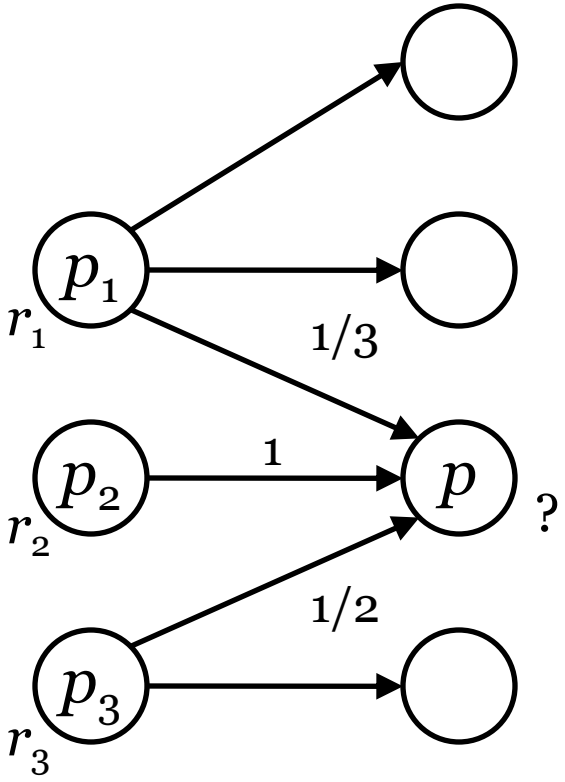
Having Fun with PageRank and MapReduce

Hadoop User Group (HUG) UK – 14th April 2009 – http://huguk.org/

Paolo Castagna – HP Labs, Bristol, UK

PageRank



$$r_p = \frac{r_1}{3} + r_2 + \frac{r_3}{2}$$

PageRank

$$r_{p_i} = \sum_{p_j \in B_{p_i}} \frac{r_{p_j}}{|p_j|}$$

recursive definition

B_{p_i} *backward links* (i.e. links to p_i)

$|p_j|$ number of *forward links* (i.e. links from p_j)

PageRank

iterative computation

$$r_{k+1}(p_i) = \sum_{p_j \in B_{p_i}} \frac{r_k(p_j)}{|p_j|}$$

$$r_0(p_i) = \frac{1}{N}$$

$r_k(p_i)$	pagerank of page p_i at k iteration
N	total number of pages

Random Surfer

- A surfer follows links at random indefinitely
- Time spent on a given page measure the importance of that page
- Problems:
  - rank sinks (accumulate too much)
  - cycles (could cause periodicity)
- Dangling pages? Jump to any other page
- Bored? Teleportation (fixes rank sinks and eliminates cycles)

Dangling Pages

if $|p_j|$ is zero?

$$r_{k+1}(p_i) = \sum_{\substack{p_j \in B_{p_i} \\ |p_j| \neq 0}} \frac{r_k(p_j)}{|p_j|} + \underbrace{\sum_{\substack{p_j \\ |p_j| = 0}} \frac{r_k(p_j)}{N}}_{\text{random jump}}$$

independent from p_i

N total number of pages

Teleportation

if there are loops or someone gets bored?

$$r_{k+1}(p_i) = d\left(\sum_{\substack{p_j \in B_{p_i} \\ |p_j| \neq 0}} \frac{r_k(p_j)}{|p_j|} + \sum_{\substack{p_j \\ |p_j| = 0}} \frac{r_k(p_j)}{N} \right) + \underbrace{(1-d)\sum_{p_j} \frac{r_k(p_j)}{N}}_{\text{random jump}}$$

independent from p_i

$$\sum_{p_j} r_k(p_j) = 1$$

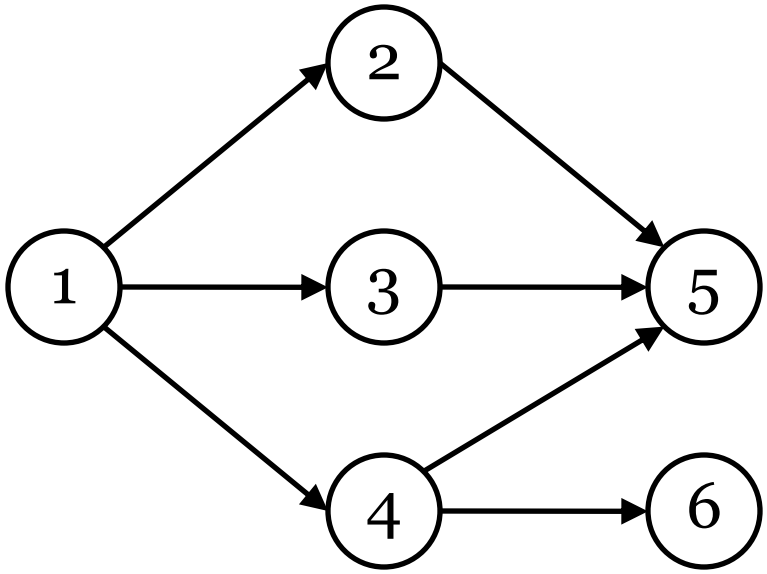
d=0.85 *dumping factor*

PageRank

$$r_{k+1}(p_i) = d\left(\sum_{\substack{p_j \in B_{p_i} \\ |p_j| \neq 0}} \frac{r_k(p_j)}{|p_j|} + \sum_{\substack{p_j \\ |p_j| = 0}} \frac{r_k(p_j)}{N} \right) + \frac{(1-d)}{N}$$

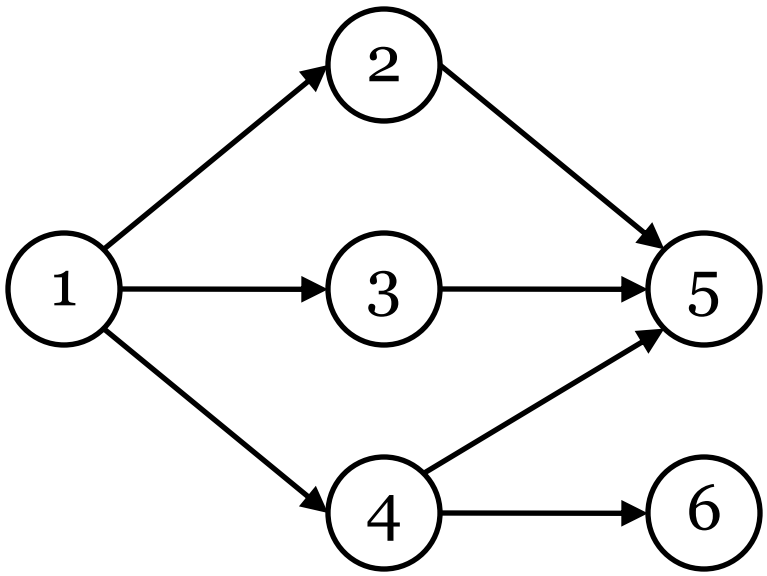
$$r_0(p_i) = \frac{1}{N}$$

Adjacency Matrix



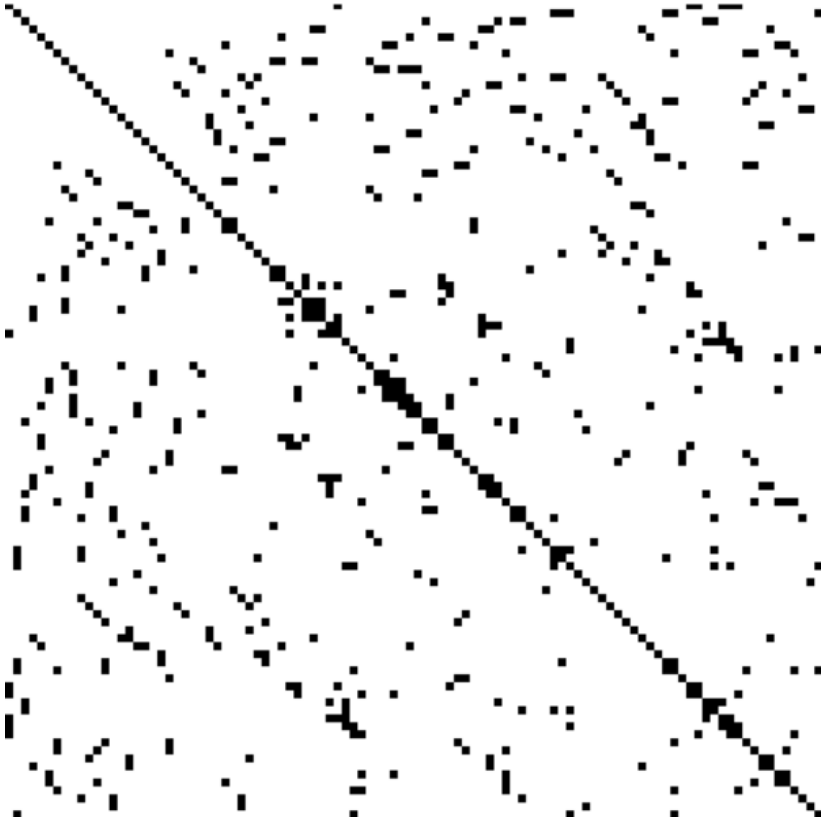
$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Hyperlink Matrix



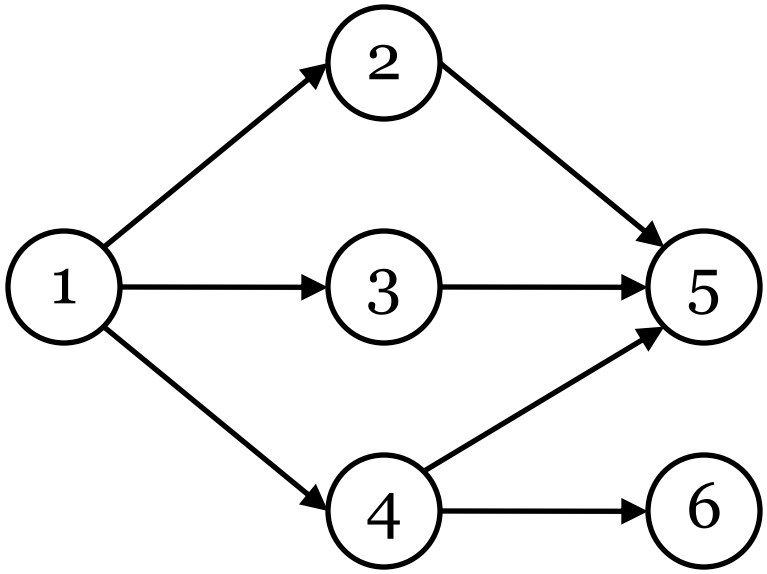
$$\mathbf{H} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Sparse Matrices



Adjacency List

better for sparse matrices



1	{ 2, 3, 4 }
2	{ 5 }
3	{ 5 }
4	{ 5, 6 }
5	{ }
6	{ }

PageRank

$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix}_{k+1}^{\mathrm{T}} = d \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix}_{\mathrm{k}}^{\mathrm{T}} \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \frac{d}{N} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix}_{\mathrm{k}}^{\mathrm{T}} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^{\mathrm{T}} + \frac{(1-d)}{N} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^{\mathrm{T}}$$

$$\mathbf{r}_{k+1}^{\mathrm{T}} = d\mathbf{r}_k^{\mathrm{T}}\mathbf{H} + \frac{d}{N}\mathbf{r}_k^{\mathrm{T}}\mathbf{a}\mathbf{e}^{\mathrm{T}} + \frac{(1-d)}{N}\mathbf{e}^{\mathrm{T}}$$

$\mathbf{a}$	dangling node vector
$\mathbf{e}^{\mathrm{T}}$	vector of all 1

Convergence

How many iterations?

$$\cong \frac{-n}{\log_{10} d}$$

How to check convergence?

$$\sum \left| r_{k+1}(p_i) - r_k(p_i) \right| < \varepsilon$$

n	number of significant digits
$\varepsilon = 10^{-n}$	tolerance

Convergence

significant digits	iterations
1	15
2	29
3	43
4	57
5	71
6	86
7	100
8	114
9	128
10	142
11	156
12	171
13	185

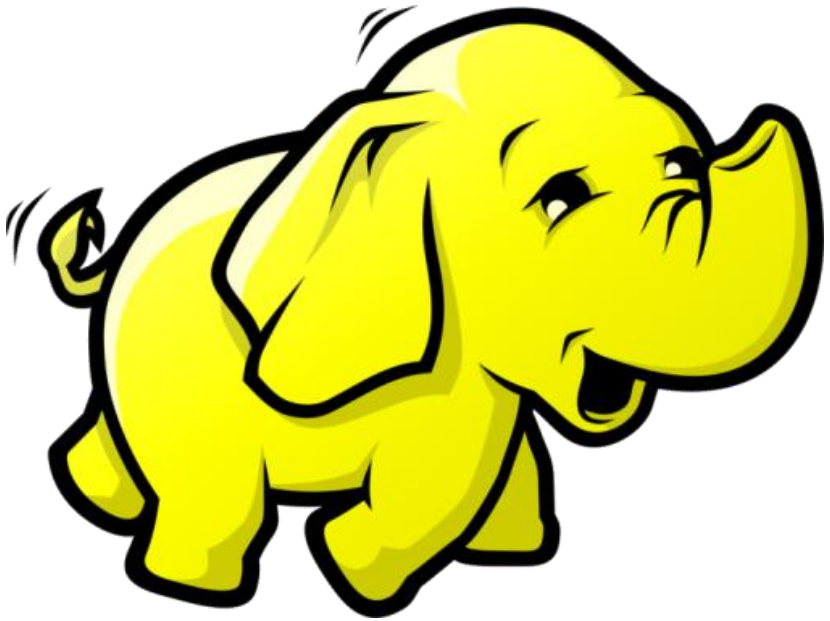
Power Method

- Slow to converge
- Each iteration complexity: O(N)
- Overall complexity: #iterations O(N) = O(N)
- Minimal storage: **H** sparse matrix, no completely dense matrices need to be stored

Storage Requirements

- Sparse hyperlink matrix **H**
  - number of non zero elements (each a double)
- Sparse binary dangling node vector
  - number of dangling nodes (each a boolean)
- PageRank values for the current iteration
  - N elements (each a double)
- (optional) PageRank values for the previous iteration to measure tolerance error
  - N elements (each a double)

Implementing PageRank with MapReduce



Adjacency List

1	r_1	2	3	4
2	r_2	5		
3	r_3	5		
4	r_4	5	6	
5	r_5			
6	r_6			

MapReduce

job 3 – page ranks from backward links

- map
  - input
    - key = p value = ($r_p, p_1, p_2, \ldots, p_n$)
  - output
    - key = p_i value = r_p/n $i = (1, 2, \ldots, n)$
    - key = p value = ($p_1, p_2, \ldots, p_n$)
- reduce
  - input
    - key = p values = (r_j/n_j)*, ($p_1, p_2, \ldots, p_n$)
  - output
    - key = p value = ($r_p, p_1, p_2, \ldots, p_n$) $r_p = d\sum_j \frac{r_j}{n_j} + r_d + \frac{(1-d)}{N}$

MapReduce

job 2 – contribution from dangling pages

- map
  - input
    - key = … value = r_p *dangling page*
  - output
    - key = 1 value = r_p N total number of pages
- combine and reduce
  - input
    - key = 1 values = (r_j)* $r_d = \frac{d}{N} \sum_j r_j$
  - output
    - key = … value = r_d only one value

MapReduce
job 1 – total number of pages

- map
  - input
    - key = p value = ...
  - output
    - key = 1 value = 1
- combine and reduce
  - input
    - key = 1 values = (v_j)* $N = \sum_j v_j$
  - output
    - key = ... value = N

Adjacency List

1	r^{k+1}_1	r^k_1	2	3	4
2	r^{k+1}_2	r^k_2	5		
3	r^{k+1}_3	r^k_3	5		
4	r^{k+1}_4	r^k_4	5	6	
5	r^{k+1}_5	r^k_5			
6	r^{k+1}_6	r^k_6			

MapReduce
job 4 – check for convergence

- map
  - input
    - key = p value = (r^{k+1}_p, r^k_p, p_1, p_2, … , p_n)
  - output
    - key = 1 value = abs (r^{k+1}_p - r^k_p)
- combine and reduce
  - input
    - key = 1 values = (v_j)* $\varepsilon = \sum_j v_j$
  - output
    - key = … value = ε ε *tolerance*

Putting all together

- job 1 – total number of pages
- for max n iterations or until convergence
  - job 2 – contribution from dangling pages
  - job 3 – page ranks from backward links
  - every y iterations
    - job 4 – check for convergence

- Total number of jobs <= $1 + 2n + n/y$

Having Fun with PageRank

- Intelligent surfer
  - Change rows of the hyperlink matrix **H** so long they remain probability distributions
  - Teleportation vector (a.k.a. personalization vector) instead of random jumps
- CiteRank to rank papers (using a time dependant decay factor to shape probability distributions of the hyperlink matrix)
- Social networks
- Ranking schemes: evaluation and comparison techniques (without involving humans?)
- Ranking schemes for directed labelled multi-graphs (a.k.a. RDF)?

Ranking Papers: CiteSeer Dataset

CiteSeer	(1)	(2)	PageRank	Title (Year)
340126	yes	yes	0.00157983	New Directions in Cryptography Invited Paper (1976)
549100	no	no	0.00121952	Structure and Complexity of Relational Queries (1982)
548351	no	no	0.00120267	Computable Queries for Relational Data Bases (1980)
527057	yes	yes	0.00114733	Optimization by Simulated Annealing (1983)
516071	no	no	0.00112389	Probabilistic Methods in Combinatorics (1974)
28289	yes	no	0.00108669	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems (1978)
552631	no	no	0.00107492	Fast Anisotropic Gauss Filtering (2001)
328445	no	yes	0.00101743	Scheduling Algorithms for Multiprogramming in a Hard-Read-Time Environment (1973)
239544	no	no	0.00096108	Discrepancy in Arithmetic Progressions (1996?)
148879	yes	no	0.00094061	Yacc: Yet Another Compiler-Compiler (1975)
311874	no	yes	0.00091705	Graph-Based Algorithms for Boolean Function Manipulation (1986)
93436	no	no	0.00090491	Privacy Enhancement for Internet Electronic Mail: Part II (1993)
219414	no	no	0.00084181	Privacy Enhancement for Internet Electronic Mail: Part III (1993)
567230	no	no	0.00080948	A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks (1986)
20336	no	no	0.00075584	Generalised Additive Models (1995)
524648	yes	no	0.00074717	Implementing Remote Procedure Calls (1984)
15205	no	no	0.00073840	Congestion Avoidance and Control (1988)
35316	no	no	0.00069750	Relational Queries Computable in Polynomial Time (1986)
76766	yes	no	0.00068785	The UNIX Time-Sharing (1974)
351230	no	no	0.00067404	History of Circumscription (1993)

CiteSeer - http://citeseer.ist.psu.edu/{*CID*}

(1) http://en.wikipedia.org/wiki/List_of_important_publications_in_computer_science

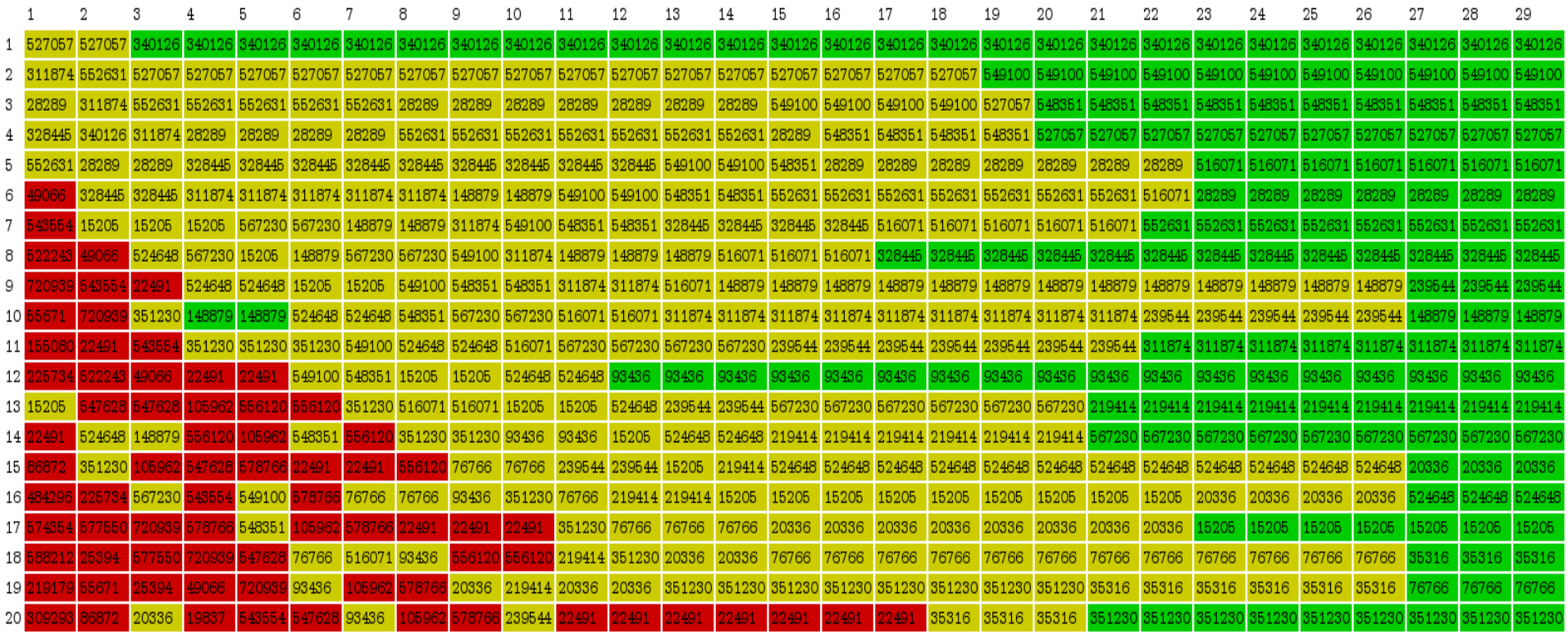
(2) http://scholar.google.com/scholar?as_q=%22+%22&num=100&as_subj=eng

Dirty Data

```
a  b  c
b  h  k  p
c
d  a  a  c
e  s
f  f  b
g
```

- Ignore duplicate links (d **a a** c) and self-references (**f f** b)
- Implicit dangling nodes (h, k, p, s)
- If the data is dirty, the Google matrix will not be stochastic and a unique solution as well as convergence are not guaranteed (with a sufficient high number of iterations, you might get as result ∞)

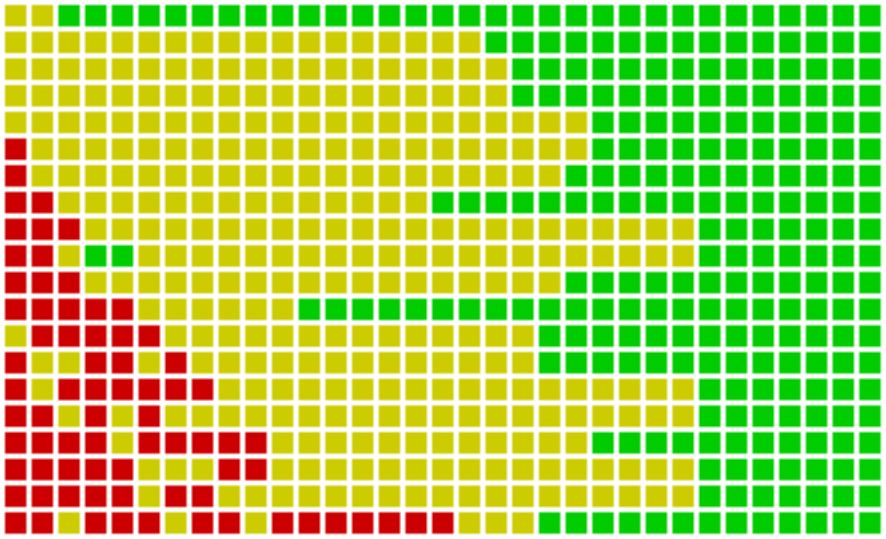
Convergence in Practice



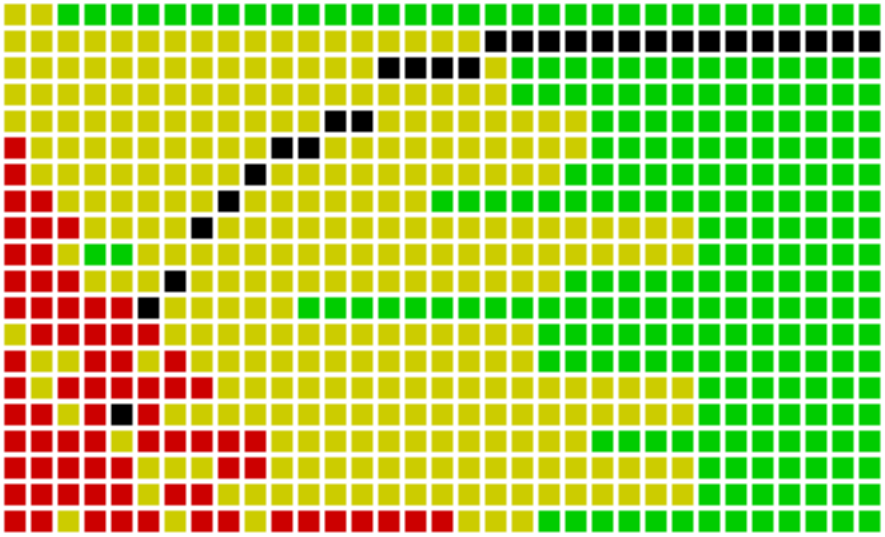
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	527057	527057	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126	340126
2	311874	552631	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057	549100	549100	549100	549100	549100	549100	549100	549100	549100	549100	549100
3	28289	311874	552631	552631	552631	552631	552631	28289	28289	28289	28289	28289	28289	28289	549100	549100	549100	549100	527057	548351	548351	548351	548351	548351	548351	548351	548351	548351	548351
4	328445	340126	311874	28289	28289	28289	28289	552631	552631	552631	552631	552631	552631	552631	28289	548351	548351	548351	548351	527057	527057	527057	527057	527057	527057	527057	527057	527057	527057
5	552631	28289	28289	328445	328445	328445	328445	328445	328445	328445	328445	328445	549100	549100	548351	28289	28289	28289	28289	28289	28289	28289	516071	516071	516071	516071	516071	516071	516071
6	49066	328445	328445	311874	311874	311874	311874	311874	148879	148879	549100	549100	548351	548351	552631	552631	552631	552631	552631	552631	552631	516071	28289	28289	28289	28289	28289	28289	28289
7	543554	15205	15205	15205	567230	567230	148879	148879	311874	549100	548351	548351	328445	328445	328445	328445	516071	516071	516071	516071	516071	552631	552631	552631	552631	552631	552631	552631	552631
8	522243	49066	524648	567230	15205	148879	567230	567230	549100	311874	148879	148879	148879	516071	516071	516071	328445	328445	328445	328445	328445	328445	328445	328445	328445	328445	328445	328445	328445
9	720939	543554	22491	524648	524648	15205	15205	549100	548351	548351	311874	311874	516071	148879	148879	148879	148879	148879	148879	148879	148879	148879	148879	148879	148879	148879	239544	239544	239544
10	55671	720939	351230	148879	148879	524648	524648	548351	567230	567230	516071	516071	311874	311874	311874	311874	311874	311874	311874	311874	311874	239544	239544	239544	239544	239544	148879	148879	148879
11	155080	22491	543554	351230	351230	351230	549100	524648	524648	516071	567230	567230	567230	567230	239544	239544	239544	239544	239544	239544	239544	311874	311874	311874	311874	311874	311874	311874	311874
12	225734	522243	49066	22491	22491	549100	548351	15205	15205	524648	524648	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436	93436
13	15205	547628	547628	105962	556120	556120	351230	516071	516071	15205	15205	524648	239544	239544	567230	567230	567230	567230	567230	567230	219414	219414	219414	219414	219414	219414	219414	219414	219414
14	22491	524648	148879	556120	105962	548351	556120	351230	351230	93436	93436	15205	524648	524648	219414	219414	219414	219414	219414	219414	567230	567230	567230	567230	567230	567230	567230	567230	567230
15	86872	351230	105962	547628	578766	22491	22491	556120	76766	76766	239544	239544	15205	219414	524648	524648	524648	524648	524648	524648	524648	524648	524648	524648	524648	524648	20336	20336	20336
16	484296	225734	567230	543554	549100	578766	76766	76766	93436	351230	76766	219414	219414	15205	15205	15205	15205	15205	15205	15205	15205	15205	20336	20336	20336	20336	524648	524648	524648
17	574354	577550	720939	578766	548351	105962	578766	22491	22491	22491	351230	76766	76766	76766	20336	20336	20336	20336	20336	20336	20336	20336	15205	15205	15205	15205	15205	15205	15205
18	588212	25394	577550	720939	547628	76766	516071	93436	556120	556120	219414	351230	20336	20336	76766	76766	76766	76766	76766	76766	76766	76766	76766	76766	76766	76766	35316	35316	35316
19	219179	55671	25394	49066	720939	93436	105962	578766	20336	219414	20336	20336	351230	351230	351230	351230	351230	351230	351230	351230	35316	35316	35316	35316	35316	35316	76766	76766	76766
20	309293	86872	20336	19837	543554	547628	93436	105962	578766	239544	22491	22491	22491	22491	22491	22491	22491	35316	35316	35316	351230	351230	351230	351230	351230	351230	351230	351230	351230

Rows: ranking positions. Columns: iterations. Cells: document ids.
Red: document should not be in the first 20 results. Yellow: document in the first 20 results, but wrong position. Green: document in the first 20 results, correct position.

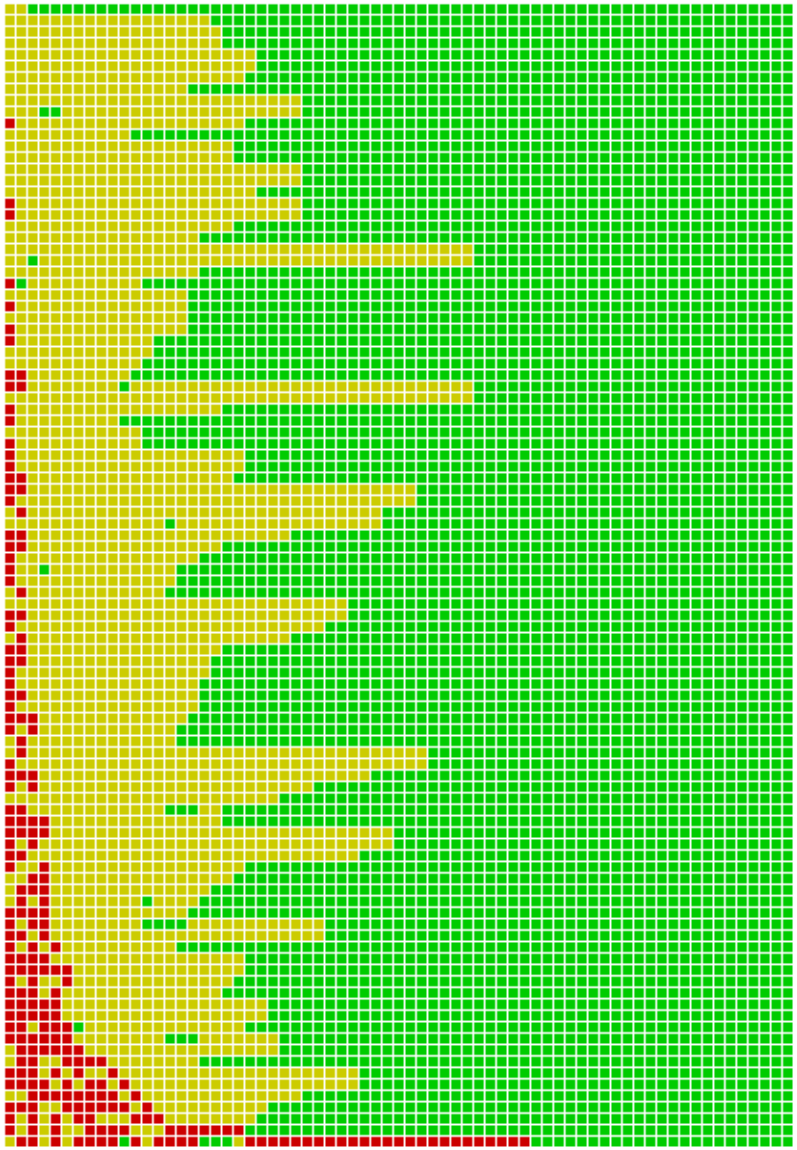
Convergence in Practice



Convergence in Practice



Convergence in Practice



References

- "*Google's PageRank and Beyond: The Science of Search Engine Rankings*"
  Amy N. Langville and Carl D. Meyer
  Princeton University Press (2006), ISBN 0-691-12202-4
  http://press.princeton.edu/titles/8216.html

- "*The anatomy of a large-scale hypertextual Web search engine*"
  Sergey Brin and Lawrence Page
  In Proc. of the Seventh International World Wide Web Conference (WWW 1998)
  http://ilpubs.stanford.edu:8090/361/

- "*The PageRank Citation Ranking: Bringing Order to the Web*"
  Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd
  Technical Report, Stanford InfoLab (1999)
  http://ilpubs.stanford.edu:8090/422/

- "*The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank*"
  Matthew Richardson and Pedro Domingos
  In Proc. of Advances in Neural Information Processing Systems (2002)
  http://www.cs.washington.edu/homes/pedrod/papers/nips01b.pdf

- "*Ranking Scientific Publications Using a Simple Model of Network Traffic*"
  Dylan Walker, Huafeng Xie, Koon-Kiu Yan, Sergei Maslov
  Journal of Statistical Mechanics (2007)
  http://arxiv.org/abs/physics/0612122v1


LABS hp