PURDUE UNIVERSITY

CS47300: Web Information Search and Management

Using Graph Structure for Retrieval

Prof. Chris Clifton

24 September 2019

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

PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

Citation Analysis

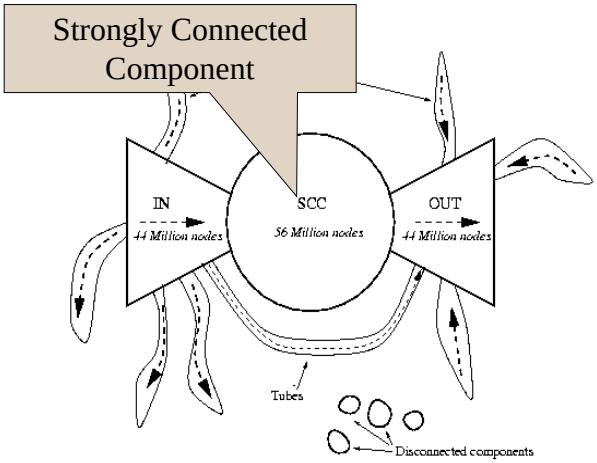
PURDUE UNIVERSITY

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

PURDUE UNIVERSITY

Bowtie Structure

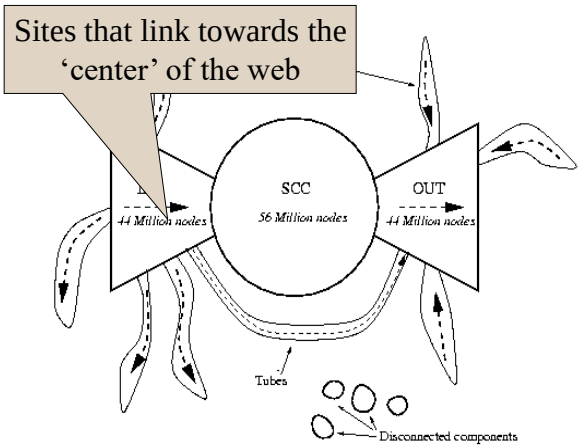


Broder et al., 2001

24

PURDUE UNIVERSITY

Bowtie Structure

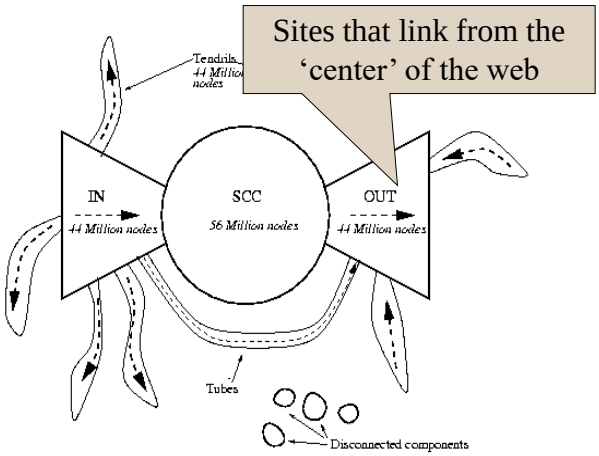


Broder et al., 2001

25

PURDUE UNIVERSITY

Bowtie Structure

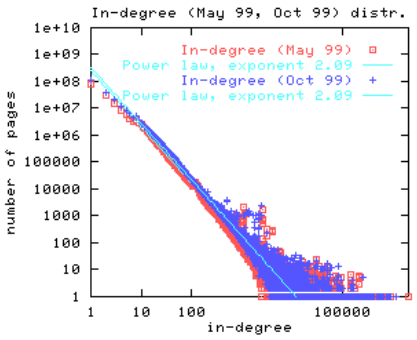


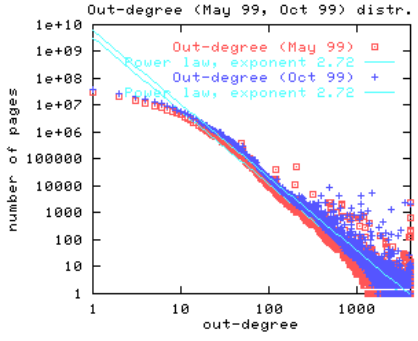
Broder et al., 2001

PURDUE UNIVERSITY

Inlinks and Outlinks

- Both degrees of incoming and outgoing links follow power law





Broder et al., 2001

PURDUE UNIVERSITY

So what good is Link Structure?

- When you can't find something, do you:
  - Keep looking in the same place?
  - Look somewhere else?
  - Give up?
  - Ask for help?
- *Other people may already know the answer!*
  - Links: Reflect human judgement

28

PURDUE UNIVERSITY

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

30

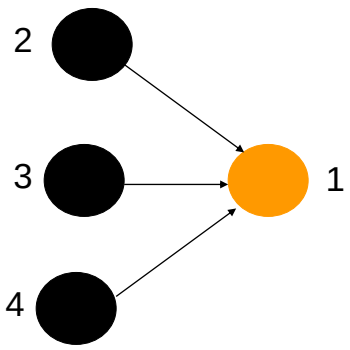
PURDUE UNIVERSITY

HITS - Kleinberg's Algorithm

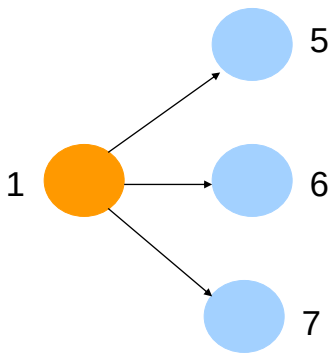
- HITS – Hypertext Induced Topic Selection
- For each vertex v ϵ V in a subgraph of interest:
  - a(v) - the authority of v
  - h(v) - the hubness of v
- A site is very authoritative if it receives many citations.
  - Citation from important sites weight more than citations from less-important sites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

PURDUE UNIVERSITY

Authority and Hubness



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



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

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)

```
1   1  ← [1,…,1] Є R^|V|
2   a_0 ← h_0 ← 1
3   t  ← 1
4   repeat
5         for each v in V
6         do a_t (v) ←   Σ_{w Є pa[v]} h_{t-1} (w)
7            h_t (v) ←   Σ_{w Є pa[v]} a_{t-1} (w)
8   t  ← t + 1
9   until || a_t – a_{t-1} || + || h_t – h_{t-1} || < ε
10  return (a_t , h_t)
```

Problems ?

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

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

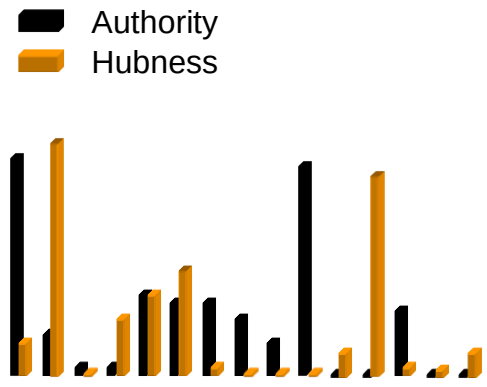
$$h(v) = \frac{h(v)}{\sum_w h(w)}$$

HubsAuthorities(G)

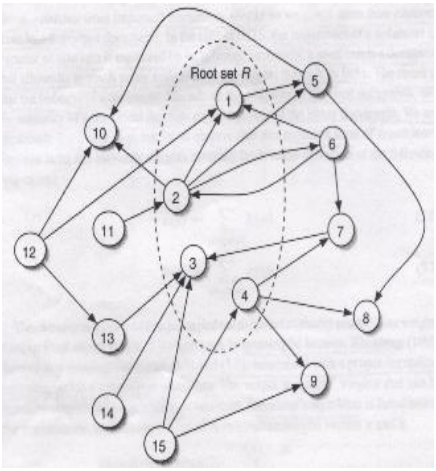
```
1   1  ← [1,…,1] Є R^|V|
2   a_0 ← h_0 ← 1
3   t  ← 1
4   repeat
5         for each v in V
6         do a_t (v) ←   Σ_{w Є pa[v]} h_{t-1} (w)
7            h_t (v) ←   Σ_{w Є pa[v]} a_{t-1} (w)
8         a_t ← a_t / || a ||
9         h_t ← h_t / || h ||
10        t  ← t + 1
11  until || a_t – a_{t-1} || + || h_t – h_{t-1} || < ε
12  return (a_t , h_t)
```

PURDUE UNIVERSITY

HITS Example Results



Authority and hubness weights



PURDUE UNIVERSITY

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

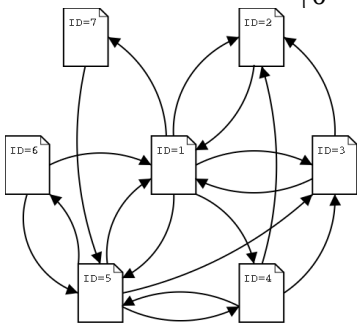
PURDUE UNIVERSITY

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**:

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



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

PURDUE UNIVERSITY

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:

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

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

PURDUE UNIVERSITY

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:

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

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

PURDUE UNIVERSITY

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**:

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

Normalization Procedure

- Recursive dependency:

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

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

PURDUE UNIVERSITY

Authority and Hub

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

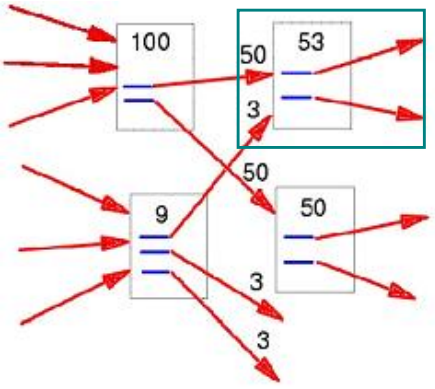
- Apply SVD to matrix **M**

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

PURDUE UNIVERSITY

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

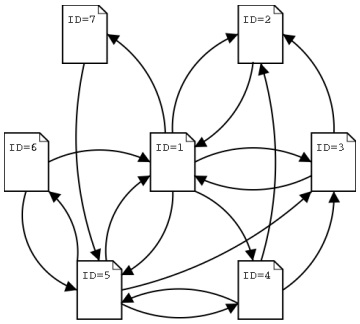


PURDUE UNIVERSITY

Matrix Notation

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

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



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

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

PURDUE UNIVERSITY

Matrix Notation

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

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

$\mathbf{r} = \alpha\, \mathbf{B}^{\mathsf{T}}\, \mathbf{r}$

α : eigenvalue

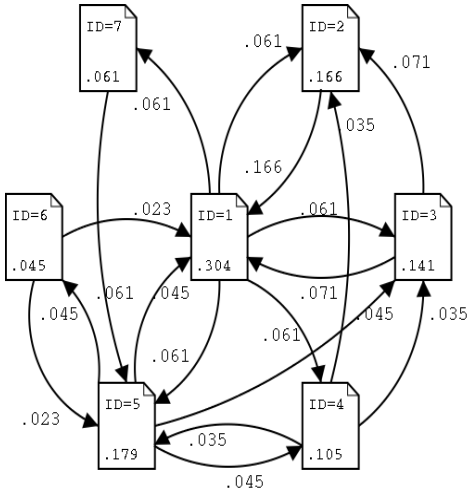
r : eigenvector of **B**

Finding Pagerank

→ find principle eigenvector of B

PURDUE UNIVERSITY

Matrix Notation



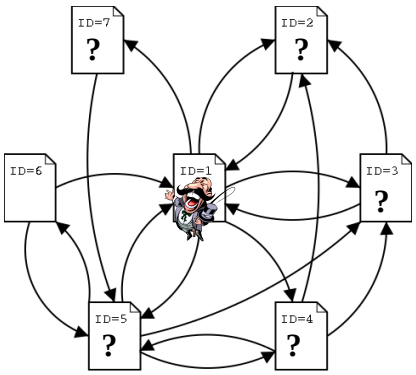
PR	ID	OutLink	InLink
0.304	1	2,3,4,5,7	2,3,5,6
0.179	5	1,3,4,6	1,4,6,7
0.166	2	1	1,3,4
0.141	3	1,2	1,4,5
0.105	4	2,3,5	1,5
0.061	7	5	1
0.045	6	1,5	5

48

PURDUE UNIVERSITY

Random Walk Model

- Consider a random walk through the Web graph



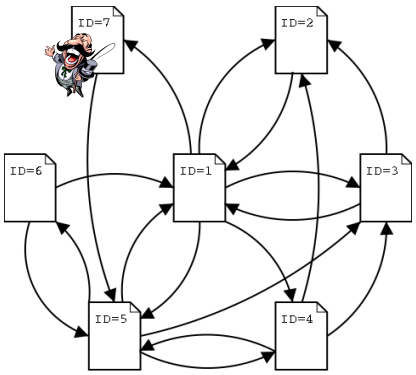
$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

49

PURDUE UNIVERSITY

Random Walk Model

- Consider a random walk through the Web graph



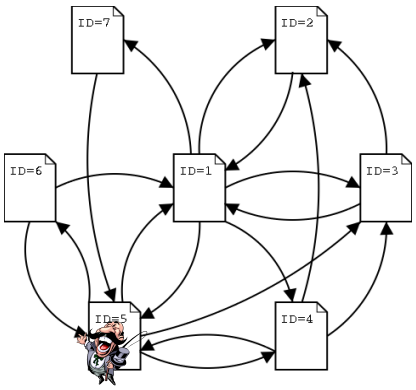
B =

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

PURDUE UNIVERSITY

Random Walk Model

- Consider a random walk through the Web graph



B =

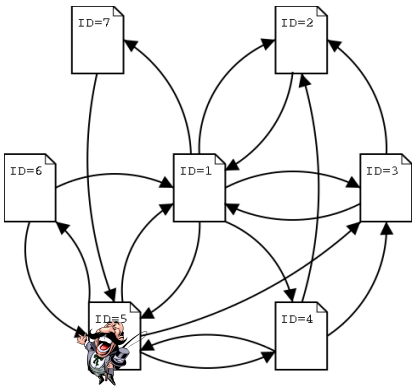
$$\begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

As T→∞, what is portion of time that the web surfer will spend at each page?

PURDUE UNIVERSITY

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

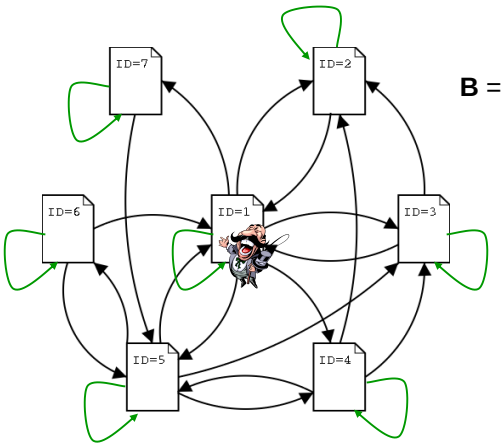
$$p(k) = \sum_i p(i)\mathbf{B}_{i,k}$$

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

PURDUE UNIVERSITY

Adding Self Loop

- Allow surfer to decide to stay on the same place



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

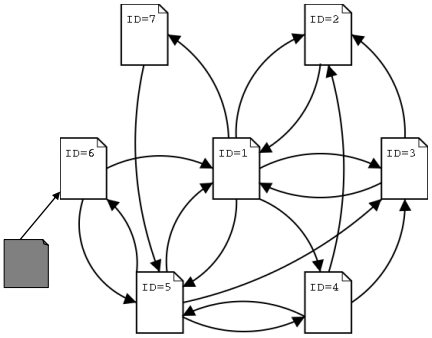
$$\mathbf{B}' = \alpha\mathbf{B} + (1-\alpha)\mathbf{I}$$

Problem

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

B =

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



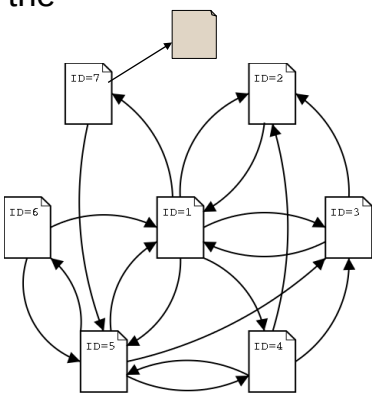
r(new page) = 0

Problem

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

B =

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



r(new page) = 1

PURDUE UNIVERSITY

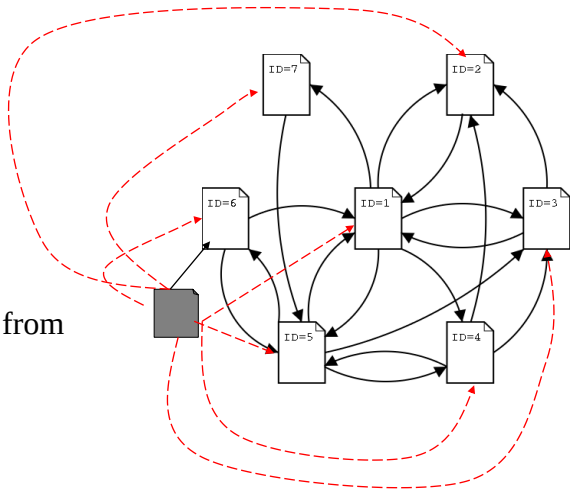
Distribution of the Mixture Model

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

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

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

Prevents the page ranks from being 0 or 1



PURDUE UNIVERSITY

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?

PURDUE UNIVERSITY

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 λ1 – λ2

d: maximum outdegree of G

PURDUE UNIVERSITY

Stability of PageRank

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