



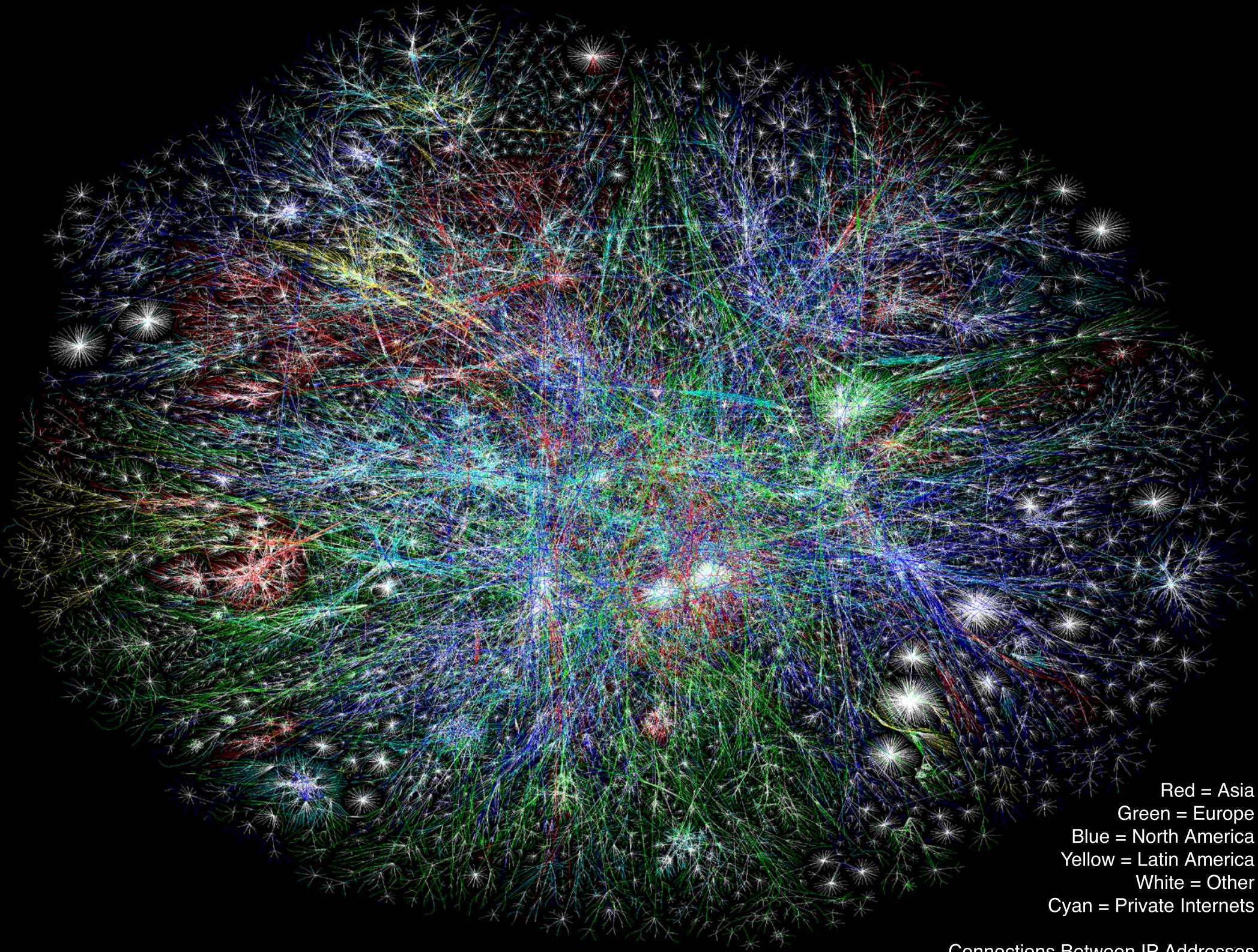
# Information Retrieval

# Web Search


Carl Meyer  
Amy Langville

Department of Mathematics  
North Carolina State University  
Raleigh, NC

MAA Short Course 3/11/2005



Connections Between IP Addresses

A man with a beard and mustache, wearing a dark sweater, is sitting on a light-colored couch. He is looking slightly to his right. Behind him are bookshelves filled with books. A white banner at the bottom of the frame contains the text "Christian Pilet".

**Christian Pilet**





















# Limitations of LSI

- Rankings are query dependent
  - Rank of each doc is recomputed for each query
- Only semantic content is used
  - Link structure completely ignored
- Difficult to add & delete documents
  - Requires updating & downdating SVD
- Determining optimal  $k$  is not easy
  - Empirical tuning required
- Doesn't scale up well
  - Impractical for www



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$

## Importance Rankings

- Attach an “importance rank”  $r_i$  to each page:  $P_i \hookrightarrow r_i$



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$

## Importance Rankings

- Attach an “importance rank”  $r_i$  to each page:  $P_i \hookrightarrow r_i$   
—  $r_i$  based on link structure (i.e., query independent)



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$

## Importance Rankings

- Attach an “importance rank”  $r_i$  to each page:  $P_i \hookrightarrow r_i$ 
  - $r_i$  based on link structure (i.e., query independent)
  - $r_i$  computed prior to any query



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$

## Importance Rankings

- Attach an “importance rank”  $r_i$  to each page:  $P_i \hookrightarrow r_i$ 
  - $r_i$  based on link structure (i.e., query independent)
  - $r_i$  computed prior to any query

## Direct Query Matching

- Query =  $(Term_1, Term_2) \longrightarrow (P_i, r_i), (P_j, r_j), (P_k, r_k), \dots$



# Using Link Structure

## Indexing

- Still must index key terms on each page  
Robots crawl the web — software does indexing
- File structure: Terms  $\longrightarrow$  Pages (similar to book index)  
 $Term_1 \rightarrow P_i, P_j, \dots$   
 $Term_2 \rightarrow P_k, P_l, \dots$   
 $\vdots$

## Importance Rankings

- Attach an “importance rank”  $r_i$  to each page:  $P_i \hookrightarrow r_i$ 
  - $r_i$  based on link structure (i.e., query independent)
  - $r_i$  computed prior to any query

## Direct Query Matching

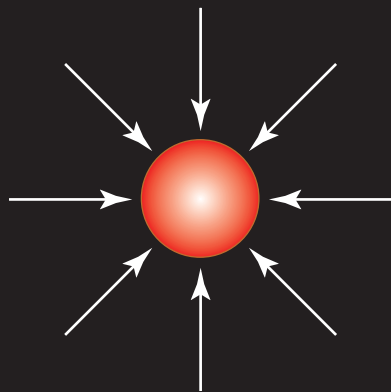
- Query =  $(Term_1, Term_2) \longrightarrow (P_i, r_i), (P_j, r_j), (P_k, r_k), \dots$

**Return  $P_i, P_j, P_k, \dots$  in order of ranks  $r_i, r_j, r_k, \dots$**

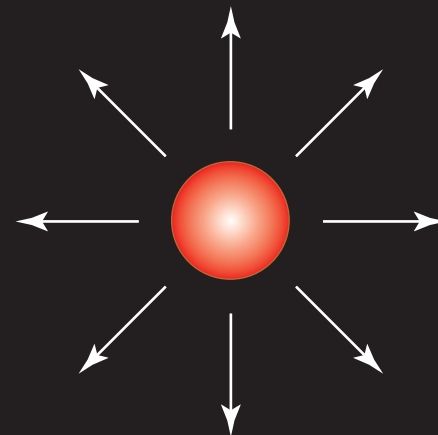


# How To Measure “Importance”

Authorities

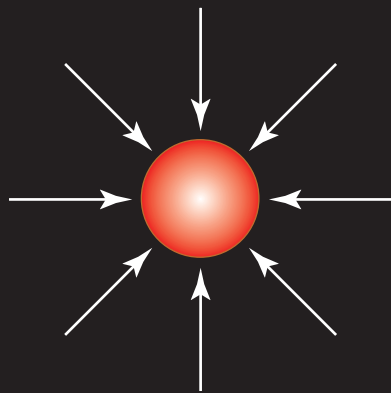


Hubs

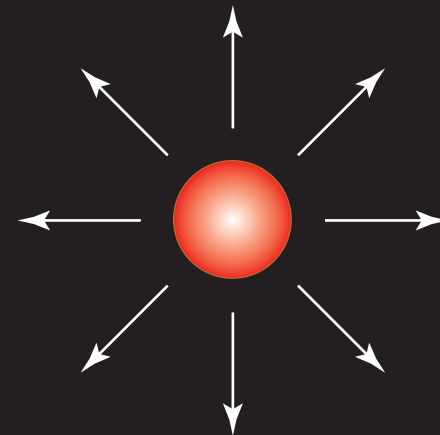


# How To Measure “Importance”

Authorities



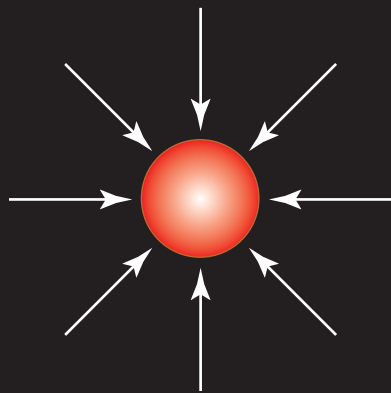
Hubs



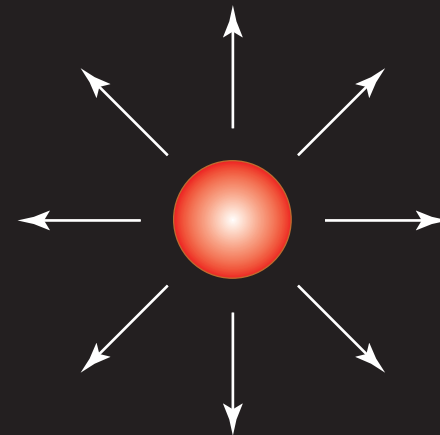
- Good hub pages point to good authority pages

# How To Measure “Importance”

Authorities



Hubs



- Good hub pages point to good authority pages
- Good authorities are pointed to by good hubs



# HITS Algorithm

Hypertext Induced Topic Search (1998)

## Determine Authority & Hub Scores

- $a_i$  = authority score for  $P_i$
- $h_i$  = hub score for  $P_i$



Jon Kleinberg



# HITS Algorithm

Hypertext Induced Topic Search (1998)



Jon Kleinberg

## Determine Authority & Hub Scores

- $a_i$  = authority score for  $P_i$
- $h_i$  = hub score for  $P_i$

## Successive Refinement

- Start with  $h_i = 1$  for all pages  $P_i \Rightarrow \mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$



# HITS Algorithm

Hypertext Induced Topic Search (1998)



Jon Kleinberg

## Determine Authority & Hub Scores

- $a_i$  = authority score for  $P_i$
- $h_i$  = hub score for  $P_i$

## Successive Refinement

- Start with  $h_i = 1$  for all pages  $P_i \Rightarrow \mathbf{h}_0 =$
- Define Authority Scores (first iterate)

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$a_i = \sum_{j:P_j \rightarrow P_i} h_j$$



# HITS Algorithm

Hypertext Induced Topic Search (1998)



Jon Kleinberg

## Determine Authority & Hub Scores

- $a_i$  = authority score for  $P_i$
- $h_i$  = hub score for  $P_i$

## Successive Refinement

- Start with  $h_i = 1$  for all pages  $P_i \Rightarrow \mathbf{h}_0 =$
- Define Authority Scores (first iterate)

$$\mathbf{h}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$a_i = \sum_{j:P_j \rightarrow P_i} h_j \Rightarrow \mathbf{a}_1 = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{L}^T \mathbf{h}_0$$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_2 = \mathbf{L}^T \mathbf{h}_1$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_2 = \mathbf{L}^T \mathbf{h}_1$

- $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_2 = \mathbf{L}^T \mathbf{h}_1$

- $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$

- $\mathbf{a}_3 = \mathbf{L}^T \mathbf{h}_2$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_2 = \mathbf{L}^T \mathbf{h}_1$

- $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$

- $\mathbf{a}_3 = \mathbf{L}^T \mathbf{h}_2$

- $\mathbf{h}_3 = \mathbf{L}\mathbf{a}_3$





# HITS Algorithm

## Refine Hub Scores

$$\bullet h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1 \quad L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

$$\begin{aligned} \bullet \mathbf{a}_2 &= \mathbf{L}^T \mathbf{h}_1 \\ \bullet \mathbf{h}_2 &= \mathbf{L} \mathbf{a}_2 \\ \bullet \mathbf{a}_3 &= \mathbf{L}^T \mathbf{h}_2 \\ \bullet \mathbf{h}_3 &= \mathbf{L} \mathbf{a}_3 \end{aligned}$$



## Combined Iterations

$$\begin{aligned} \bullet \mathbf{A} &= \mathbf{L}^T \mathbf{L} \text{ (authority matrix)} & \mathbf{a}_k &= \mathbf{A} \mathbf{a}_{k-1} \\ \bullet \mathbf{H} &= \mathbf{L} \mathbf{L}^T \text{ (hub matrix)} & \mathbf{h}_k &= \mathbf{H} \mathbf{h}_{k-1} \end{aligned}$$



# HITS Algorithm

## Refine Hub Scores

- $h_i = \sum_{j:P_i \rightarrow P_j} a_j \Rightarrow \mathbf{h}_1 = \mathbf{L}\mathbf{a}_1$

$$L_{ij} = \begin{cases} 1 & P_i \rightarrow P_j \\ 0 & P_i \not\rightarrow P_j \end{cases}$$

## Successively Re-refine Authority & Hub Scores

- $\mathbf{a}_2 = \mathbf{L}^T \mathbf{h}_1$ 
  - $\mathbf{h}_2 = \mathbf{L}\mathbf{a}_2$ 
    - $\mathbf{a}_3 = \mathbf{L}^T \mathbf{h}_2$ 
      - $\mathbf{h}_3 = \mathbf{L}\mathbf{a}_3$



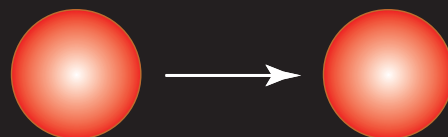
## Combined Iterations

- $\mathbf{A} = \mathbf{L}^T \mathbf{L}$  (authority matrix)       $\mathbf{a}_k = \mathbf{A}\mathbf{a}_{k-1} \rightarrow$  e-vector      (direction)
- $\mathbf{H} = \mathbf{L}\mathbf{L}^T$  (hub matrix)       $\mathbf{h}_k = \mathbf{H}\mathbf{h}_{k-1} \rightarrow$  e-vector      (direction)



# Compromise

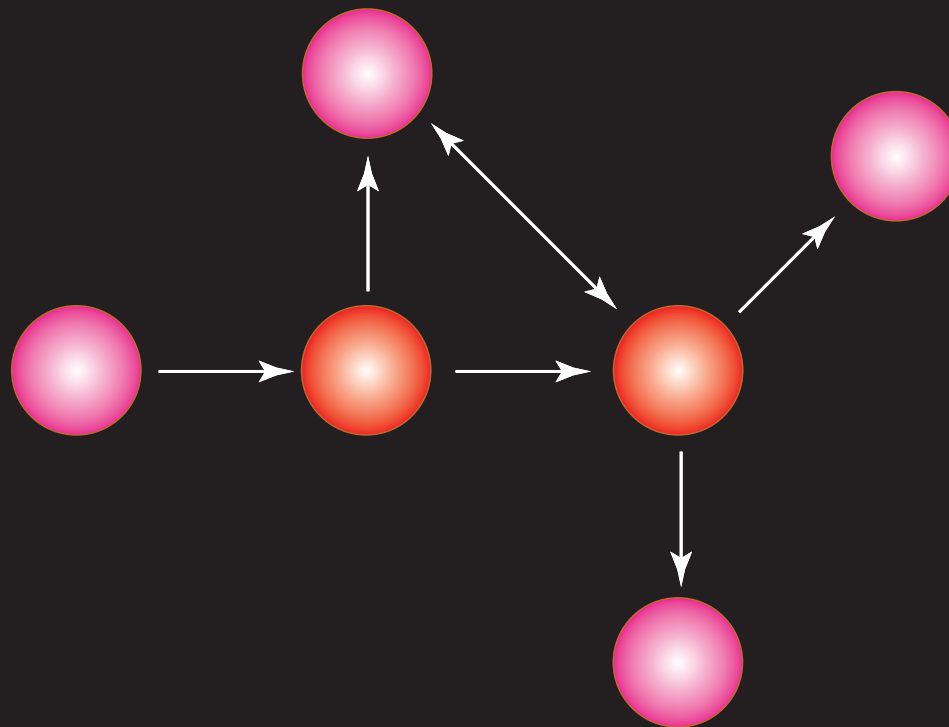
1. Do direct query matching





# Compromise

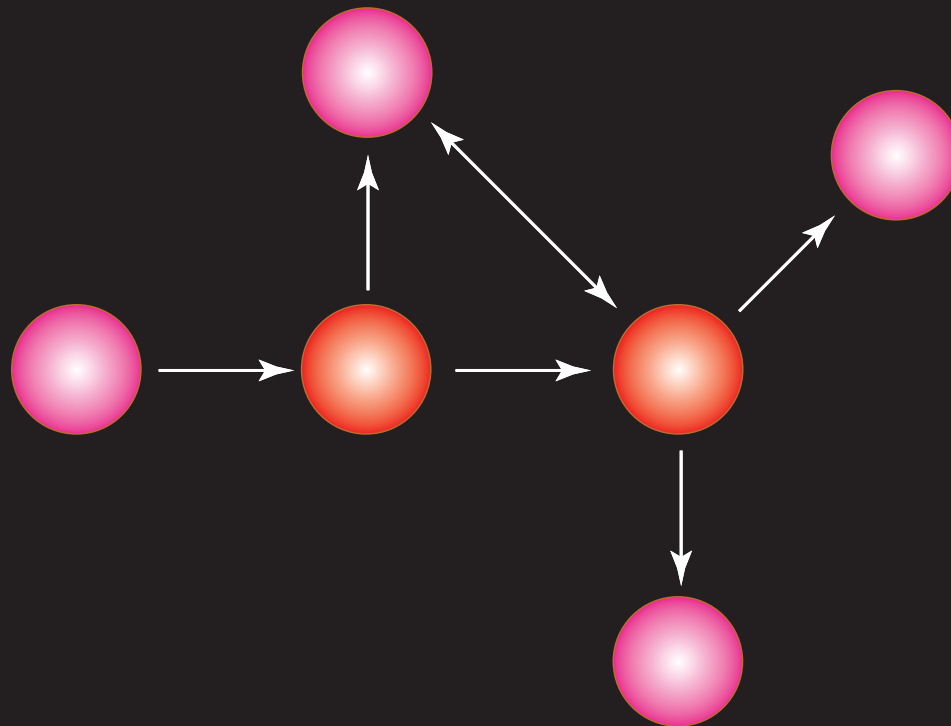
1. Do direct query matching
2. Build neighborhood graph





# Compromise

1. Do direct query matching
2. Build neighborhood graph



3. Compute authority & hub scores for just the neighborhood



# Pros & Cons

## Advantages

- Returns satisfactory results



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility for making refinements



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility for making refinements

## Disadvantages

- Too much has to happen while client is waiting



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility for making refinements

## Disadvantages

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility for making refinements

## Disadvantages

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query
  - Two eigenvector computations needed for each query



# Pros & Cons

## Advantages

- Returns satisfactory results
  - Client gets both authority & hub scores
- Some flexibility for making refinements

## Disadvantages

- Too much has to happen while client is waiting
  - Custom built neighborhood graph needed for each query
  - Two eigenvector computations needed for each query
- Scores can be manipulated by creating artificial hubs

# Newsweek

March 29, 2004 : \$3.95

newsweek.msnbc.com

The Next Frontiers

## The New Age of Google

The Search Giant Has Changed  
Our Lives. Can Anybody  
Catch These Guys? **By Steven Levy**

Google founders Larry Page and Sergey Brin



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

**PageRank  $r(P)$  Is Not Query Dependent**



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
    - No computation at query time



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
  - No computation at query time

## $r(P)$ Depends On Ranks Of Pages Pointing To $P$

- Importance is not number of in-links or out-links



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
  - No computation at query time

## $r(P)$ Depends On Ranks Of Pages Pointing To $P$

- Importance is not number of in-links or out-links
  - One link to  $P$  from Yahoo! is important



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
    - No computation at query time

## $r(P)$ Depends On Ranks Of Pages Pointing To $P$

- Importance is not number of in-links or out-links
  - One link to  $P$  from Yahoo! is important
    - Many links to  $P$  from me is not



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
  - No computation at query time

## $r(P)$ Depends On Ranks Of Pages Pointing To $P$

- Importance is not number of in-links or out-links
  - One link to  $P$  from Yahoo! is important
  - Many links to  $P$  from me is not

## PageRank Shares The Vote

- Yahoo! casts many “votes”  $\implies$  value of vote from  $Y$  is diluted



# Google's PageRank

(Lawrence Page & Sergey Brin 1998)

## PageRank $r(P)$ Is Not Query Dependent

- Depends primarily on link structure of web
  - Off-line calculations
  - No computation at query time

## $r(P)$ Depends On Ranks Of Pages Pointing To $P$

- Importance is not number of in-links or out-links
  - One link to  $P$  from Yahoo! is important
  - Many links to  $P$  from me is not

## PageRank Shares The Vote

- Yahoo! casts many “votes”  $\implies$  value of vote from  $Y$  is diluted
  - If Yahoo! “votes” for  $n$  pages
  - then  $P$  receives only  $r(Y)/n$  credit from  $Y$



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$

## Successive Refinement

Start with  $r_0(P_i) = 1/n$  for all pages  $P_1, P_2, \dots, P_n$



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$

## Successive Refinement

Start with  $r_0(P_i) = 1/n$  for all pages  $P_1, P_2, \dots, P_n$

Iteratively refine rankings for each page



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$

## Successive Refinement

Start with  $r_0(P_i) = 1/n$  for all pages  $P_1, P_2, \dots, P_n$

Iteratively refine rankings for each page

$$r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|}$$



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$

## Successive Refinement

Start with  $r_0(P_i) = 1/n$  for all pages  $P_1, P_2, \dots, P_n$

Iteratively refine rankings for each page

$$r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|}$$

$$r_2(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_1(P)}{|P|}$$



# PageRank

## The Definition

$$r(P) = \sum_{P \in \mathcal{B}_P} \frac{r(P)}{|P|}$$

$\mathcal{B}_P = \{\text{all pages pointing to } P\}$

$|P| = \text{number of out links from } P$

## Successive Refinement

Start with  $r_0(P_i) = 1/n$  for all pages  $P_1, P_2, \dots, P_n$

Iteratively refine rankings for each page

$$r_1(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_0(P)}{|P|}$$

$$r_2(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_1(P)}{|P|}$$

$\vdots$

$$r_{j+1}(P_i) = \sum_{P \in \mathcal{B}_{P_i}} \frac{r_j(P)}{|P|}$$



# In Matrix Notation

**After Step  $j$**

$$\boldsymbol{\pi}_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$



# In Matrix Notation

**After Step  $j$**

$$\boldsymbol{\pi}_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\boldsymbol{\pi}_{j+1}^T = \boldsymbol{\pi}_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$



# In Matrix Notation

**After Step  $j$**

$$\pi_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{PageRank} = \lim_{j \rightarrow \infty} \pi_j^T = \pi^T \quad (\text{provided limit exists})$$



# In Matrix Notation

## After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{PageRank} = \lim_{j \rightarrow \infty} \pi_j^T = \pi^T \quad (\text{provided limit exists})$$

## Maybe It's A Markov Chain?

$$\text{If } \mathbf{P} = [p_{ij}] \text{ is a stochastic matrix} \quad ( p_{ij} \geq 0 \text{ and } \sum_j p_{ij} = 1 )$$



# In Matrix Notation

## After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{PageRank} = \lim_{j \rightarrow \infty} \pi_j^T = \pi^T \quad (\text{provided limit exists})$$

## Maybe It's A Markov Chain?

If  $\mathbf{P} = [p_{ij}]$  is a stochastic matrix (  $p_{ij} \geq 0$  and  $\sum_j p_{ij} = 1$  )

Each  $\pi_j^T$  is a probability vector (  $\pi_i \geq 0$  and  $\sum_i \pi_i = 1$  )



# In Matrix Notation

## After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{PageRank} = \lim_{j \rightarrow \infty} \pi_j^T = \pi^T \quad (\text{provided limit exists})$$

## Maybe It's A Markov Chain?

If  $\mathbf{P} = [p_{ij}]$  is a stochastic matrix (  $p_{ij} \geq 0$  and  $\sum_j p_{ij} = 1$  )

Each  $\pi_j^T$  is a probability vector (  $\pi_i \geq 0$  and  $\sum_i \pi_i = 1$  )

$\pi_{j+1}^T = \pi_j^T \mathbf{P}$  is random walk on the graph defined by links



# In Matrix Notation

## After Step $j$

$$\pi_j^T = [r_j(P_1), r_j(P_2), \dots, r_j(P_n)]$$

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \quad \text{where} \quad p_{ij} = \begin{cases} 1/|P_i| & \text{if } i \rightarrow j \\ 0 & \text{otherwise} \end{cases}$$

$$\text{PageRank} = \lim_{j \rightarrow \infty} \pi_j^T = \pi^T \quad (\text{provided limit exists})$$

## Maybe It's A Markov Chain?

If  $\mathbf{P} = [p_{ij}]$  is a stochastic matrix (  $p_{ij} \geq 0$  and  $\sum_j p_{ij} = 1$  )

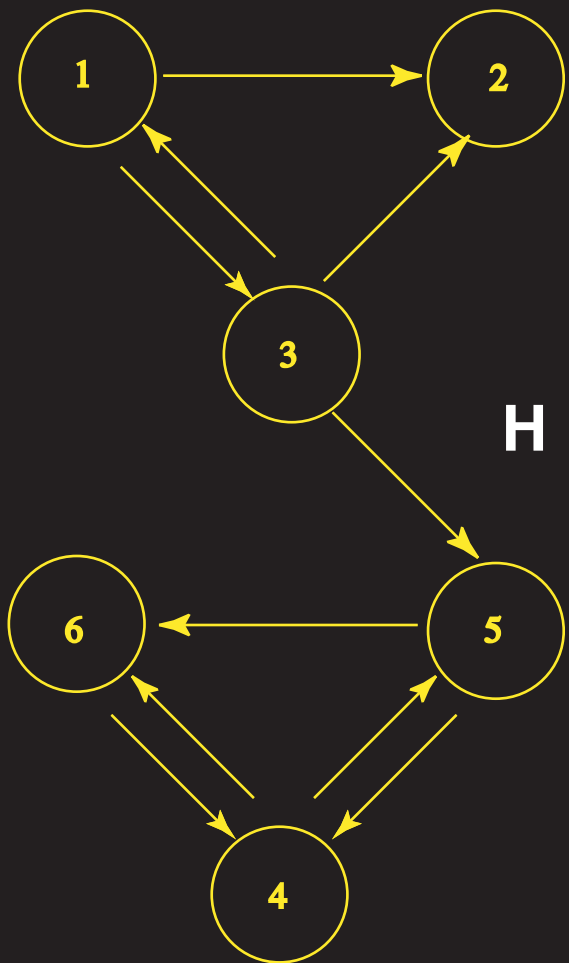
Each  $\pi_j^T$  is a probability vector (  $\pi_i \geq 0$  and  $\sum_i \pi_i = 1$  )

$\pi_{j+1}^T = \pi_j^T \mathbf{P}$  is random walk on the graph defined by links

$\pi^T = \lim_{j \rightarrow \infty} \pi_j^T =$  steady-state probability distribution



# Tiny Web

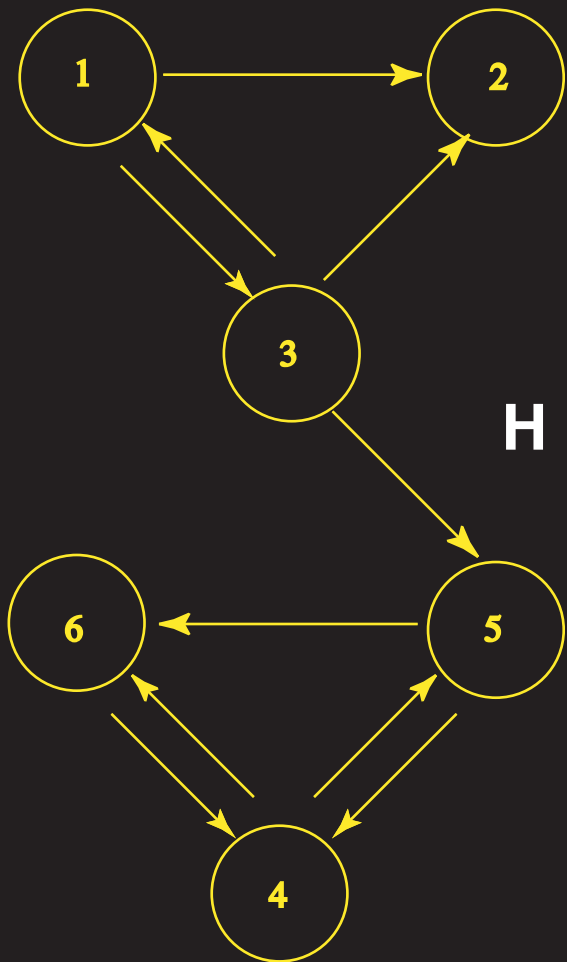


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{pmatrix}$$



# Tiny Web

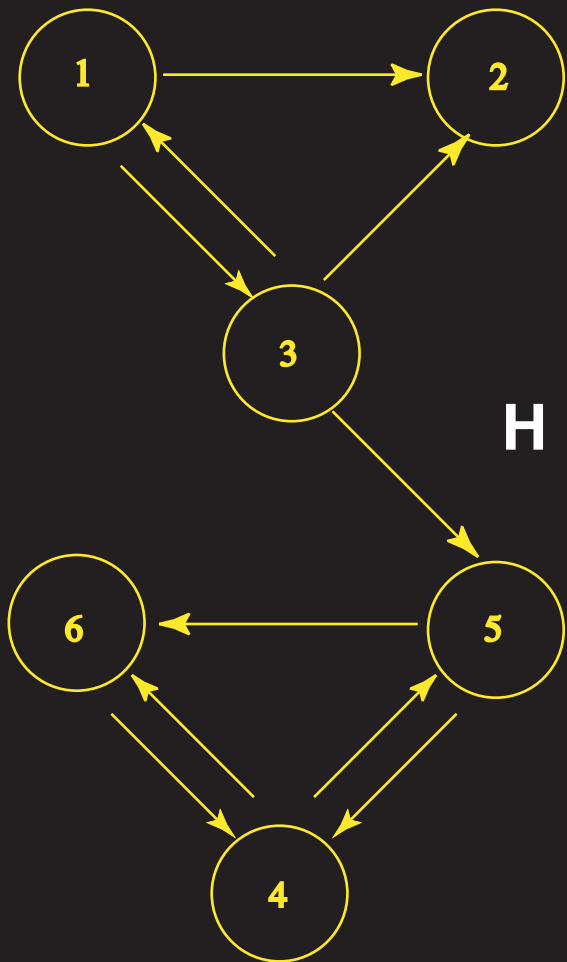


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \end{pmatrix}$$



# Tiny Web

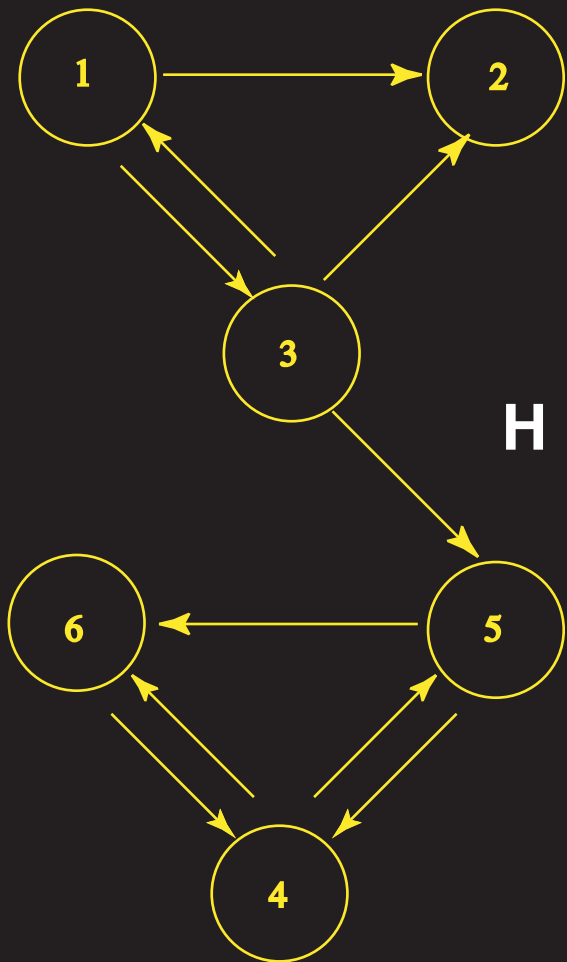


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix}$$



# Tiny Web

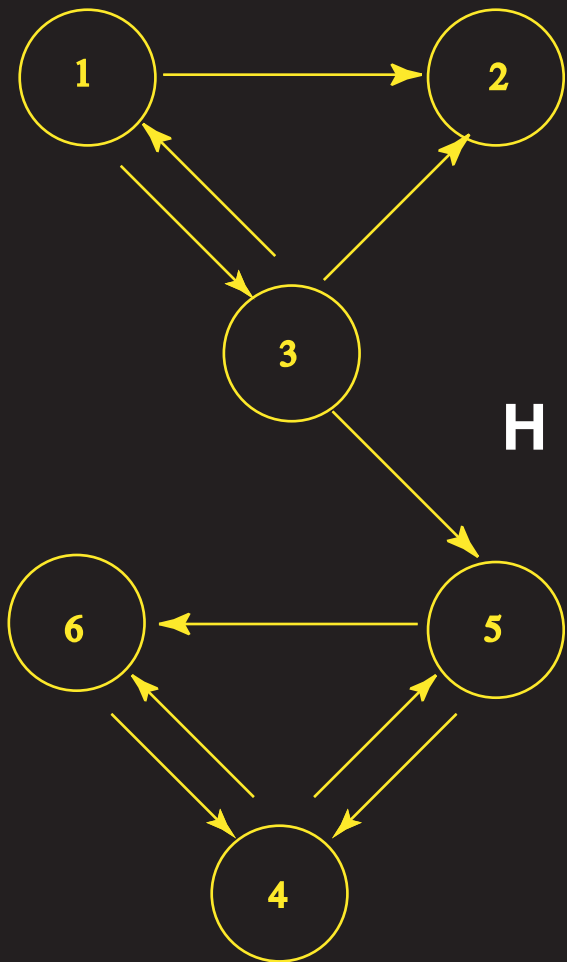


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



# Tiny Web

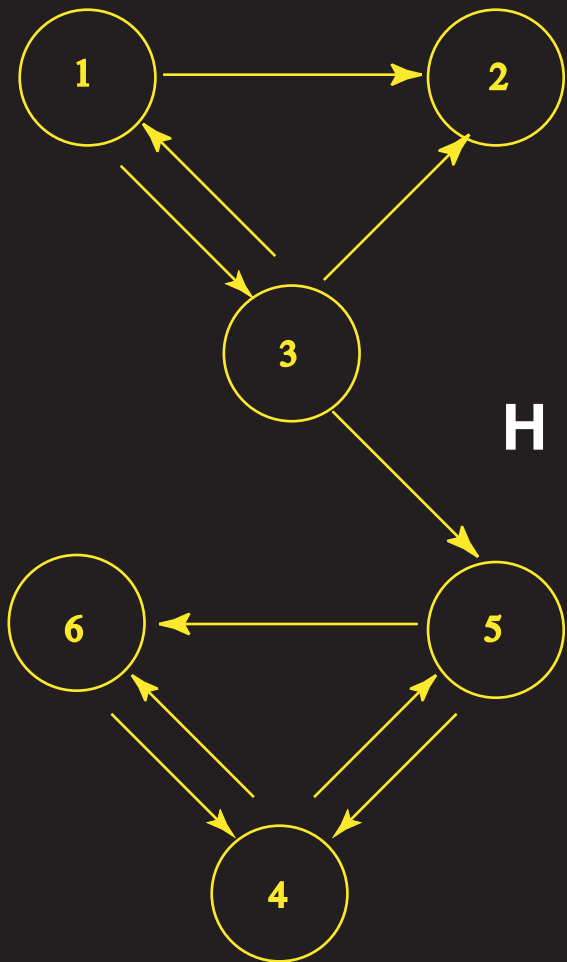


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$



# Tiny Web

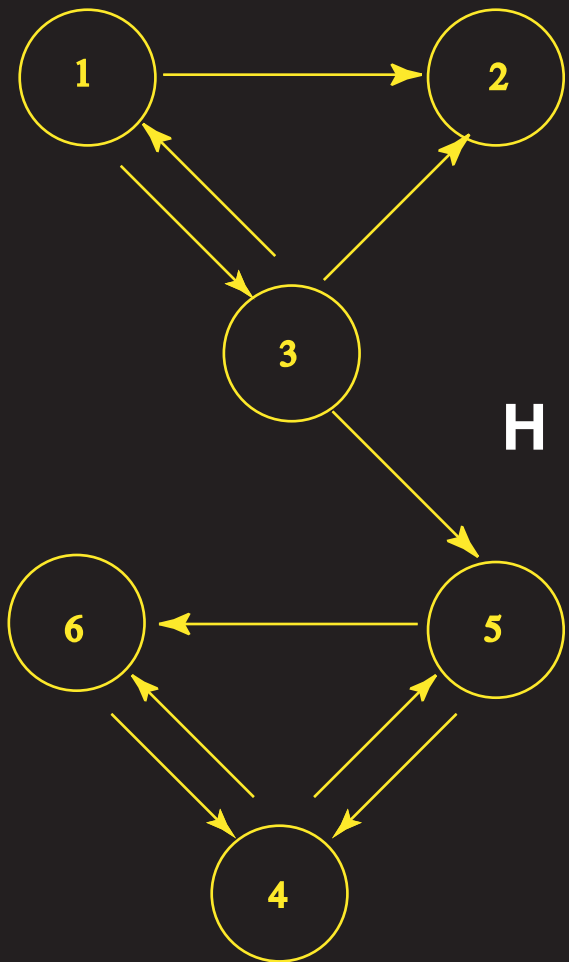


**H** =

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$



# Tiny Web

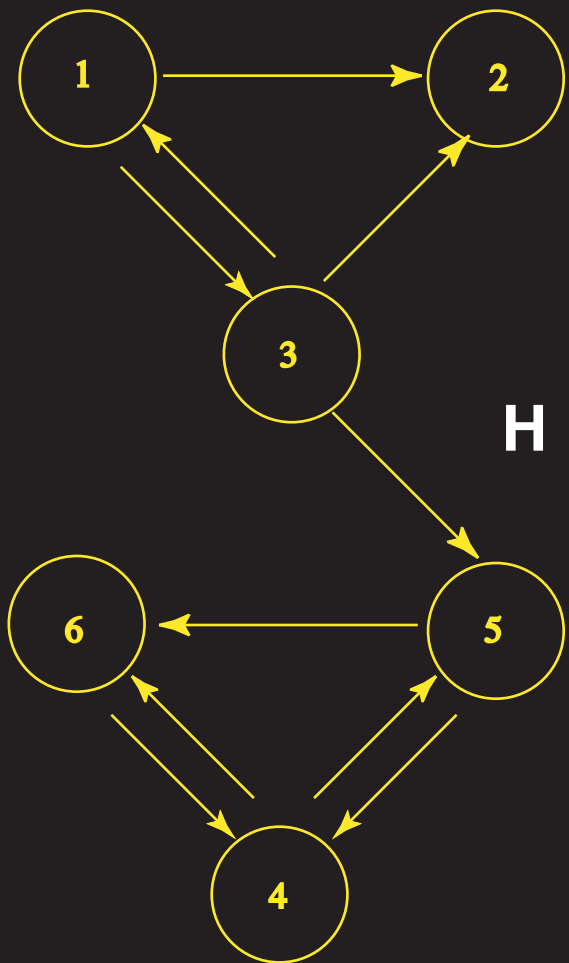


**H** =

$$\begin{matrix} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ P_6 & 0 & 0 & 0 & 1 & 0 & 0 \end{matrix}$$



# Tiny Web



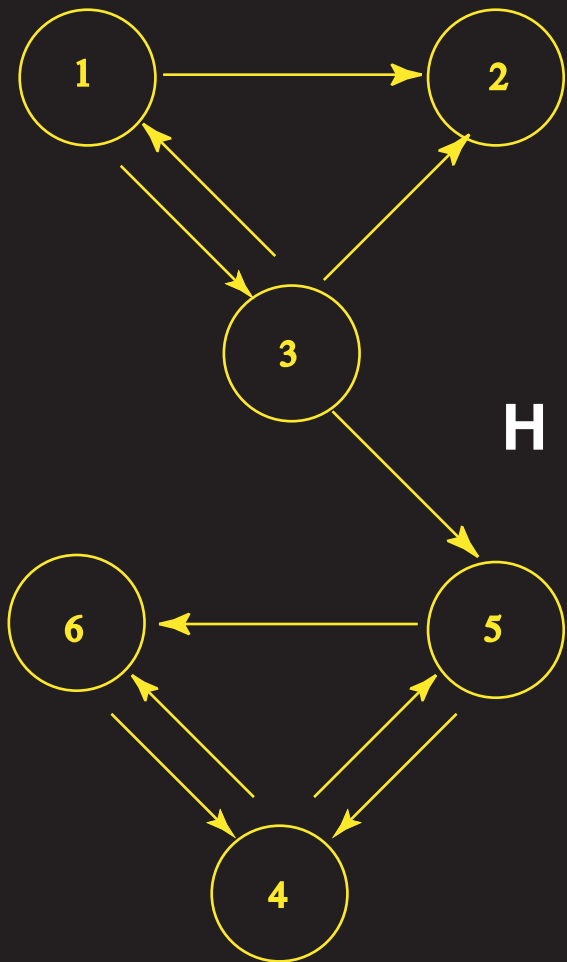
$\mathbf{H} =$

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- Dead end page (nothing to click on) — a “dangling node”



# Tiny Web



$\mathbf{H} =$

$$\begin{matrix} P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{matrix} \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- Dead end page (nothing to click on) — a “dangling node”

✓  $\pi^T$  not well defined



# The Fix

- Replace zero rows with  $(1/n)\mathbf{e}^T = (1/n, 1/n, \dots, 1/n)$

$$\mathbf{S} = \begin{matrix} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \end{pmatrix} \\ P_2 & \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \\ P_3 & \begin{pmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \end{pmatrix} \\ P_4 & \begin{pmatrix} 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \\ P_5 & \begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix} \\ P_6 & \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$



# The Fix

- Replace zero rows with  $(1/n)\mathbf{e}^T = (1/n, 1/n, \dots, 1/n)$

$$\mathbf{S} = \begin{matrix} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \end{pmatrix} \\ P_2 & \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \\ P_3 & \begin{pmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \end{pmatrix} \\ P_4 & \begin{pmatrix} 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix} \\ P_5 & \begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 & 1/2 \end{pmatrix} \\ P_6 & \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

$$\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\mathbf{e}^T}{6} = \mathbf{H} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \frac{1}{6} (1 \ 1 \ 1 \ 1 \ 1 \ 1)$$



# Another Problem

- S is reducible

$$\mathbf{S} = \begin{array}{c|ccc|ccc} & P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ \hline P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ \hline P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ P_6 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

✓  $\pi^T$  not well defined



# Yet More Problems

**Could get trapped into a cycle**  $(P_i \rightarrow P_j \rightarrow P_i)$



# Yet More Problems

**Could get trapped into a cycle**  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence



# Yet More Problems

**Could get trapped into a cycle**  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## **Convergence Requirement**

Markov chain must be irreducible and aperiodic



# Yet More Problems

**Could get trapped into a cycle**  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## **Convergence Requirement**

Markov chain must be irreducible and aperiodic

- This means  $\mathbf{P}$  must be a primitive matrix

✓ No eigenvalues other than  $\lambda = 1$  on unit circle



# Yet More Problems

**Could get trapped into a cycle**  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## Convergence Requirement

Markov chain must be irreducible and aperiodic

- This means  $\mathbf{P}$  must be a primitive matrix
  - ✓ No eigenvalues other than  $\lambda = 1$  on unit circle
  - ✓  $\mathbf{P}^k > 0$  for some  $k$



# Yet More Problems

Could get trapped into a cycle  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## Convergence Requirement

Markov chain must be irreducible and aperiodic

- This means  $\mathbf{P}$  must be a primitive matrix

✓ No eigenvalues other than  $\lambda = 1$  on unit circle

✓  $\mathbf{P}^k > \mathbf{0}$  for some  $k$

## The Google Fixes

- $\mathbf{P} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{e} \mathbf{e}^T / n$        $\alpha \approx .85$



# Yet More Problems

Could get trapped into a cycle  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## Convergence Requirement

Markov chain must be irreducible and aperiodic

- This means  $\mathbf{P}$  must be a primitive matrix

✓ No eigenvalues other than  $\lambda = 1$  on unit circle

✓  $\mathbf{P}^k > 0$  for some  $k$

## The Google Fixes

- $\mathbf{P} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{e} \mathbf{e}^T / n$        $\alpha \approx .85$

- $\mathbf{P} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{e} \mathbf{v}^T$        $\mathbf{v}^T =$  positive probability vector



# Yet More Problems

Could get trapped into a cycle  $(P_i \rightarrow P_j \rightarrow P_i)$

✓  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  won't convergence

## Convergence Requirement

Markov chain must be irreducible and aperiodic

- This means  $\mathbf{P}$  must be a primitive matrix

✓ No eigenvalues other than  $\lambda = 1$  on unit circle

✓  $\mathbf{P}^k > \mathbf{0}$  for some  $k$

## The Google Fixes

- $\mathbf{P} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{e} \mathbf{e}^T / n$        $\alpha \approx .85$

- $\mathbf{P} = \alpha \mathbf{S} + (1 - \alpha) \mathbf{e} \mathbf{v}^T$        $\mathbf{v}^T =$  positive probability vector

- $\mathbf{P} = \alpha \mathbf{H} + (\alpha \mathbf{a} + (1 - \alpha) \mathbf{e}) \mathbf{v}^T$

# THE WALL STREET JOURNAL.

© 2003 Dow Jones & Company. All Rights Reserved

WEDNESDAY, FEBRUARY 26, 2003 - VOL. CCXLI NO. 39 - ★★★ \$1.00

WSJ.com

## What's News—

### Business and Finance

### World-Wide

**NEWS CORP.** and Liberty are no longer working together on a joint offer to take control of Hughes, with News Corp. proceeding on its own and Liberty considering an independent bid. The move threatens to cloud the process of finding a new owner for the GM unit.

(Article on Page A3)

**The SEC signaled** it may file civil charges against Morgan Stanley, alleging it doled out IPO shares based partly on investors' commitments to buy more stock.

(Article on Page C1)

**Ahold's problems deepened** as U.S. authorities opened inquiries into accounting at the Dutch company's U.S. Foodservice unit.

**Fleming said** the SEC upgraded to a formal investigation an inquiry into the food wholesaler's trade practices with suppliers.

(Articles on Page A2)

**Consumer confidence** fell to its lowest level since 1993, hurt by energy costs, the terrorism threat and a stagnant job market.

(Article on Page A3)

**The industrials rebounded** on

**BUSH IS PREPARING** to present Congress a huge bill for Iraq costs.

The total could run to \$95 billion depending on the length of the possible war and occupation. As horse-trading began at the U.N. to win support for a war resolution, the president again made clear he intends to act with or without the world body's imprimatur. Arms inspectors said Baghdad provided new data, including a report of a possible biological bomb. Gen. Franks assumed command of the war-operations center in Qatar. Allied warplanes are aggressively taking out missile sites that could threaten the allied troop buildup. (Column 4 and Pages A4 and A6)

*Turkey's parliament debated legislation to let the U.S. deploy 62,000 to open a northern front. Kurdish soldiers lined roads in a show of force as U.S. officials traveled into Iraq's north for an opposition conference.*

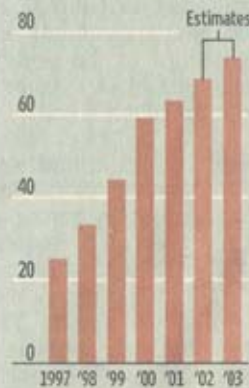
**Powell said** North Korea hasn't restarted a reactor and plutonium-processing facility at Yongbyon, hinting such forbearance might constitute an overture. But saber rattling continued a day after a missile test timed for the inauguration in Seoul. Pyongyang accused U.S. spy planes of violating its airspace and told its army to prepare for U.S. attack. (Page A14)

**The FBI came under** withering bipartisan criticism in a Senate Judiciary report in which Sen. Specter

## Web Master

### As the Web spreads...

Total Internet users, by household, in millions



Sources: Forrester Research; Nielsen NetRatings

### Google's U.S. presence expands

Top search engines, in millions of unique visitors<sup>1</sup>



<sup>1</sup>Including visitors from home and work, in January 2003

Top shopping-referral sites, in millions of referrals<sup>2</sup>



<sup>2</sup>Number of people the sites send to major online stores, including only visitors from home, for Q4 2002

## Bush to Seek up to \$95 Billion To Cover Costs of War on Iraq

By GREG JAFFE  
And JOHN D. MCKINNON

WASHINGTON—The Bush administration is preparing supplemental spending requests totaling as much as \$95 billion for a war with Iraq, its aftermath and new expenses to fight terrorism, officials said.

The total could be as low as \$60 billion because Pentagon budget planners don't know how long a military conflict will last, whether U.S. allies will contribute more than token sums to the effort and what damage Saddam Hussein might do

to his own country to retaliate against conquering forces.

Budget planners also are awaiting the outcome of an intense internal debate over whether to include \$13 billion in the requests to Congress that the Pentagon says it needs to fund the broader war on terrorism, as well as for stepped up homeland security. The White House Office of Management and Budget argues that the money might not be necessary. President Bush, Defense Secretary Donald Rumsfeld and budget director Mitchell Daniels Jr. met yesterday to discuss the matter but didn't reach a final agreement. Mr. Rumsfeld plans to continue pressing his

## Cat and Mouse

### As Google Becomes Web's Gatekeeper, Sites Fight to Get In

Search Engine Punishes Firms That Try to Game System; Outlawing the 'Link Farms'

Exoticleatherwear Gets Cut Off

By MICHAEL TOTT  
And MYLENE MANGALINDAN

Joy Holman sells provocative leather clothing on the Web. She wants what nearly everyone doing business online wants: more exposure on Google.

So from the time she launched exoticleatherwear.com last May, she tried all sorts of tricks to get her site to show up among the first listings when a user of Google Inc.'s popular search engine typed in "women's leatherwear" or "leather apparel." She buried hidden words in her Web pages intended to fool Google's computers. She signed up with a service that promised to have hundreds of sites link to her online store—thereby boosting a crucial measure in Google's system of ranking sites.

The techniques worked—







# Back To Tiny Web

## The Google Matrix

$$\mathbf{P} = \alpha \mathbf{H} + (\alpha \mathbf{a} + (1 - \alpha) \mathbf{e}) \mathbf{v}^T \quad (\text{with } \alpha = .9 \text{ and } \mathbf{v} = \mathbf{e})$$

$$= \begin{bmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{bmatrix}$$

## The PageRank Vector

$$\pi_{j+1}^T = \pi_j^T \mathbf{P} \rightarrow \pi^T$$

$$\pi^T = \begin{pmatrix} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{4} & \mathbf{5} & \mathbf{6} \\ \mathbf{.03721} & \mathbf{.05396} & \mathbf{.04151} & \mathbf{.3751} & \mathbf{.206} & \mathbf{.2862} \end{pmatrix}$$



# Computing $\pi^T$

## A Big Problem

$$\text{Solve } \pi^T = \pi^T \mathbf{P}$$

(eigenvector problem)



# Computing $\pi^T$

## A Big Problem

$$\text{Solve } \pi^T = \pi^T \mathbf{P}$$

(eigenvector problem)

$$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$$

(too big for direct solves)



# Cleve's Corner Page



# Computing $\pi^T$

## A Big Problem

Solve  $\pi^T = \pi^T \mathbf{P}$  (eigenvector problem)

$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$  (too big for direct solves)

Start with  $\pi_0^T = \mathbf{e}/n$  and iterate  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  (power method)



# Computing $\pi^T$

## A Big Problem

Solve  $\pi^T = \pi^T \mathbf{P}$  (eigenvector problem)

$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$  (too big for direct solves)

Start with  $\pi_0^T = \mathbf{e}/n$  and iterate  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  (power method)

## Convergence Time

Measured in days



# Computing $\pi^T$

## A Big Problem

Solve  $\pi^T = \pi^T \mathbf{P}$  (eigenvector problem)

$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$  (too big for direct solves)

Start with  $\pi_0^T = \mathbf{e}/n$  and iterate  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  (power method)

## Convergence Time

Measured in days

## A Bigger Problem — Updating

Pages & links are added, deleted, changed continuously



# Computing $\pi^T$

## A Big Problem

Solve  $\pi^T = \pi^T \mathbf{P}$  (eigenvector problem)

$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$  (too big for direct solves)

Start with  $\pi_0^T = \mathbf{e}/n$  and iterate  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  (power method)

## Convergence Time

Measured in days

## A Bigger Problem — Updating

Pages & links are added, deleted, changed continuously

Google says just start from scratch every 3 to 4 weeks



# Computing $\pi^T$

## A Big Problem

Solve  $\pi^T = \pi^T \mathbf{P}$  (eigenvector problem)

$\pi^T (\mathbf{I} - \mathbf{P}) = \mathbf{0}$  (too big for direct solves)

Start with  $\pi_0^T = \mathbf{e}/n$  and iterate  $\pi_{j+1}^T = \pi_j^T \mathbf{P}$  (power method)

## Convergence Time

Measured in days

## A Bigger Problem — Updating

Pages & links are added, deleted, changed continuously

Google says just start from scratch every 3 to 4 weeks

Prior results don't help to restart



# Conclusions

✦ Elegant Blend of NA, LA, Graph Theory, MC, & CS ✦



# Conclusions

- ✦ Elegant Blend of NA, LA, Graph Theory, MC, & CS ✦
- ✦ Google Now Uses Many Other “Metrics” to augment PR ✦



# Conclusions

- ✦ Elegant Blend of NA, LA, Graph Theory, MC, & CS ✦
- ✦ Google Now Uses Many Other “Metrics” to augment PR ✦
- ✦ Search Is Opening New Areas Ripe For Inovative Ideas ✦
- ✦ Exciting Work That Is Changing The World ✦



# Conclusions

- ✦ Elegant Blend of NA, LA, Graph Theory, MC, & CS ✦
- ✦ Google Now Uses Many Other “Metrics” to augment PR ✦
- ✦ Search Is Opening New Areas Ripe For Inovative Ideas ✦
- ✦ Exciting Work That Is Changing The World ✦

**✦ Thanks For Your Attention ✦**