

# CS54701: Information Retrieval

*Link Analysis*

16 February 2016

Prof. Jennifer Neville



## 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



# Citation Analysis

Probabilistic Latent Semantic Analysis - Hofmann (ResearchIndex) - Microsoft Internet Explorer

Alternate document: [Details](#) **Probabilistic Latent Semantic Indexing (99)** Thomas Hofmann

**Probabilistic Latent Semantic Analysis (1999)** (Make Corrections) [46 citations](#)

Thomas Hofmann  
Proc. of Uncertainty in Artificial Intelligence, UA199

[View or download](#)  
[brown.edu/about/.../hofmann/UA199.ps](#)  
Cached: [PS](#) [PDF](#) [Image](#) [Update](#) [Help](#)

From [whitbang.com/~wochen/course](#) (more)  
(Enter author homepage)

[CiteSeer](#) [Home/Search](#) [Bookmark](#) [Context](#) [Related](#) [DRIIP Metadata](#)

[\(Enter summary\)](#) [Rate this article](#) 1 2 3 4 5 (best) [Comment on this article](#)

**Abstract:** Probabilistic Latent Semantic Analysis is a novel statistical technique for the analysis of two-mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas. Compared to standard Latent Semantic Analysis which stems from linear algebra and performs a Singular Value Decomposition of co-occurrence tables, the proposed method is based on a mixture decomposition derived from a latent class... [\(Update\)](#)

**Cited by:** [More](#)  
A Hierarchical Model for Clustering and - Categorizing Documents [Gaussier \(2002\)](#) (Correct)  
Probabilistic Models for Hierarchical Clustering and - Society Eric Gaussier (2002) (Correct)  
Name That Song!: A Probabilistic Approach - To Querying On [\(Correct\)](#)

**Similar documents (at the sentence level):**  
**0.7%** Probabilistic Latent Semantic Indexing - Hofmann (1999) (Correct)

**Active bibliography (related documents):** [More](#) [All](#)  
**0.3** Unsupervised Learning from Dyadic Data - Hofmann, al (1998) (Correct)  
**0.7** Language Model Adaptation - Gotoh (2000) (Correct)  
**0.7** Topic-Based Language Models Using EM - Gildea, Hofmann (1999) (Correct)

**Similar documents based on text:** [More](#) [All](#)  
**0.3** Bayesian Latent Semantic Analysis of Multimedia Databases - de Freitas, Barnard (2001) (Correct)  
**0.3** Essay Assessment with Latent Semantic Analysis - Miller (2003) (Correct)  
**0.3** Latent Dirichlet Allocation - Blei, Ng, Jordan (2001) (Correct)

**Related documents from co-citation:** [More](#) [All](#)  
**15** Indexing by latent semantic analysis - Deerwester, Dumais et al. - 1990  
**14** Maximum Likelihood from Incomplete Data via the EM Algorithm (context) - Dempster, Laird et al. - 1977  
**10** Latent Dirichlet allocation - Blei, Ng et al. - 2002

3



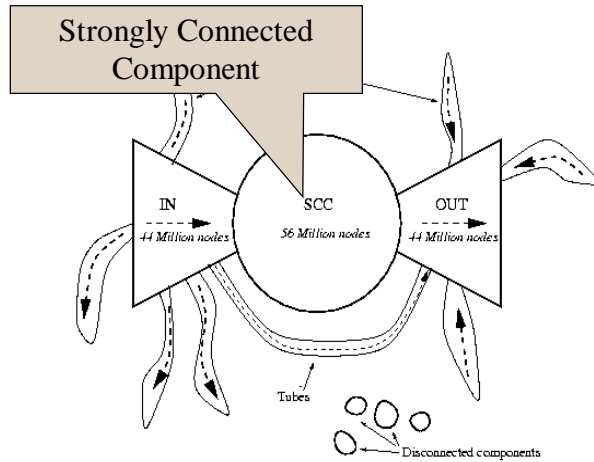
# Ad-Hoc Retrieval: Beyond the Words

- 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

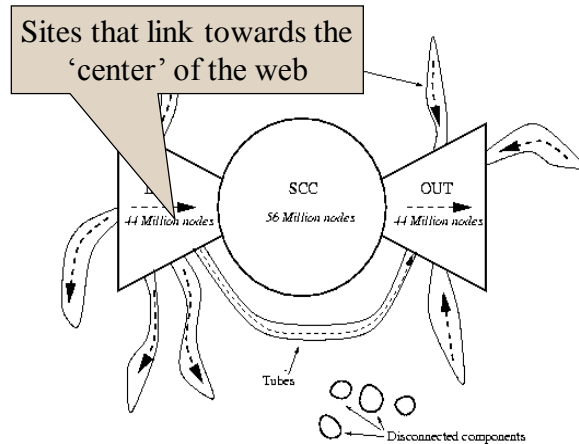


Broder et al., 2001

6



# Bowtie Structure

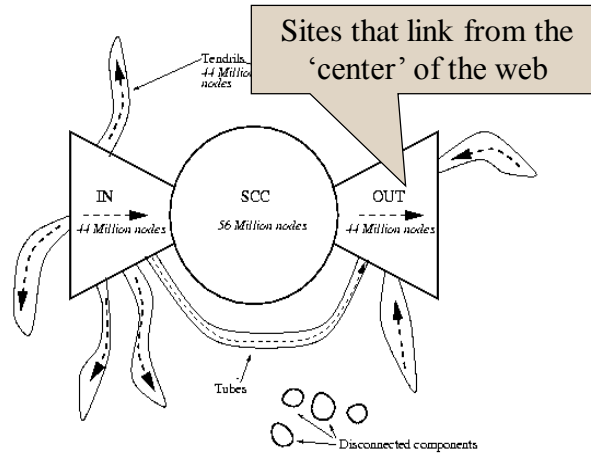


Broder et al., 2001

7



# Bowtie Structure



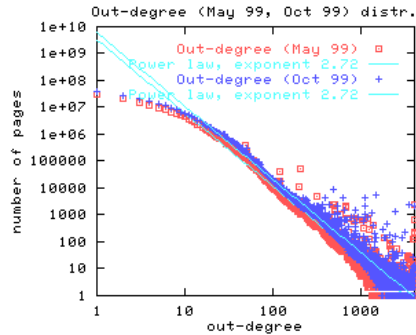
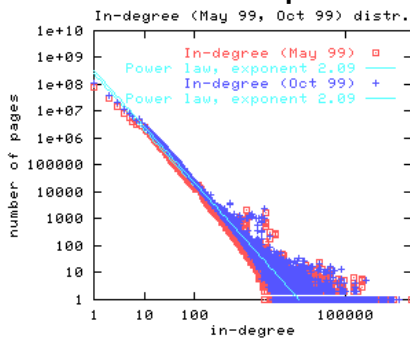
Broder et al., 2001

8



# Inlinks and Outlinks

- Both degrees of incoming and outgoing links follow power law



Broder et al., 2001

9



## 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



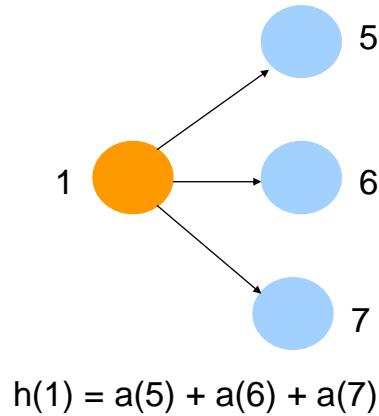
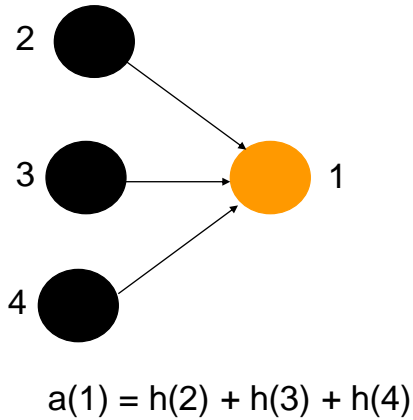
## HITS - Kleinberg's Algorithm

- HITS – Hypertext Induced Topic Selection
- For each vertex  $v \in V$  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



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)

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

2  $\mathbf{a}_0 \leftarrow \mathbf{h}_0 \leftarrow \mathbf{1}$

3  $t \leftarrow 1$

4 repeat

5     for each  $v$  in  $V$

6         do  $\mathbf{a}_t(v) \leftarrow \sum_{w \in pa[v]} \mathbf{h}_{t-1}(w)$

7          $\mathbf{h}_t(v) \leftarrow \sum_{w \in ch[v]} \mathbf{a}_{t-1}(w)$

8  $t \leftarrow t + 1$

9 until  $\|\mathbf{a}_t - \mathbf{a}_{t-1}\| + \|\mathbf{h}_t - \mathbf{h}_{t-1}\| < \epsilon$

10 return  $(\mathbf{a}_t, \mathbf{h}_t)$

**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)<sup>|V|</sup>

- 1  $1 \leftarrow [1, \dots, 1] \in \mathbb{R}$
- 2  $a_0 \leftarrow h_0 \leftarrow 1$
- 3  $t \leftarrow 1$
- 4 repeat
- 5     for each  $v$  in  $V$
- 6         do  $a_t(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$
- 7              $h_t(v) \leftarrow \sum_{w \in ch[v]} a_{t-1}(w)$
- 8              $a_t \leftarrow a_t / \|a\|$
- 9              $h_t \leftarrow h_t / \|h\|$
- 10          $t \leftarrow t + 1$
- 11 until  $\|a_t - a_{t-1}\| + \|h_t - h_{t-1}\| < \epsilon$
- 12 return  $(a_t, h_t)$

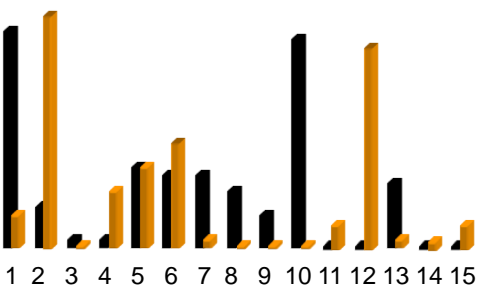
14



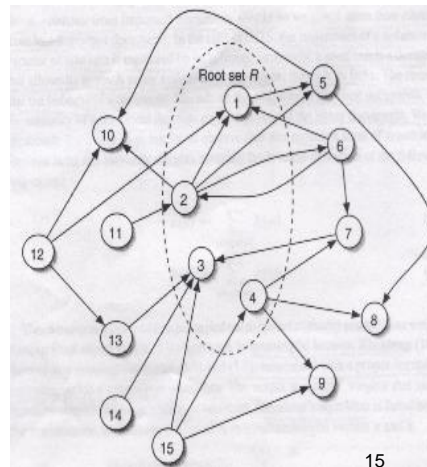
# HITS Example Results



- Authority
- Hubness



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

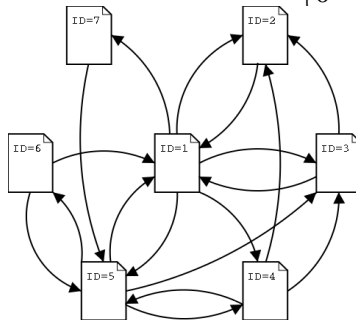


# Authority and Hub

- Column vector  $\mathbf{a}$ :  $a_i$  is the authority score for the  $i$ -th site
- Column vector  $\mathbf{h}$ :  $h_i$  is the hub score for the  $i$ -th site

- Matrix  $\mathbf{M}$ :

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



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

17





# Authority and Hub

- Vector  $\mathbf{a}$ :  $a_i$  is the authority score for the  $i$ -th site
- Vector  $\mathbf{h}$ :  $h_i$  is the hub score for the  $i$ -th site

- Matrix  $\mathbf{M}$ :

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

- Recursive dependency:

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

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

18



# Authority and Hub

- Column vector  $\mathbf{a}$ :  $a_i$  is the authority score for the  $i$ -th site
- Column vector  $\mathbf{h}$ :  $h_i$  is the hub score for the  $i$ -th site

- Matrix  $\mathbf{M}$ :

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

- Recursive dependency:

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

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



$$\mathbf{a} = \mathbf{M}^T \mathbf{h}$$

$$\mathbf{h} = \mathbf{M} \mathbf{a}$$

19



# Authority and Hub

- Column vector  $\mathbf{a}$ :  $a_i$  is the authority score for the  $i$ -th site
- Column vector  $\mathbf{h}$ :  $h_i$  is the hub score for the  $i$ -th site
- Matrix  $\mathbf{M}$ :

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

Normalization Procedure

- Recursive dependency:

$$\begin{aligned} a(v) &\leftarrow \sum_{w \in \text{pa}[v]} h(w) & \mathbf{a}_t &= \alpha_t \mathbf{M}^T \mathbf{h}_t \\ h(v) &\leftarrow \sum_{w \in \text{ch}[v]} a(w) & \mathbf{h}_t &= \beta_t \mathbf{M} \mathbf{a}_t \end{aligned}$$

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  $\mathbf{M}$

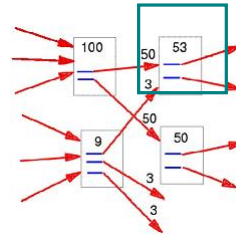
$$\mathbf{M} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^T = \sum_i \lambda_i \mathbf{u}_i \mathbf{v}_i^T \longrightarrow \mathbf{a} = \mathbf{u}_1, \mathbf{h} = \mathbf{v}_1$$

21



# PageRank

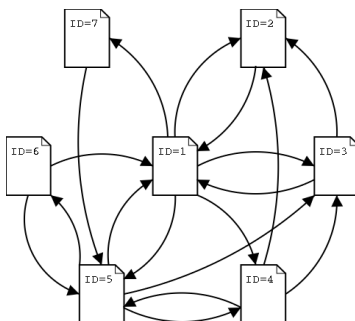
- Introduced by Page et al (1998)
  - The weight is assigned by the rank of parents
- Difference with HITS
  - HITS takes Hubness & Authority weights
  - The page rank is proportional to its parents' rank, but inversely proportional to its parents' outdegree



# Matrix Notation

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

$$B_{i,j} = \begin{cases} \frac{1}{\sum_j M_{i,j}} & \sum_j M_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$



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

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



# Matrix Notation

$r : r_i$  represents the rank score for the  $i$ -th web page

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

$$r = \alpha B^T r$$

$\alpha$  : eigenvalue

$r$  : eigenvector of  $B$

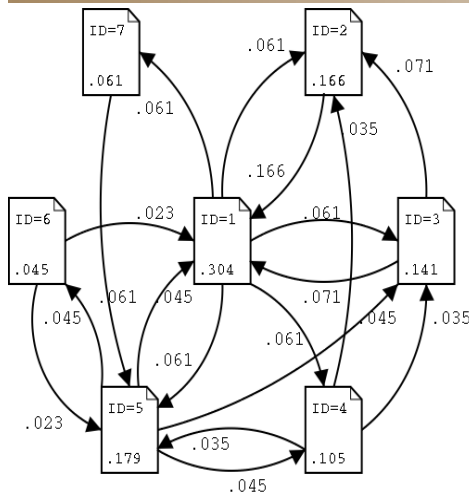
Finding Pagerank

→ finding principle eigenvector of  $B$

25



# Matrix Notation



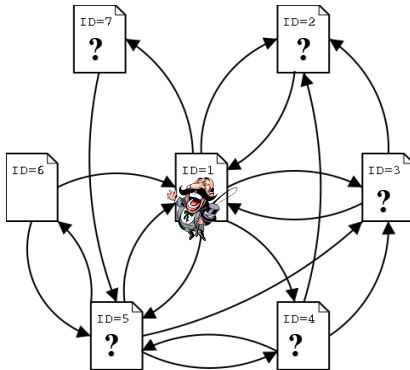
PR	ID	OutLink	InLink
<b>0.304</b>	<b>1</b>	<b>2,3,4,5,7</b>	<b>2,3,5,6</b>
<b>0.179</b>	<b>5</b>	<b>1,3,4,6</b>	<b>1,4,6,7</b>
<b>0.166</b>	<b>2</b>	<b>1</b>	<b>1,3,4</b>
<b>0.141</b>	<b>3</b>	<b>1,2</b>	<b>1,4,5</b>
<b>0.105</b>	<b>4</b>	<b>2,3,5</b>	<b>1,5</b>
<b>0.061</b>	<b>7</b>	<b>5</b>	<b>1</b>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

26



# Random Walk Model

- Consider a random walk through the Web graph



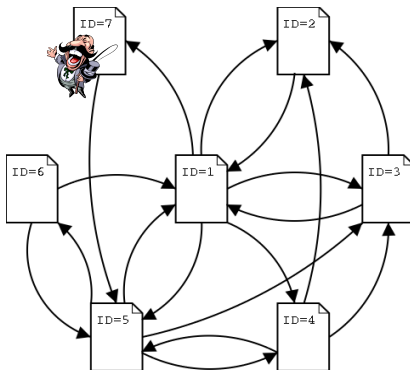
$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

27



# Random Walk Model

- Consider a random walk through the Web graph



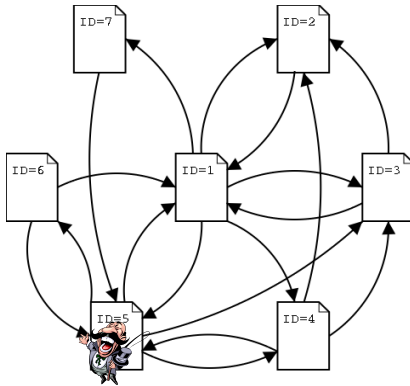
$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

28



# Random Walk Model

- Consider a random walk through the Web graph



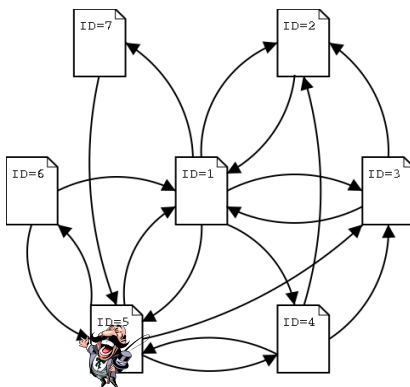
$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

29



# Random Walk Model

- Consider a random walk through the Web graph



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

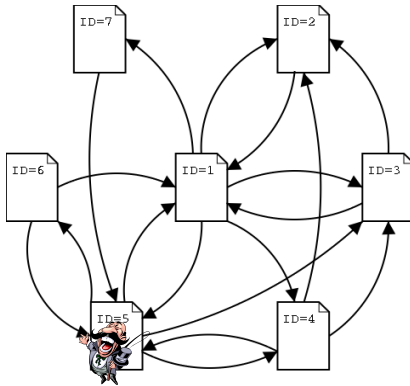
$T \rightarrow \infty$ , what is portion of time that the surfer will spend time on each site?

30



# Random Walk Model

- Consider a random walk through the Web graph



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

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

$$p(k) = \sum_i p(i) B_{i,k}$$

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

31

**PURDUE**  
UNIVERSITY

## CS54701: Information Retrieval

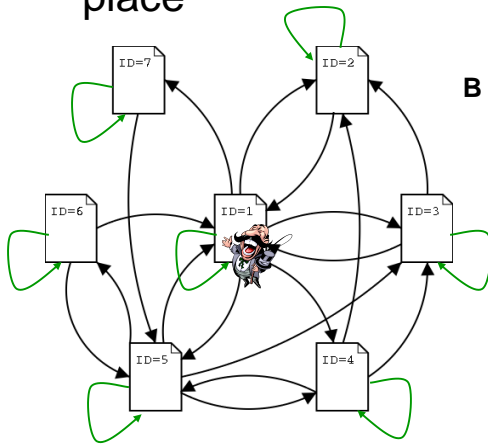
*Link Analysis*  
18 February 2016  
Prof. Chris Clifton





# Adding Self Loops

- Allow surfer to decide to stay on the same place



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

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

33

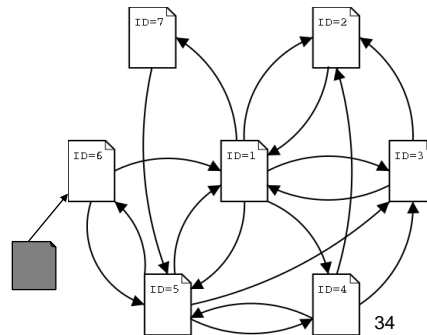


# Problem

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

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

$$r(\text{new page}) = 0$$



34





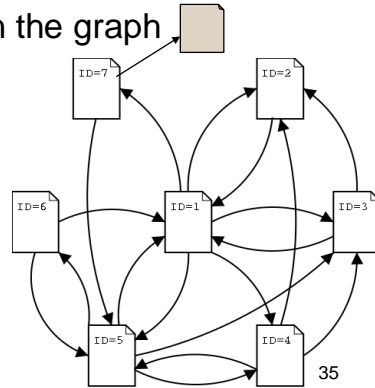
# Problem

## • “Rank Sink” Problem

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

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

$$r(\text{new page}) = 1$$



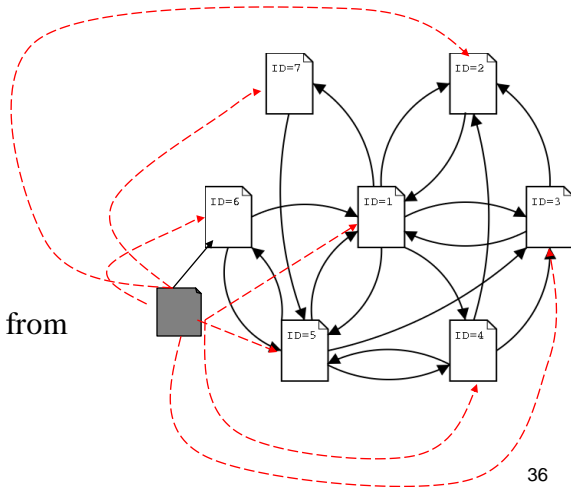
# Distribution of the Mixture Model

$$H_{i,j} = 1/n$$

$$B' = \epsilon H + (1 - \epsilon)B$$

$$r = B'^T r$$

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?

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  $\bar{\delta}$  that produces a constant perturbation of the dominant eigenvector

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

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

$\bar{\delta}$ : eigengap  $\lambda_1 - \lambda_2$

$d$ : maximum outdegree of  $G$

38



# Stability of PageRank

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

$V$ : the set of vertices touched by the perturbation

- The parameter  $\epsilon$  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