \mathbf{v}_1$$

21

PageRank

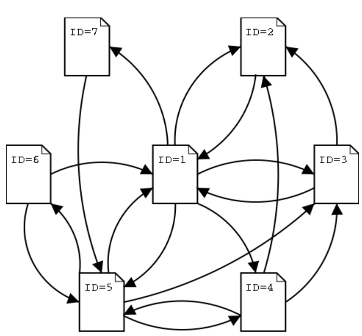
- Introduced by Page et al (1998)
  - The weight is assigned by the rank of parents
- Difference with HITS
  - HITS takes Hubness & Authority weights
  - The page rank is proportional to its parents' rank, but inversely proportional to its parents' outdegree



Matrix Notation

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

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



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

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

24

Matrix Notation

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

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

r = α **B**ᵀ **r**

α : eigenvalue

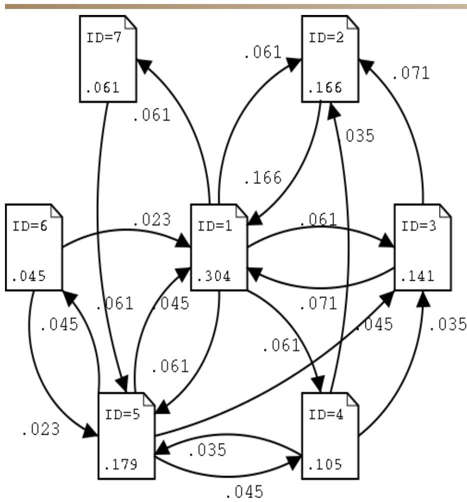
r : eigenvector of **B**

Finding Pagerank

→ finding principle eigenvector of B

25

Matrix Notation

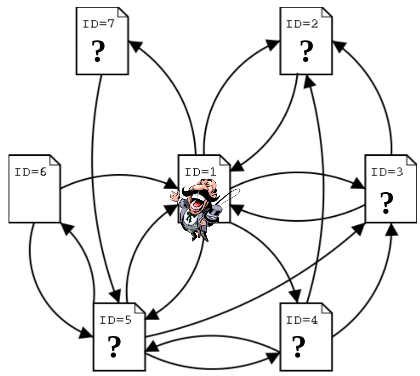


PR	ID	OutLink	InLink
0.304	1	2,3,4,5,7	2,3,5,6
0.179	5	1,3,4,6	1,4,6,7
0.166	2	1	1,3,4
0.141	3	1,2	1,4,5
0.105	4	2,3,5	1,5
0.061	7	5	1
0.045	6	1,5	5

26

Random Walk Model

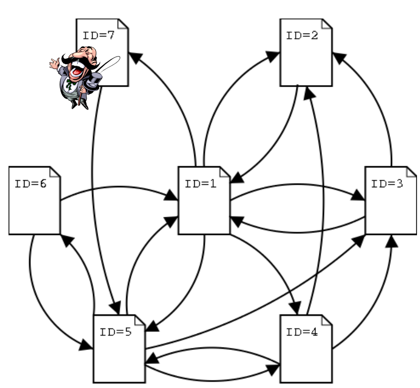
- Consider a random walk through the Web graph



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

Random Walk Model

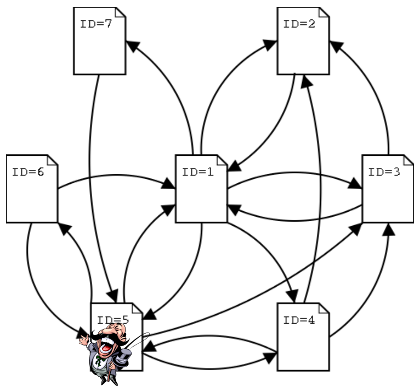
- Consider a random walk through the Web graph



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

Random Walk Model

- Consider a random walk through the Web graph

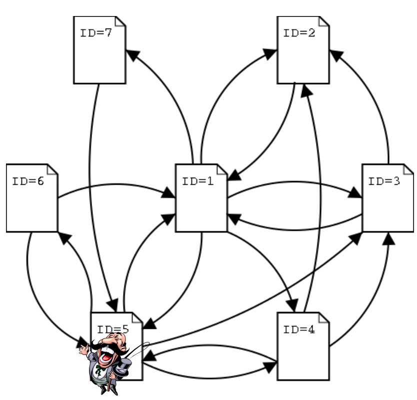


$$\mathbf{B} = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

29

Random Walk Model

- Consider a random walk through the Web graph



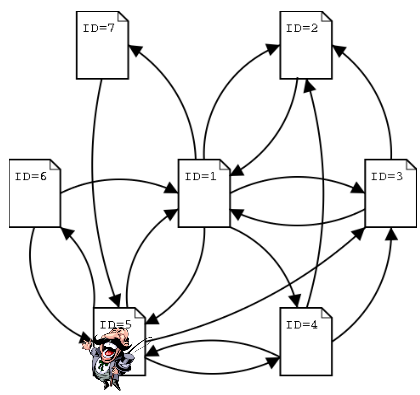
$$\mathbf{B} = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$T \rightarrow \infty$, what is portion of time that the surfer will spend time on each site?

30

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

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

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

31

PURDUE UNIVERSITY

CS54701: Information Retrieval

Link Analysis

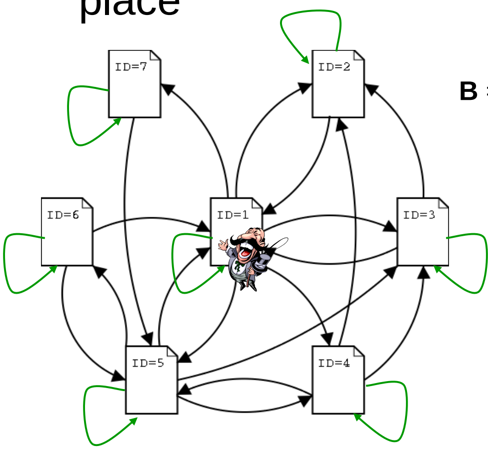
18 February 2016

Prof. Chris Clifton

Indiana Center for Database Systems ™

Adding Self Loops

- Allow surfer to decide to stay on the same place



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

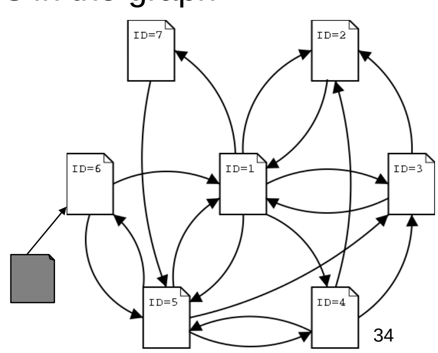
$$\mathbf{B}' = \alpha\mathbf{B} + (1-\alpha)\mathbf{I}$$

33

Problem

- "Rank Sink" Problem
  - In general, many Web pages have no inlinks/outlinks
  - Results in dangling edges in the graph

$$\mathbf{B} = \begin{array}{ccccccc|c} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

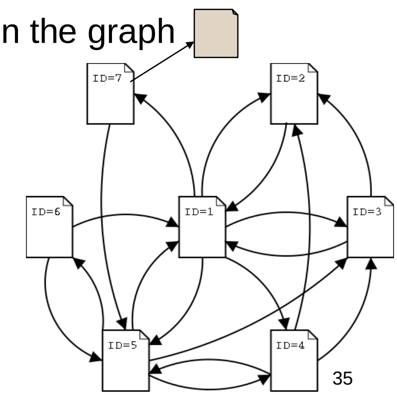


r(new page) = 0

34

Problem

- "Rank Sink" Problem
  - In general, many Web pages have no inlinks/outlinks
  - Results in dangling edges in the graph



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

r(new page) = 1

35

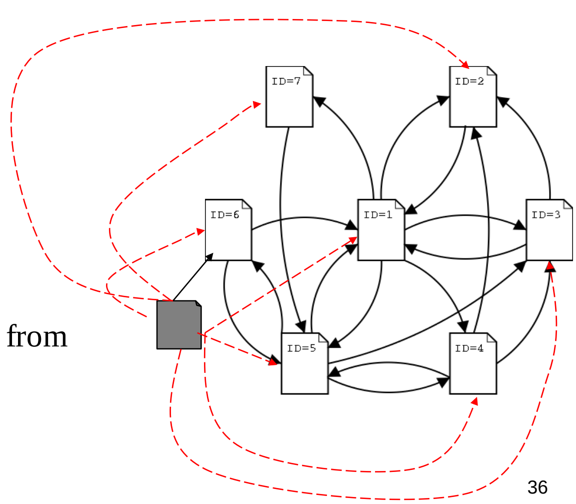
Distribution of the Mixture Model

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

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

$$\downarrow$$

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



Prevents the page ranks from being 0 or 1

36

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?

37

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

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

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

δ: eigengap $\lambda_1 - \lambda_2$

d: maximum outdegree of G

38

Stability of PageRank

$$\|\tilde{r} - r\| \le \frac{2\sum_{j\in \mathrm{V}} r(j)}{\epsilon}$$ 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

39