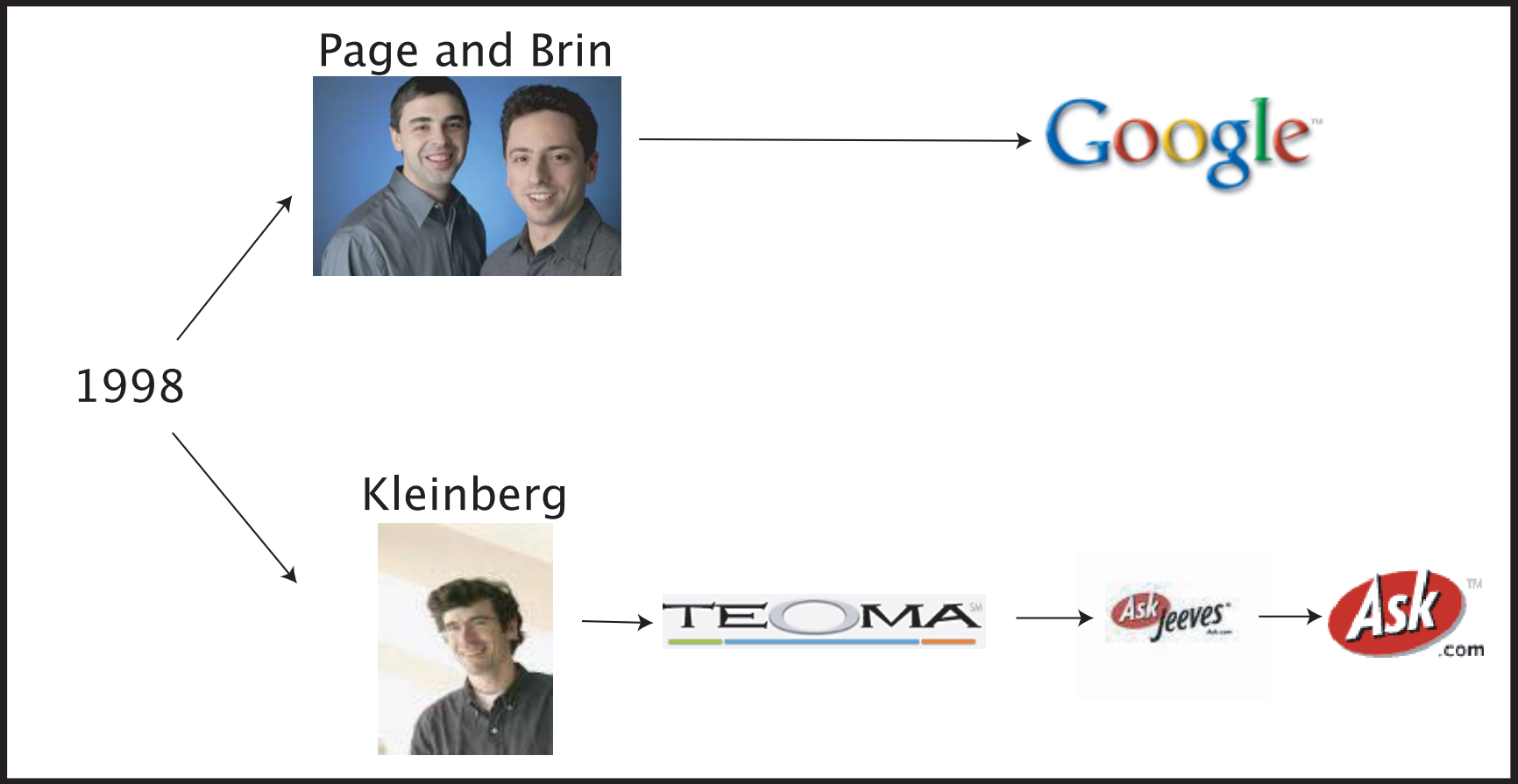
Google's PageRank and Beyond

THE SCIENCE OF SEARCH ENGINE RANKINGS

AMY N. LANGVILLE and CARL D. MEYER

1998: enter Link Analysis

- uses *hyperlink* structure to focus the relevant set
- combine traditional IR score with popularity score



Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = web IR

How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, average page size of 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!

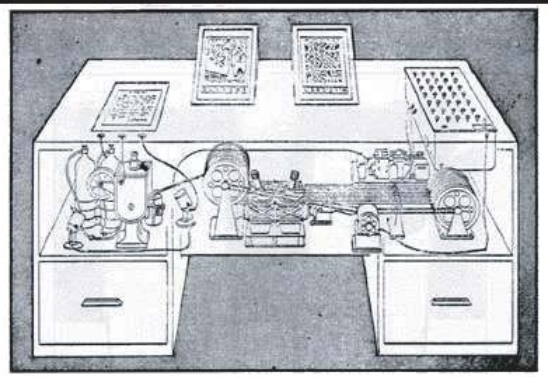
A Herculean Task!

Web Information Retrieval

IR before the Web = traditional IR
IR on the Web = **web IR**

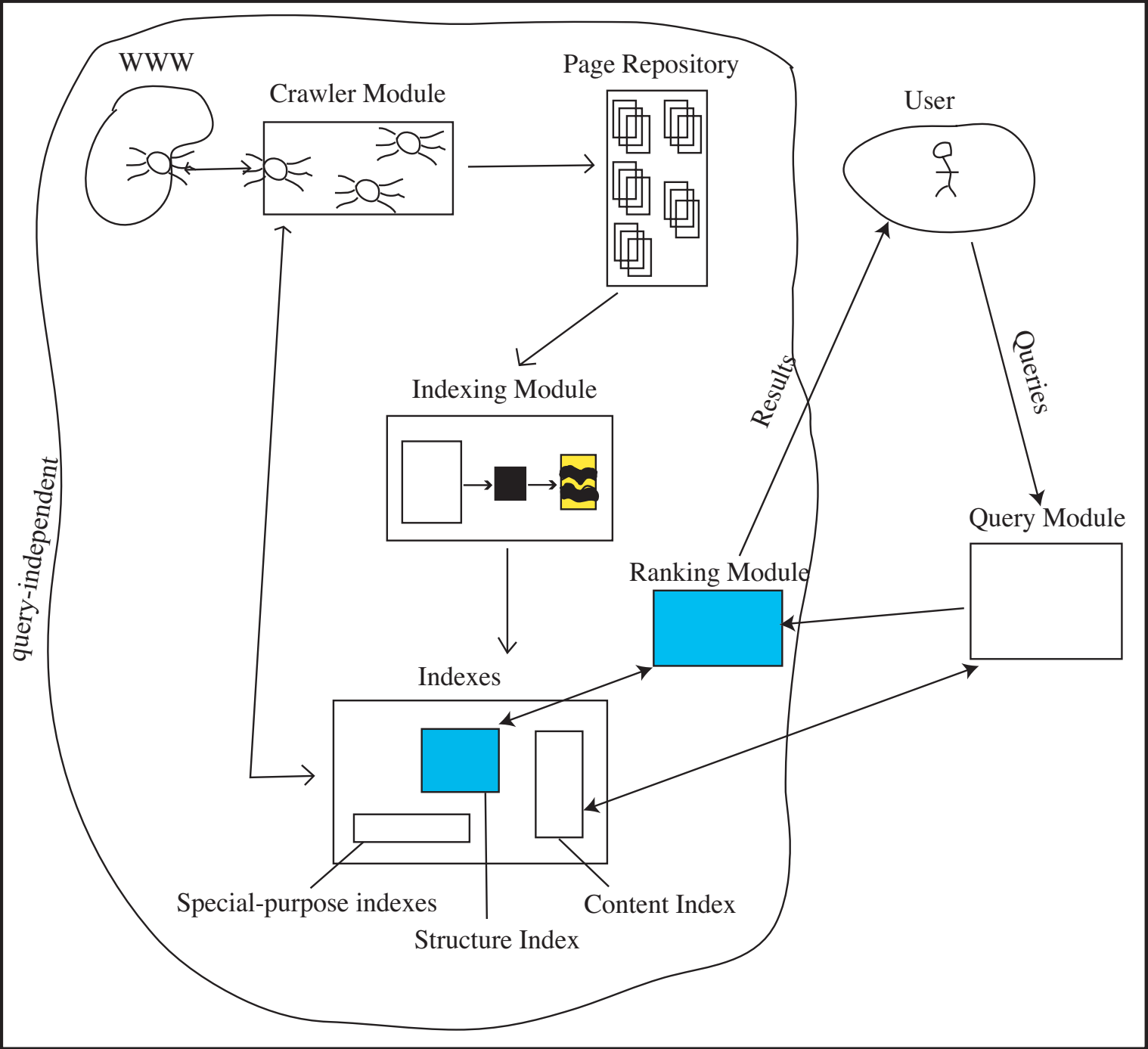
How is the Web different from other document collections?

- It's huge.
  - over 10 billion pages, each about 500KB
  - 20 times size of Library of Congress print collection
  - Deep Web - 400 X bigger than Surface Web
- It's dynamic.
  - content changes: 40% of pages change in a week, 23% of .com change daily
  - size changes: billions of pages added each year
- It's self-organized.
  - no standards, review process, formats
  - errors, falsehoods, link rot, and spammers!
- Ah, but it's *hyperlinked* !
  - Vannevar Bush's 1945 memex



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Elements of a Web Search Engine



The Ranking Module (generates popularity scores)

- Measure the importance of each page

The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations

The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed

The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
- Google's PageRank© technology distinguishes it from all competitors

The Ranking Module (generates popularity scores)

- Measure the importance of each page
- The measure should be Independent of any query
  - Primarily determined by the link structure of the Web
  - Tempered by some content considerations
- Compute these measures off-line long before any queries are processed
- Google's PageRank© technology distinguishes it from all competitors

Google's PageRank = Google's $$$$$

Sign in

Google

Web Images Video News Maps more »

business intelligence

Search

Advanced Search
Preferences

Web Results **1** - **10** of about **122,000,000** for **business intelligence**. (**0.10** seconds)

SAS Business Intelligence Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (**BI**) is a **business** management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/**Business_intelligence** - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for **Business Intelligence**
The **Business Intelligence** resource for **business** and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.**businessintelligence**.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ...
Business intelligence and **business** performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle **Business Intelligence** Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ **business_intelligence**/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.**business-intelligence**.co.uk/ - 18k - Cached - Similar pages

Intelligent Enterprise: Better Insight for **Business** Decisions

Sponsored Links

SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com

Business Intelligence
See what **business intelligence** can do for you (free interactive demo).
www.InformationBuilders.com

MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp **Business Intelligence** All Inclusive
www.mcseclasses.com

Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com

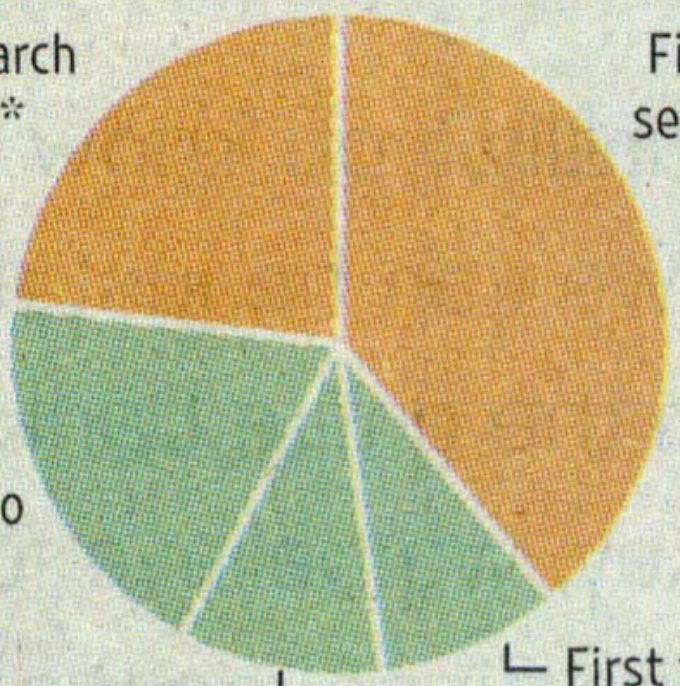
Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com

Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

The Wall Street Journal, April 13, 2007

Take Your Pick

Amount of Internet search results that Web surfers typically scan before selecting one



*Top results without reading through the whole page

Note: Sample size is 2,369 people

Sources: JupiterResearch; iProspect

Sign in

Web Images Video News Maps more »

Google

business intelligence

Search

Advanced Search
Preferences

Web

Results **1** - **10** of about **122,000,000** for **business intelligence**. (**0.10** seconds)

SAS **Business Intelligence** Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (**BI**) is a **business** management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/**Business_intelligence** - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for **Business Intelligence**
The **Business Intelligence** resource for **business** and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.**businessintelligence**.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ...
Business intelligence and **business** performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle **Business Intelligence** Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ **business_intelligence**/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.**business-intelligence**.co.uk/ - 18k - Cached - Similar pages

Intelligent Enterprise: Better Insight for Business Decisions

Sponsored Links

SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com

Business Intelligence
See what **business intelligence** can do for you (free interactive demo).
www.InformationBuilders.com

MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp **Business Intelligence** All Inclusive
www.mcseclasses.com

Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com

Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com

Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

Sign in

Google

Web Images Video News Maps more »

business intelligence

Search

Advanced Search
Preferences

Web

Results **1** - **10** of about **122,000,000** for **business intelligence**. (**0.10** seconds)

SAS **Business Intelligence** Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (**BI**) is a **business** management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/**Business_intelligence** - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for **Business Intelligence**
The **Business Intelligence** resource for **business** and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.**businessintelligence**.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ...
Business intelligence and **business** performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle **Business Intelligence** Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ **business_intelligence**/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.**business-intelligence**.co.uk/ - 18k - Cached - Similar pages

Intelligent Enterprise: Better Insight for **Business** Decisions

Sponsored Links

SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com

Business Intelligence
See what **business intelligence** can do for you (free interactive demo).
www.InformationBuilders.com

MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp **Business Intelligence** All Inclusive
www.mcseclasses.com

Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com

Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com

Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com


Newsweek
March 29
$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
PLUS: The Future of Digital Voting
Google founders Larry Page and Sergey Brin

Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - ▷ Off-line calculations — No query time computation
- Let the Web vote with in-links
  - ▷ But not by simple link counts
    - — One link to P from Yahoo! is important
    - — Many links to P from me is not
- Share The Vote
  - ▷ Yahoo! casts many "votes"
    - — value of vote from $Yahoo$! is diluted
  - ▷ If Yahoo! "votes" for n pages
    - — Then P receives only $r(Y)/n$ credit from Y

Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - ▷ Off-line calculations — No query time computation
- Let the Web vote with in-links
  - ▷ But not by simple link counts
    - — One link to P from Yahoo! is important
    - — Many links to P from me is not
- Share The Vote
  - ▷ Yahoo! casts many "votes"
    - — value of vote from $Yahoo$! is diluted
  - ▷ If Yahoo! "votes" for n pages
    - — Then P receives only $r(Y)/n$ credit from Y

Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - ▷ Off-line calculations — No query time computation
- Let the Web vote with in-links
  - ▷ But not by simple link counts
    - — One link to P from Yahoo! is important
    - — Many links to P from me is not
- Share The Vote
  - ▷ Yahoo! casts many "votes"
    - — value of vote from $Yahoo$! is diluted
  - ▷ If Yahoo! "votes" for n pages
    - — Then P receives only $r(Y)/n$ credit from Y

Google's PageRank

(Lawrence Page & Sergey Brin 1998)

The Google Goals

- Create a PageRank $r(P)$ that is not query dependent
  - ▷ Off-line calculations — No query time computation
- Let the Web vote with in-links
  - ▷ But not by simple link counts
    - — One link to P from Yahoo! is important
    - — Many links to P from me is not
- Share The Vote
  - ▷ Yahoo! casts many "votes"
    - — value of vote from $Yahoo$! is diluted
  - ▷ If Yahoo! "votes" for n pages
    - — Then P receives only $r(Y)/n$ credit from Y

PageRank

The Definition

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

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

$|P|$ = number of out links from P

PageRank

The Definition

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

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

$|P|$ = number of out links from P

Successive Refinement

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

PageRank

The Definition

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

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

$|P|$ = number of out links from P

Successive Refinement

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

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

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

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

$|P|$ = number of out links from P

Successive Refinement

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

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

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

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

$|P|$ = number of out links from P

Successive Refinement

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

Iteratively refine rankings for each page

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

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

$$\ddots$$

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

In Matrix Notation

After Step k

- $\boldsymbol{\pi}_k^T = [r_k(P_1),\, r_k(P_2),\, \cdots,\, r_k(P_n)]$

In Matrix Notation

After Step k

- $\boldsymbol{\pi}_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

- $\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T \mathbf{H}$ where $h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$

In Matrix Notation

After Step k

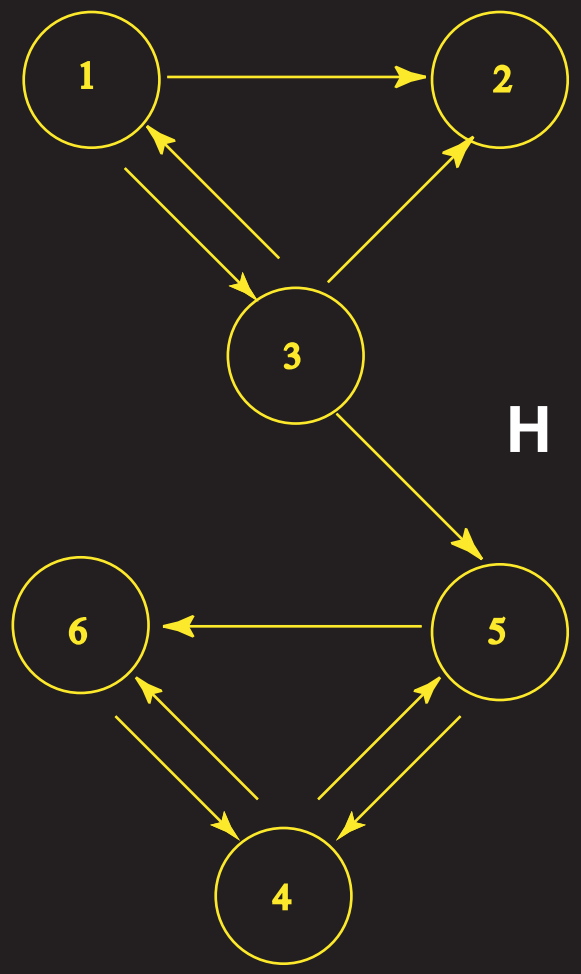
- $\boldsymbol{\pi}_k^T = [r_k(P_1), r_k(P_2), \cdots, r_k(P_n)]$

- $\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T \mathbf{H}$ where $h_{ij} = \begin{cases} 1/|P_i| & \text{if } i \to j \\ 0 & \text{otherwise} \end{cases}$

- PageRank vector = $\boldsymbol{\pi}^T = \lim_{k \to \infty} \boldsymbol{\pi}_k^T$ = eigenvector for $\mathbf{H}$

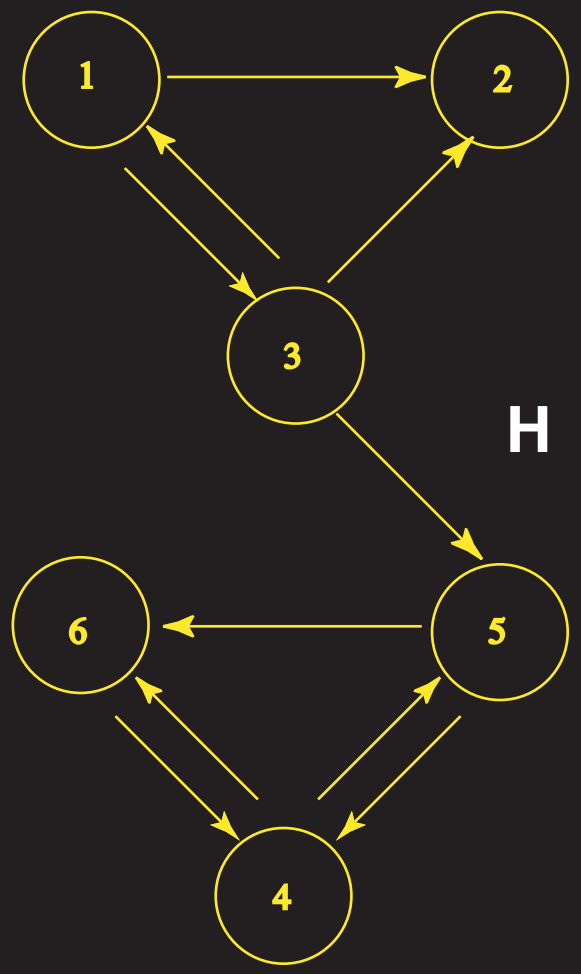
Provided that the limit exists

Tiny Web



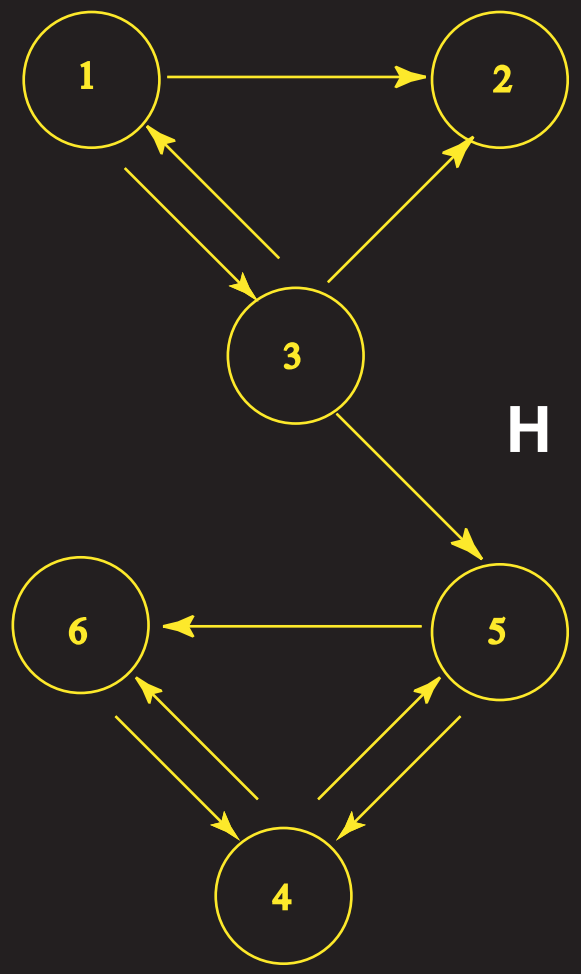
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



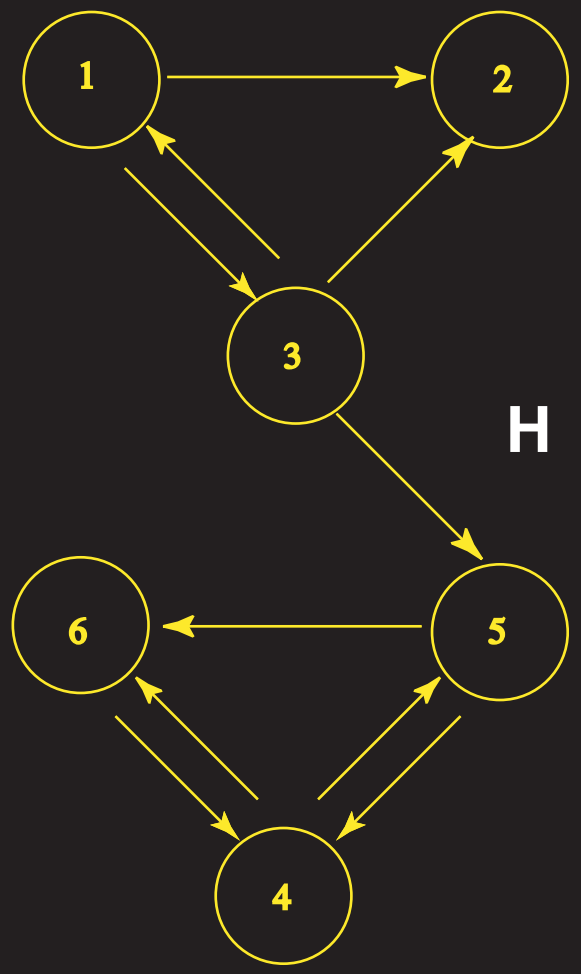
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



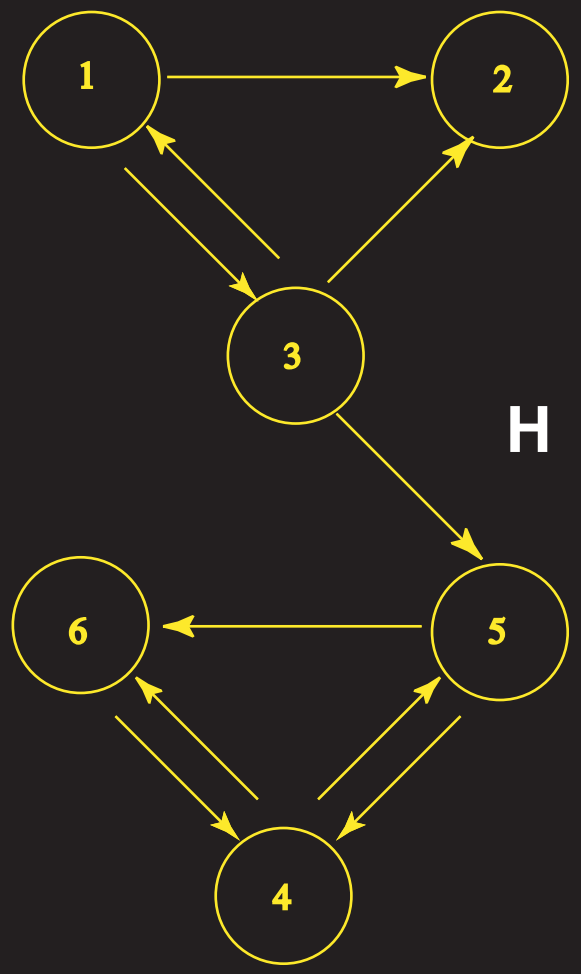
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



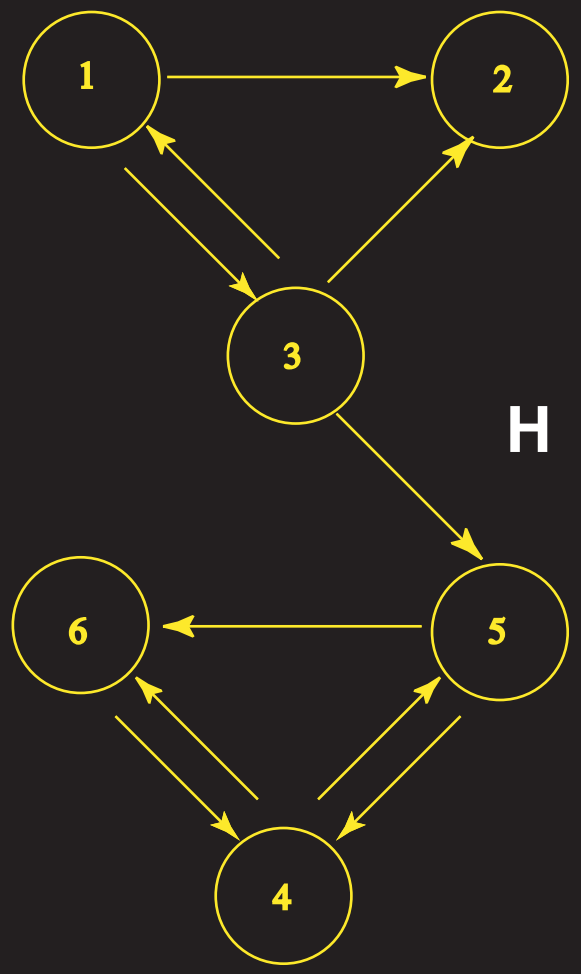
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



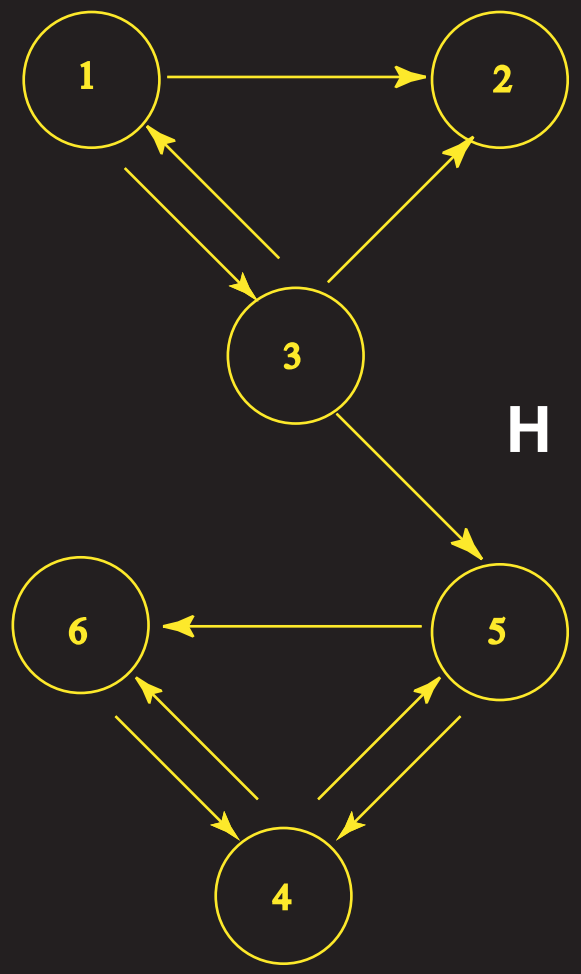
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



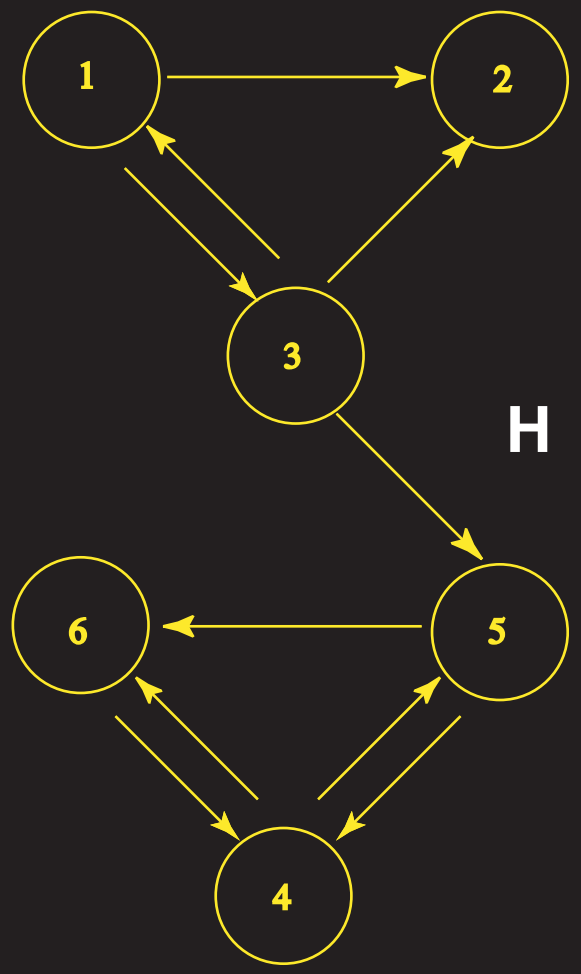
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

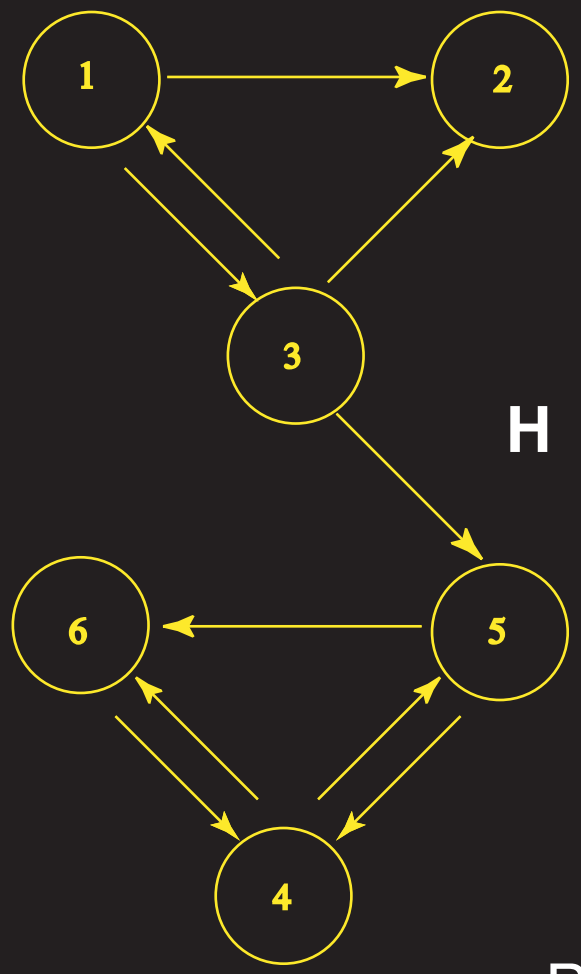
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph

Tiny Web



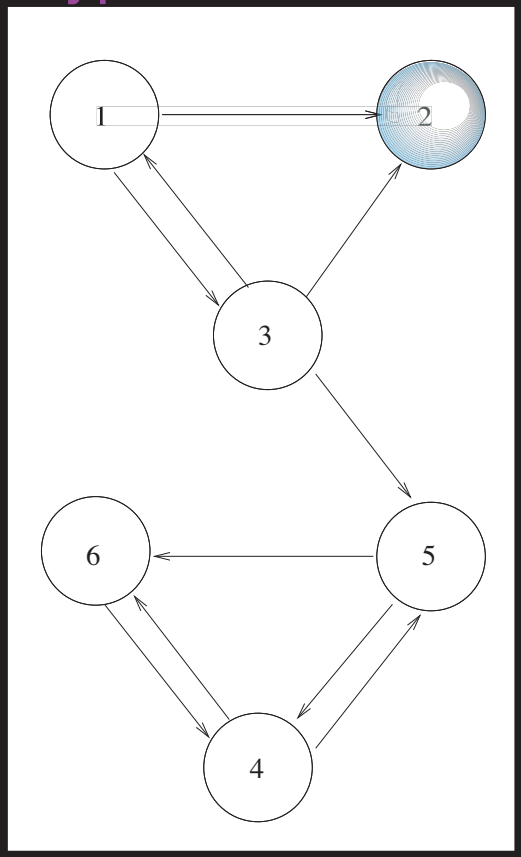
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote

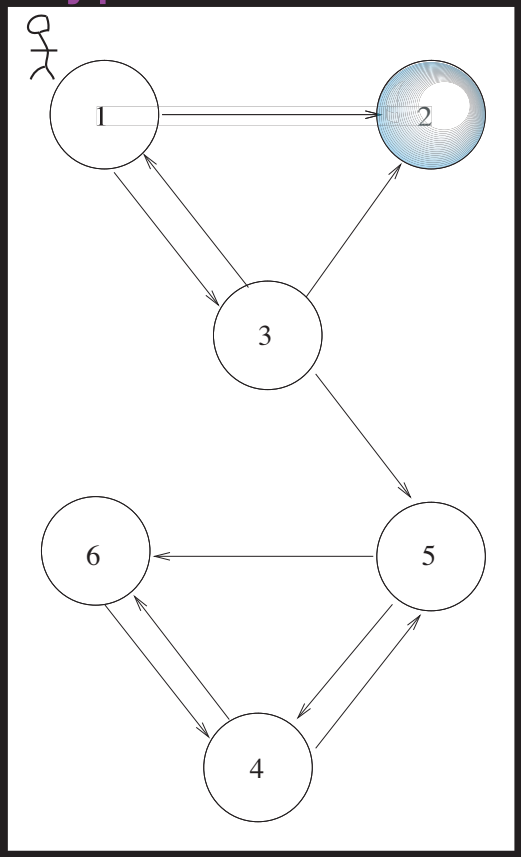


Markov chain

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

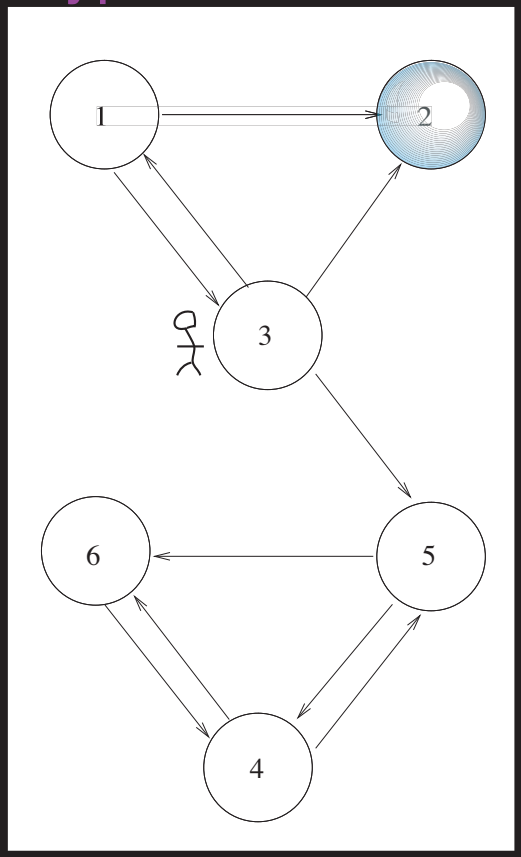
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

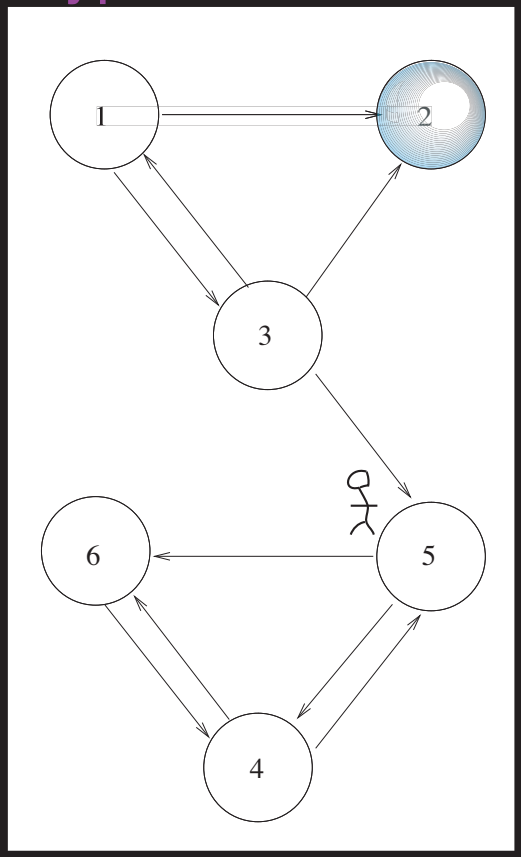
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

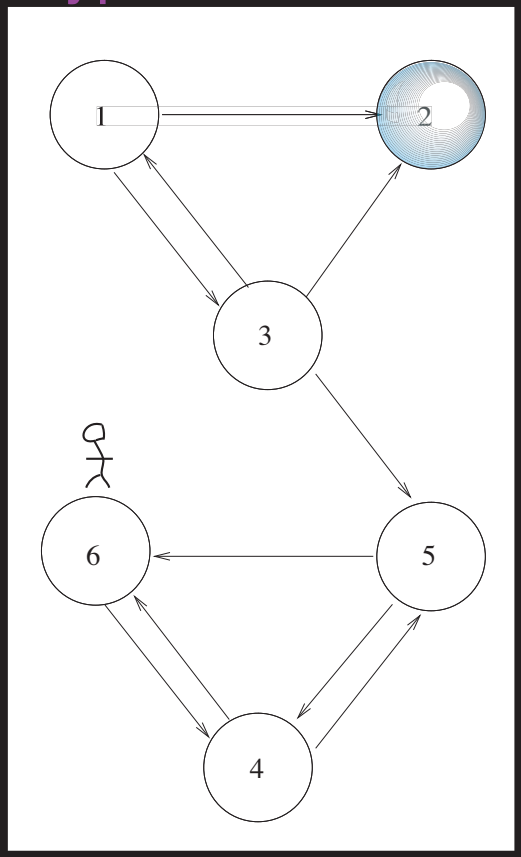
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

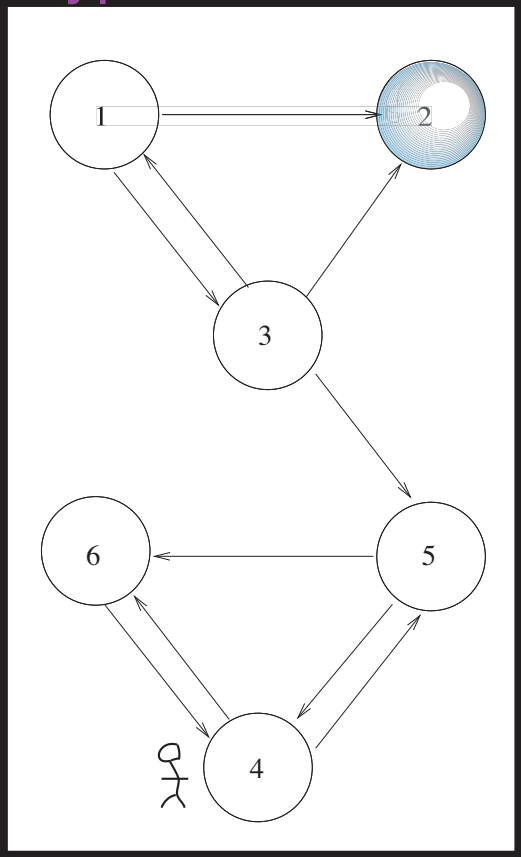
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

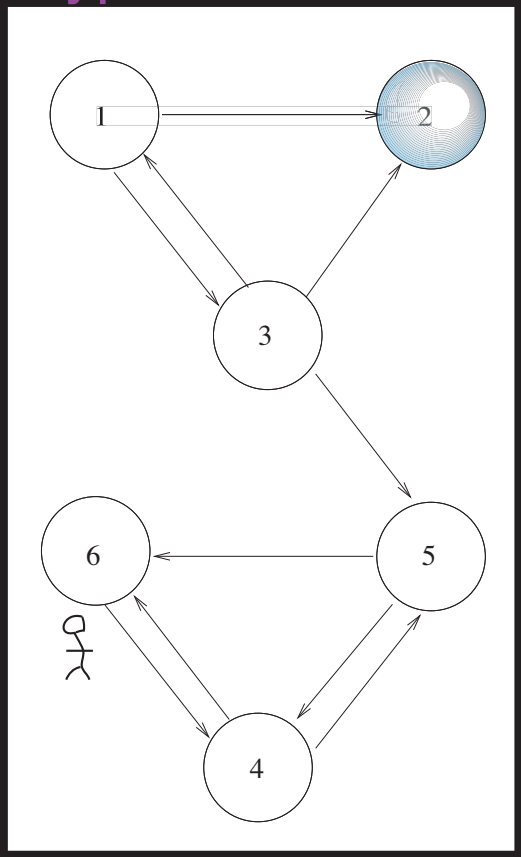
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

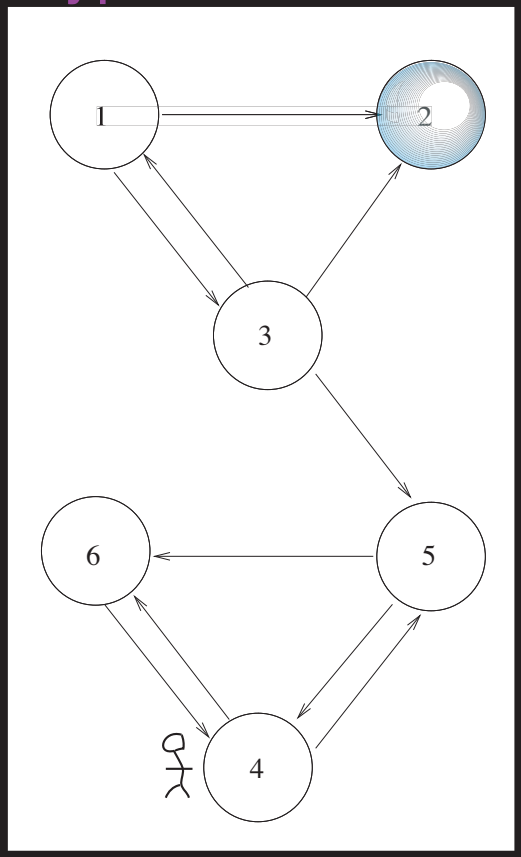
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

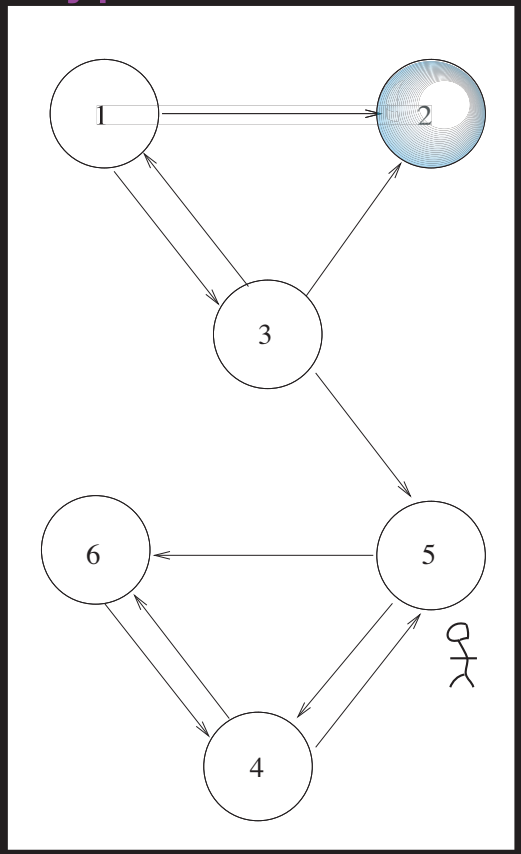
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

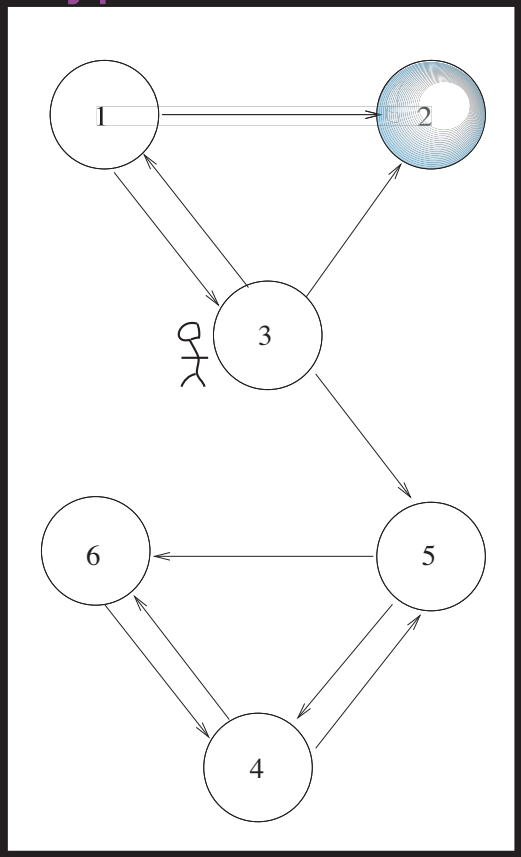
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

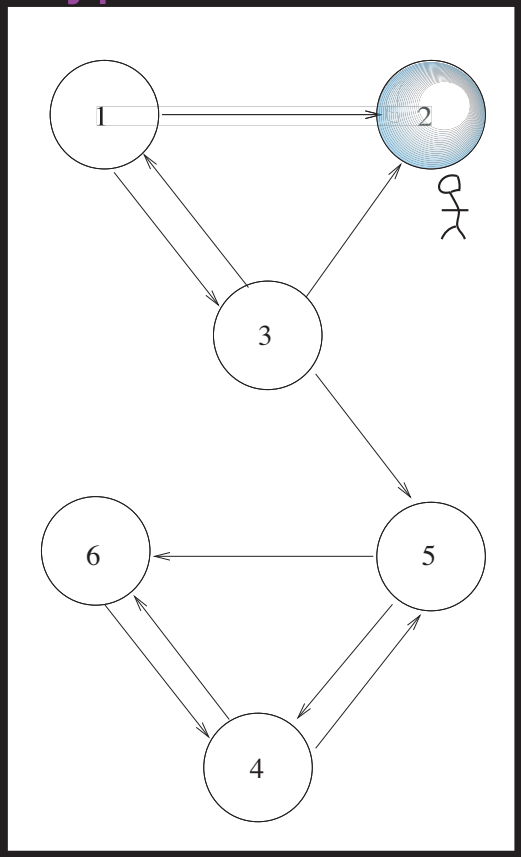
Hyperlink as vote



Ranking with a Random Surfer

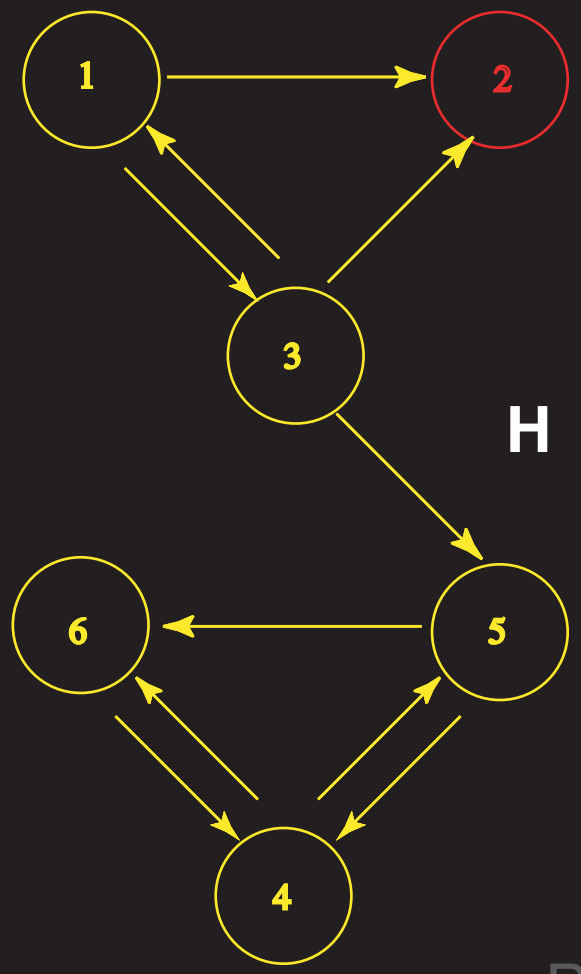
- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



page 2 is a dangling node

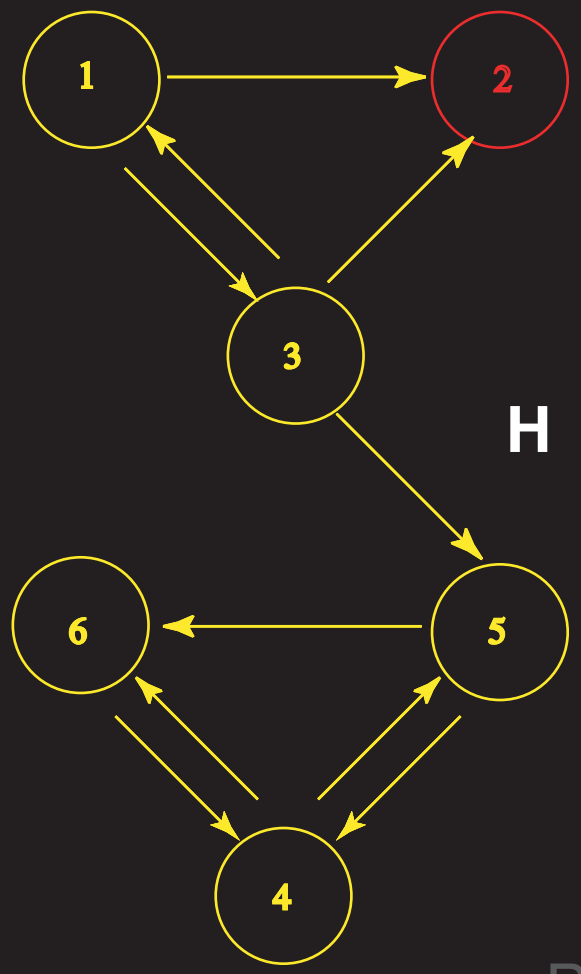
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i
- Dead end page (nothing to click on) — a “dangling node”

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- ▷ A random walk on the Web Graph
- ▷ PageRank = π_i = amount of time spent at P_i
- ▷ Dead end page (nothing to click on) — a "dangling node"
- ▷ $\boldsymbol{\pi}^T = (0, 1, 0, 0, 0, 0)$ = e-vector $\Longrightarrow$ Page P_2 is a "rank sink"

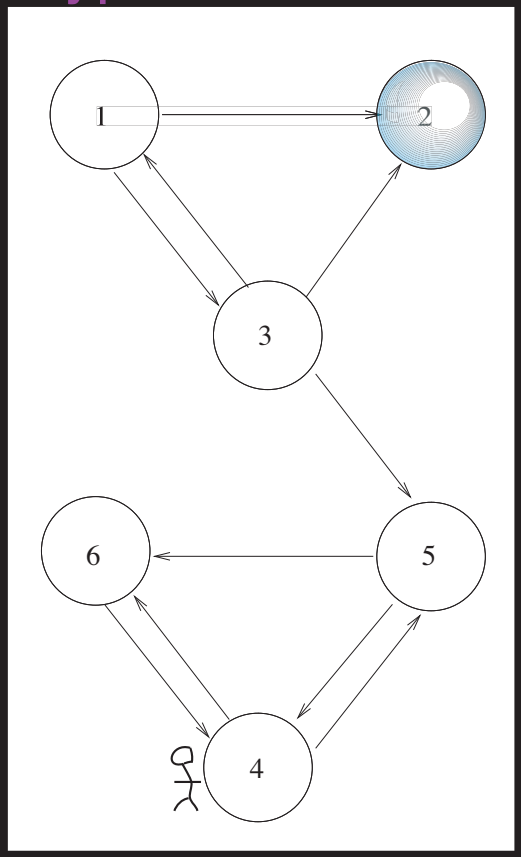
The Fix

Allow Web Surfers To Make Random Jumps

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



surfer “teleports”

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

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

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

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

- $\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\,\mathbf{e}^T}{6}$ is now row stochastic $\implies \rho(\mathbf{S}) = 1$

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

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

- $\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\,\mathbf{e}^T}{6}$ is now row stochastic $\Longrightarrow$ $\rho(\mathbf{S}) = 1$
- Perron says $\exists$ $\boldsymbol{\pi}^T \geq 0$ s.t. $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{S}$ with $\sum_i \pi_i = 1$

Nasty Problem

The Web Is Not Strongly Connected

Nasty Problem

The Web Is Not Strongly Connected

- S is reducible

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

Nasty Problem

The Web Is Not Strongly Connected

∴ S is reducible

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

- Reducible $\Longrightarrow$ PageRank vector is not well defined
- Frobenius says $\mathbf{S}$ needs to be *irreducible* to ensure a unique $\pi^T > 0$ s.t. $\pi^T = \pi^T \mathbf{S}$ with $\sum_i \pi_i = 1$

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \rightarrow P_j \rightarrow P_i)$

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \rightarrow P_j \rightarrow P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\boldsymbol{\pi}_{k+1}^T = \boldsymbol{\pi}_k^T \mathbf{S}$ fails to convergence

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive
- No eigenvalues other than $\lambda = 1$ on unit circle

Irreducibility Is Not Enough

Could Get Trapped Into A Cycle $(P_i \to P_j \to P_i)$

- The powers $\mathbf{S}^k$ fail to converge
- $\pi_{k+1}^T = \pi_k^T \mathbf{S}$ fails to convergence

Convergence Requirement

- Perron–Frobenius requires $\mathbf{S}$ to be primitive
- No eigenvalues other than $\lambda = 1$ on unit circle
- Frobenius proved $\mathbf{S}$ is primitive $\iff \mathbf{S}^k > 0$ for some k

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector $\qquad \pi^T$ = left-hand Perron vector of $\mathbf{G}$

Ranking with a Random Surfer

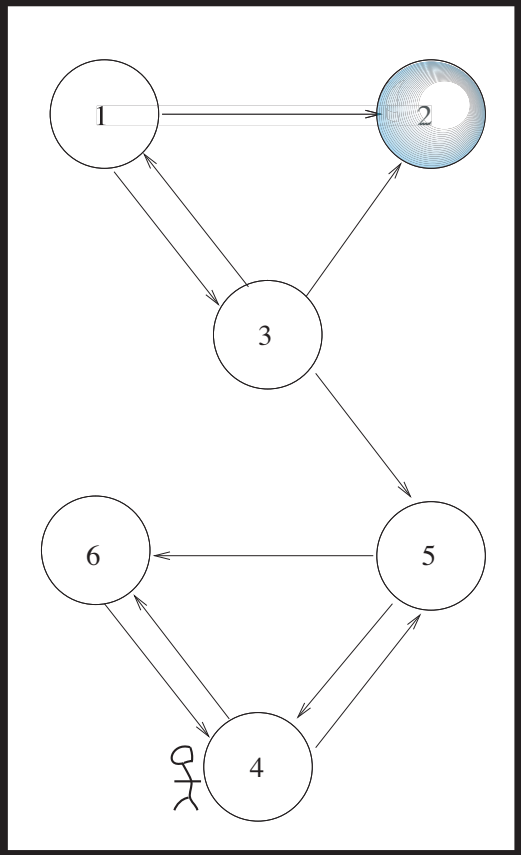
- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or $proportion\ of\ time$, the surfer spends on each page to create ranking of webpages.

Ranking with a Random Surfer

- If a page is "important," it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or *proportion of time*, the surfer spends on each page to create ranking of webpages.

Proportion of Time

Page 1 = .04
Page 2 = .05
Page 3 = .04
Page 4 = .38
Page 5 = .20
Page 6 = .29



Ranked List of Pages

Page 4
Page 6
Page 5
Page 2
Page 1
Page 3

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector $\qquad \pi^T$ = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ Sparse computations with the original link structure

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector $\qquad \pi^T$ = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ Convergence rate controllable by Google engineers

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector $\qquad \pi^T$ = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ Convergence rate controllable by Google engineers
- $\mathbf{v}^T$ can be any positive probability vector in $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T$

The Google Fix

Allow A Random Jump From Any Page

- $\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} > 0, \quad \mathbf{E} = \mathbf{e}\mathbf{e}^T/n, \quad 0 < \alpha < 1$
- $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T > 0 \qquad \mathbf{u} = \alpha\mathbf{a} + (1-\alpha)\mathbf{e}, \quad \mathbf{v}^T = \mathbf{e}^T/n$
- PageRank vector $\qquad \pi^T$ = left-hand Perron vector of $\mathbf{G}$

Some Happy Accidents

- $\mathbf{x}^T\mathbf{G} = \alpha\mathbf{x}^T\mathbf{H} + \beta\mathbf{v}^T$ Sparse computations with the original link structure
- $\lambda_2(\mathbf{G}) = \alpha$ Convergence rate controllable by Google engineers
- $\mathbf{v}^T$ can be any positive probability vector in $\mathbf{G} = \alpha\mathbf{H} + \mathbf{u}\mathbf{v}^T$
- The choice of $\mathbf{v}^T$ allows for personalization

PageRank Issues

- Computing PageRank: simulation, eigensystem, linear system; accuracy
- power law distribution: sensitivity, spamming
- link strategies
- overuse