



# Analisi e Visualizzazione delle Reti Complesse

**NS15 - Link Analysis and Web  
Search**

**Prof. Rossano Schifanella**



# Agenda

- Searching the Web
- Link Analysis
  - HITS: Hubs and Authorities (+ Spectral Analysis)
  - Page Rank (+ Spectral Analysis)
  - Random Walks and PR
- Practical implications
  - Modern Web search
  - Link Analysis beyond the Web

# PageRank and HITS

- **Centrality measures** for nodes in directed networks
- Sergey Brin and Larry Page introduced PageRank in 1998 as a key ingredient of Google
- Jon Kleinberg introduced HITS in 1999
- Both are based on eigenvector centrality and designed for web information retrieval
- In NetworkX:

```
PR_dict = nx.pagerank(D)      # D must be a DiGraph
H_dict, A_dict = nx.hits(G)    # G should be a DiGraph
```

## Reading material:

- [Page, Lawrence; Brin, Sergey; Motwani, Rajeev and Winograd, Terry, The PageRank citation ranking: Bringing order to the Web. 1999](#)
- [Jon Kleinberg, Authoritative sources in a hyperlinked environment Journal of the ACM 46 \(5\): 604-32, 1999. doi:10.1145/324133.324140.](#)
- [A. Langville and C. Meyer, "A survey of eigenvector methods of web information retrieval.", SIAM Review, vol. 47, No. 1](#)

# Searching the Web

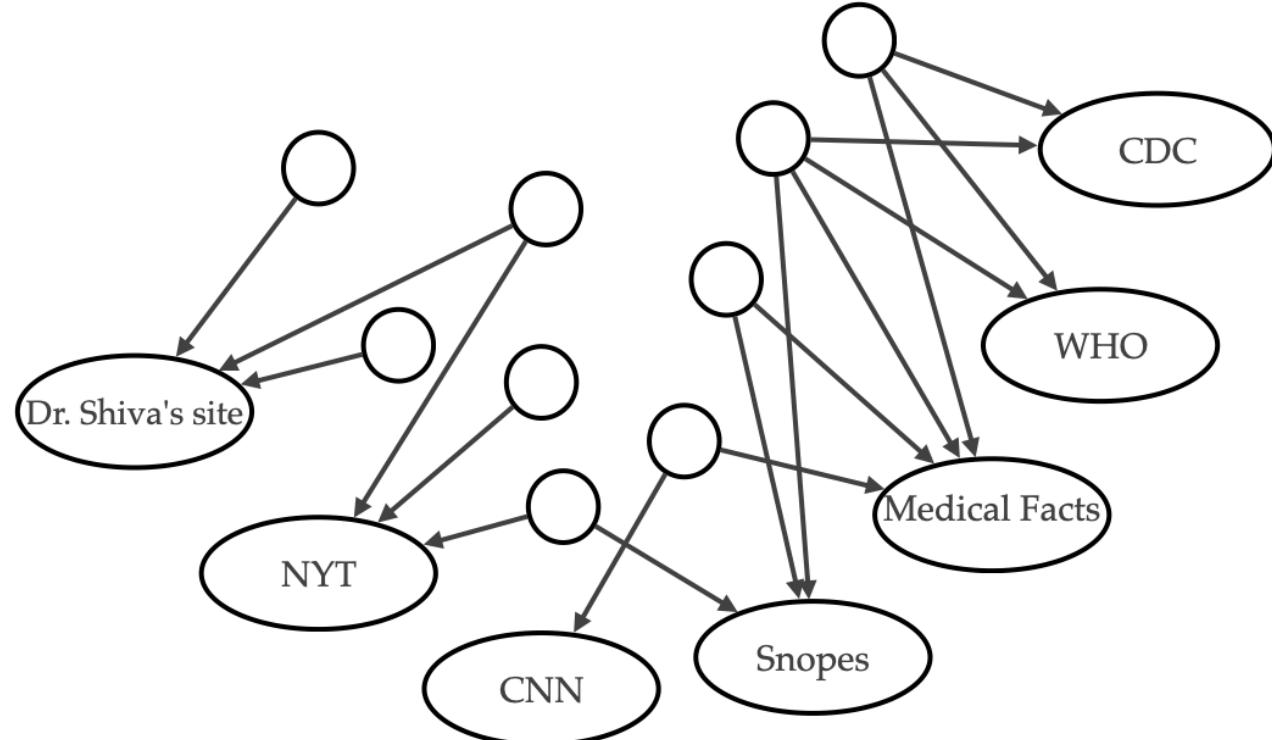
- Search Engine
  - problem: how to rank (web) pages related to a given topic
- Information Retrieval
  - automated strategies to search in libraries, scientific papers, repositories, and so on
  - in response to keywords based queries
- List of keywords is "inexpressive" (e.g., polysemy, synonymy)
- **Diversity:** given a topic we find pages created by many authors
- Pages are **dynamic** and always changing
- **Filters:** what is important?
- Can the structure of the Web, dominated by links, help us to find such filters?
  - first attempt: count words in documents
  - can we do better?

# HITS: Link Analysis using Hubs and Authorities

- Disclaimer: the term "Hub" is used slightly differently here w.r.t. previous lectures
- Information contained "between" pages can be used as well
- Count **in-links**:
  - select documents on a given topic
  - **in-links are a measure of authority** of a page on such a topic: it is an implicit endorsement from the community of web pages' authors

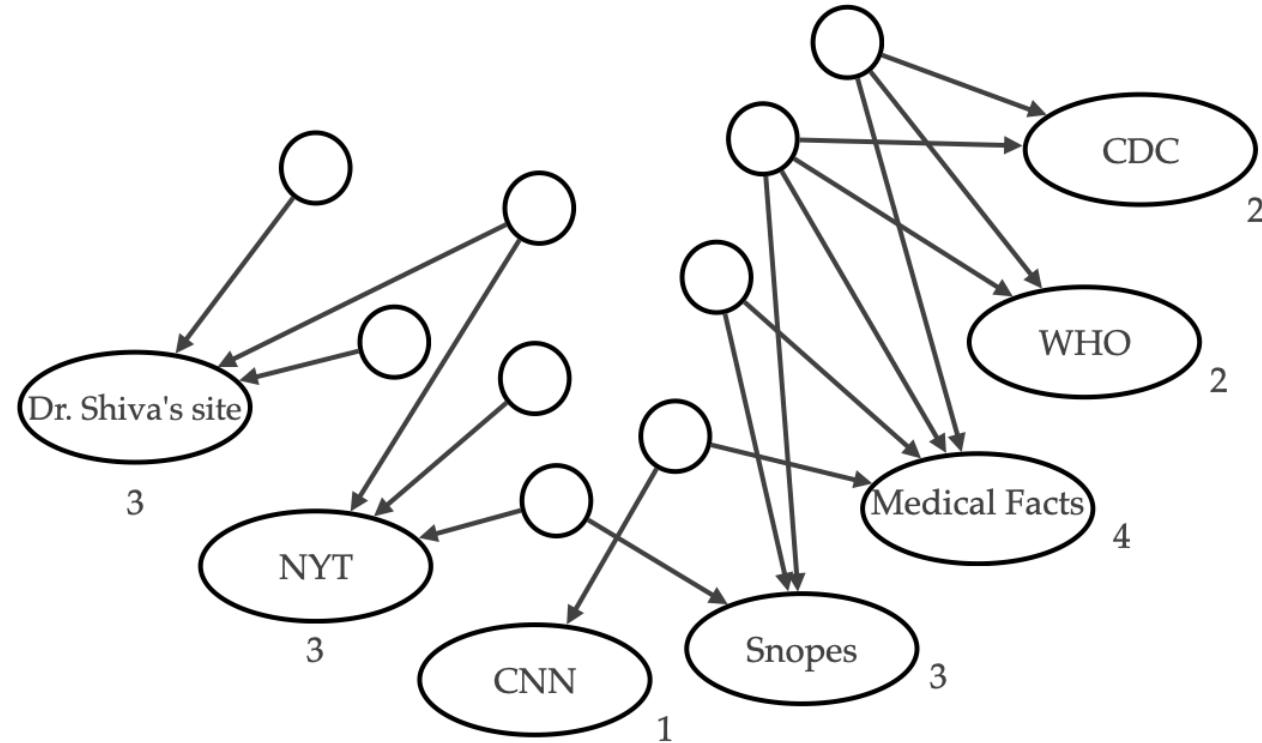
## Finding lists

- query: "covid-19"
- we could have intuitively correct authority values with in-degrees
- Lists: pages that provide many different out-links to other pages



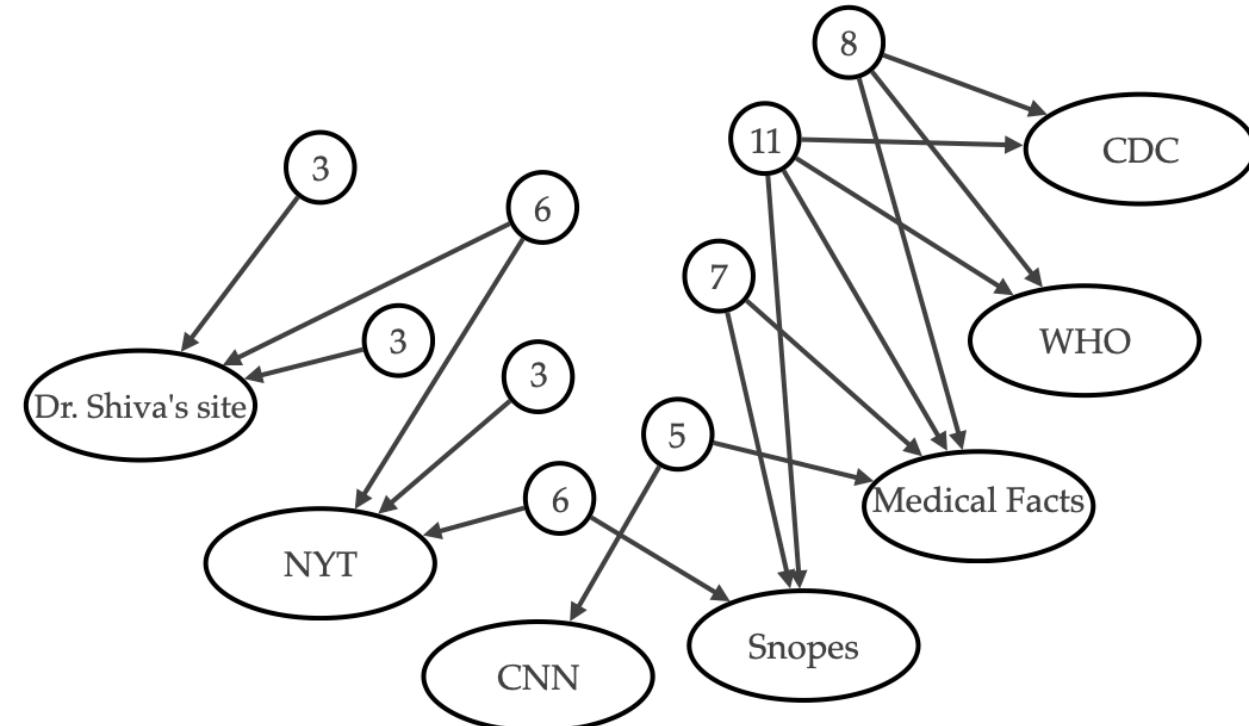
## List values

- Hub values for list pages: the sum of in-links received by all pages they link to
- Assumption: list pages have a better sense for where the good results are
- Authorities are often competitors!



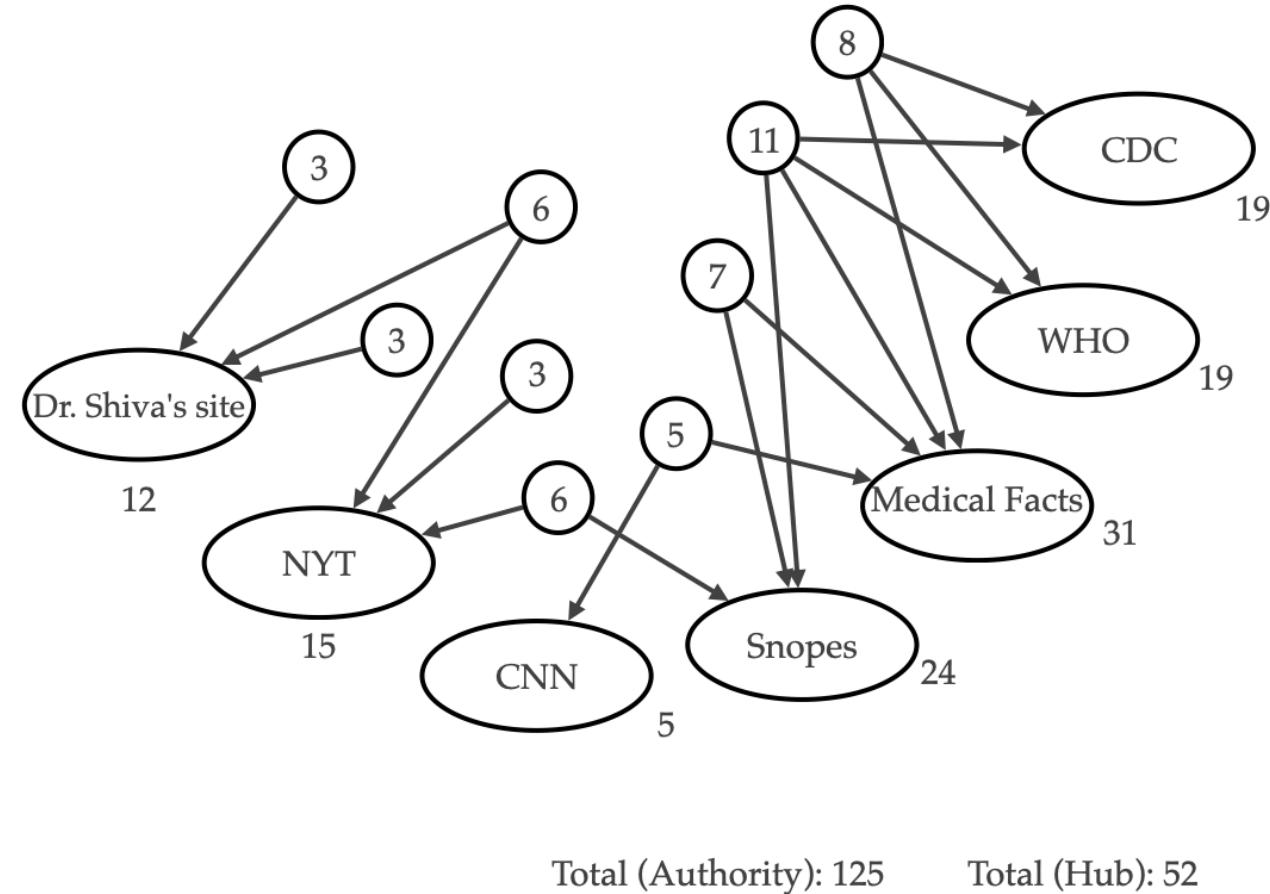
## The principle of repeated improvement

- Now we can weight links from hubs more heavily
- Recalculate
- Why stop here? we can refine values at both sides



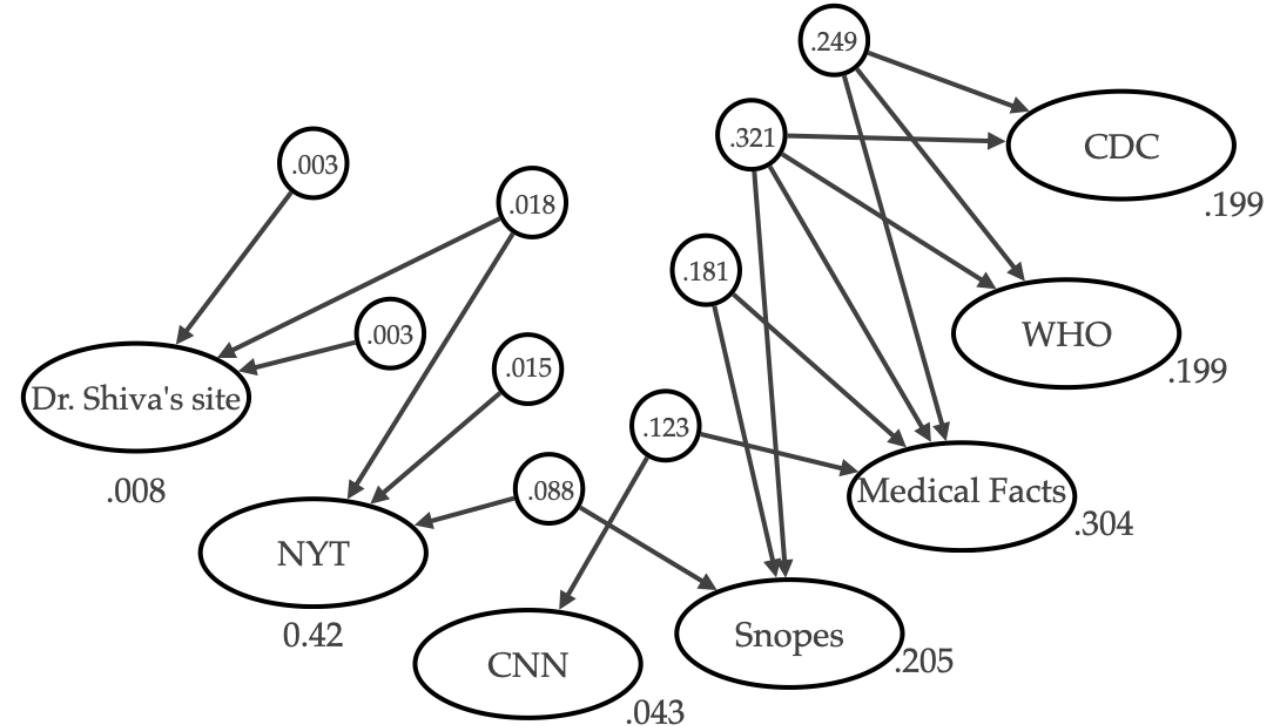
## Hubs and Authorities

- step 0: every node has both a hub and an authority value (all initialized to 1)
- step k (Update Rules):
- authorities are updated after k-1's hub values
- hubs are updated on new authority values
- normalize hub and authority values



## Stabilization

- We will prove that normalized values converge when  $i \rightarrow \infty$
- Stabilization: initial values are not important
- limiting values for hubs and authorities are properties of the links structure
- Different form of game theoretical concept of equilibrium



# Introduction to spectral analysis

- We need to analyze the methods to compute hubs and authorities values
- Pre-requisites:
  - linear algebra
  - vector and matrix multiplication
- Limiting values are coordinates in eigenvectors for given eigenvalues in matrices derived from our graphs
- Eigenvalues/eigenvectors calculation to study the structure of networks
  - spectral analysis

# Spectral Analysis of Hubs and Authorities

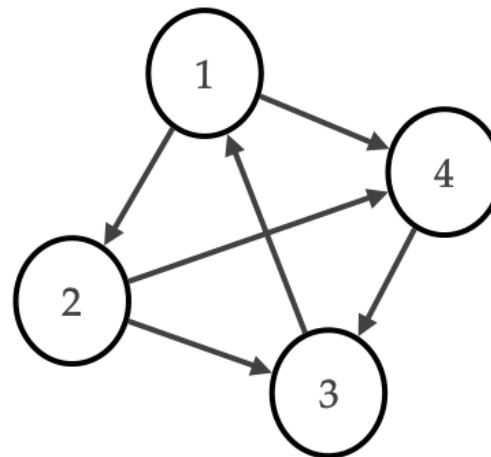
Adjacency Matrix  $M$

$1 \dots n$  nodes

$M : n \times n$

$$M_{ij} = \begin{cases} 1, & \text{if } (i, j) \text{ is an edge} \\ 0, & \text{otherwise} \end{cases}$$

- The adjacency matrix is not necessarily efficient for computational representations
  - for practical use, consider adjacency lists or edge lists instead



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

## Update rules

$\mathbf{h}$ ,  $\mathbf{a}$ : n-dimensional vectors (respectively, hub and authorities values)

### Hub Update Rule

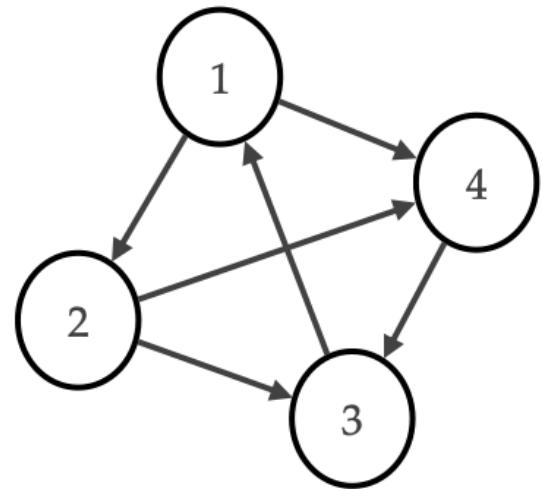
$$h_i \leftarrow \sum_{j=1}^n M_{ij} a_j = M_{i1} a_1 + M_{i2} a_2 + \dots + M_{in} a_n$$

$$\mathbf{h} \leftarrow \mathbf{M} \cdot \mathbf{a}$$

### Authority Update Rule

$$a_i \leftarrow \sum_{j=1}^n M_{ji} h_j$$

$$\mathbf{a} \leftarrow \mathbf{M}^T \cdot \mathbf{h}$$



$$\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix}$$

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



## Understanding the k-step hub-authority computation

initialization:  $\mathbf{h}^{<0>} = \underbrace{(1, 1, \dots, 1)}_n$

after  $k$  applications of the rule:  $\mathbf{a}^{<k>}, \mathbf{h}^{<k>}$

# Understanding the k-step hub-authority computation

First step

$$\begin{aligned}\mathbf{a}^{<1>} &= \mathbf{M}^T \mathbf{h}^{<0>} \\ \mathbf{h}^{<1>} &= \mathbf{M} \mathbf{a}^{<1>} = \mathbf{M} \mathbf{M}^T \mathbf{h}^{<0>}\end{aligned}$$

Second step

$$\begin{aligned}\mathbf{a}^{<2>} &= \mathbf{M}^T \mathbf{h}^{<1>} = (\mathbf{M} \mathbf{M}^T) \mathbf{M}^T \mathbf{h}^{<0>} \\ \mathbf{h}^{<2>} &= \mathbf{M} \mathbf{a}^{<2>} = (\mathbf{M} \mathbf{M}^T) (\mathbf{M} \mathbf{M}^T) \mathbf{h}^{<0>} = (\mathbf{M} \mathbf{M}^T)^2 \mathbf{h}^{<0>}\end{aligned}$$

...

$k^{th}$  step

$$\begin{aligned}\mathbf{a}^{<k>} &= (\mathbf{M} \mathbf{M}^T)^{k-1} \mathbf{M}^T \mathbf{h}^{<0>} \\ \mathbf{h}^{<k>} &= (\mathbf{M} \mathbf{M}^T)^k \mathbf{h}^{<0>}\end{aligned}$$

$\mathbf{a}, \mathbf{h}$  vectors: multiplication of an initial vector  $\mathbf{h}^{<0>}$  by larger powers of  $\mathbf{M}^T \mathbf{M}$  and  $\mathbf{M} \mathbf{M}^T$ .

## Multiplications and Eigenvectors

**normalization:** we can find constants  $c$  and  $d$  s.t.  $\frac{\mathbf{h}^{<k>}}{c^k}$  and  $\frac{\mathbf{a}^{<k>}}{d^k}$ .

We want to prove that they converge for  $k \rightarrow \infty$

Focus on the hub vectors:

If  $\frac{\mathbf{h}^{<k>}}{c^k} = \frac{(\mathbf{M}\mathbf{M}^T)^k \cdot \mathbf{h}^{<0>}}{c^k}$  converges to a limit  $\mathbf{h}^{<*>}$ , then I can expect that:

$$c \cdot \mathbf{h}^{<*>} = (\mathbf{M}\mathbf{M}^T) \cdot \mathbf{h}^{<*>}$$

Hence, we need to prove that the sequence of  $\frac{\mathbf{h}^{<k>}}{c^k}$  converges to the eigenvector of  $\mathbf{M}\mathbf{M}^T$

## Eigenvectors and square matrices

- Observe that  $\mathbf{M}\mathbf{M}^T$  is a symmetric matrix
- Fact 1: "Any symmetric matrix  $n \times n$  has a set of  $n$  eigenvectors  $\mathbf{z}_1, \dots, \mathbf{z}_n$  that are orthogonal and all unit vectors - that is they form a basis for the space  $\mathbb{R}^n \implies \mathbf{z}_i \cdot \mathbf{z}_j = 0$  and  $\mathbf{z}_i \cdot \mathbf{z}_i = 1$ "
- That means that for our symmetric  $\mathbf{M}\mathbf{M}^T$  we can find:
  - $n$  mutual orthogonal eigenvectors:  $\mathbf{z}_1, \dots, \mathbf{z}_n$  (the spectrum of  $\mathbf{M}\mathbf{M}^T$ )
  - $n$  corresponding eigenvalues:  $c_1, \dots, c_n$
  - Let's sort eigenvectors s.t. corresponding eigenvalues:  $c_1 \geq c_2 \geq \dots \geq c_n$
  - Assume (for now):  $c_1 > c_2$

- Let's consider  $\mathbf{x} \in \mathbb{R}^n$

$$\begin{aligned}(\mathbf{M}\mathbf{M}^T)\mathbf{x} &= (\mathbf{M}\mathbf{M}^T)(p_1\mathbf{z}_1 + p_2\mathbf{z}_2 + \dots + p_n\mathbf{z}_n) \\&= p_1(\mathbf{M}\mathbf{M}^T)\mathbf{z}_1 + p_2(\mathbf{M}\mathbf{M}^T)\mathbf{z}_2 + \dots + p_n(\mathbf{M}\mathbf{M}^T)\mathbf{z}_n \\&= p_1c_1\mathbf{z}_1 + p_2c_2\mathbf{z}_2 + \dots + p_nc_n\mathbf{z}_n\end{aligned}$$

- We will use this equation to analyze multiplication by larger powers of  $(\mathbf{M}\mathbf{M}^T)$

$$(\mathbf{M}\mathbf{M}^T)^k\mathbf{x} = p_1c_1^k\mathbf{z}_1 + p_2c_2^k\mathbf{z}_2 + \dots + p_nc_n^k\mathbf{z}_n$$

# Convergence of the Hub-Authority computation

vector of hub scores at step  $k$ :

$$\mathbf{h}^{<\mathbf{k}>} = (\mathbf{M}\mathbf{M}^T)^k \cdot \mathbf{h}^{<0>}$$

$$\mathbf{h}^{<0>} = q_1 \mathbf{z}_1 + q_2 \mathbf{z}_2 + \dots + q_n \mathbf{z}_n$$

$$\mathbf{h}^{<\mathbf{k}>} = c_1^k q_1 \mathbf{z}_1 + c_2^k q_2 \mathbf{z}_2 + \dots + c_n^k q_n \mathbf{z}_n$$

Let's divide both sides by  $c_1^k$ :

$$\frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = \frac{c_1^k q_1 \mathbf{z}_1}{c_1^k} + \frac{c_2^k q_2 \mathbf{z}_2}{c_1^k} + \dots + \frac{c_n^k q_n \mathbf{z}_n}{c_1^k}$$

$$\text{assumption: } c_1 > c_2 \Rightarrow \lim_{k \rightarrow \infty} \left( \frac{c_2}{c_1} \right)^k = 0$$

Then:

$$\lim_{k \rightarrow \infty} \frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = q_1 \mathbf{z}_1$$

## Wrapping up

- (i) a limit in the direction of  $\mathbf{z}_1$  is reached regardless of initial values of  $\mathbf{h}^{<0>}$ :

let's suppose that  $\mathbf{h}^{<0>} = \mathbf{x}$  and that is a positive vector:

$$\mathbf{x} = p_1\mathbf{z}_1 + p_2\mathbf{z}_2 + \dots + p_n\mathbf{z}_n \Rightarrow (\mathbf{M}\mathbf{M}^T)^k \mathbf{x} = c_1^k p_1\mathbf{z}_1 + c_2^k p_2\mathbf{z}_2 + \dots + c_n^k p_n\mathbf{z}_n$$

$$\lim_{k \rightarrow \infty} \frac{\mathbf{h}^{<k>}}{c_1^k} = p_1\mathbf{z}_1$$

- (ii) coefficient  $p_1$  (or  $q_1$ ) must be  $\neq 0$ : assuring that  $p_1\mathbf{z}_1$  (or  $q_1\mathbf{z}_1$ ) are non zero vectors, in the direction of  $\mathbf{z}_1$

- (iii) relax assumption:  $c_1 > c_2$

in general we can have  $l > 1$  eigenvalues s.t.  $c_1 = c_2 = \dots = c_l$  until we find that  $c_1 > c_{l+1}$

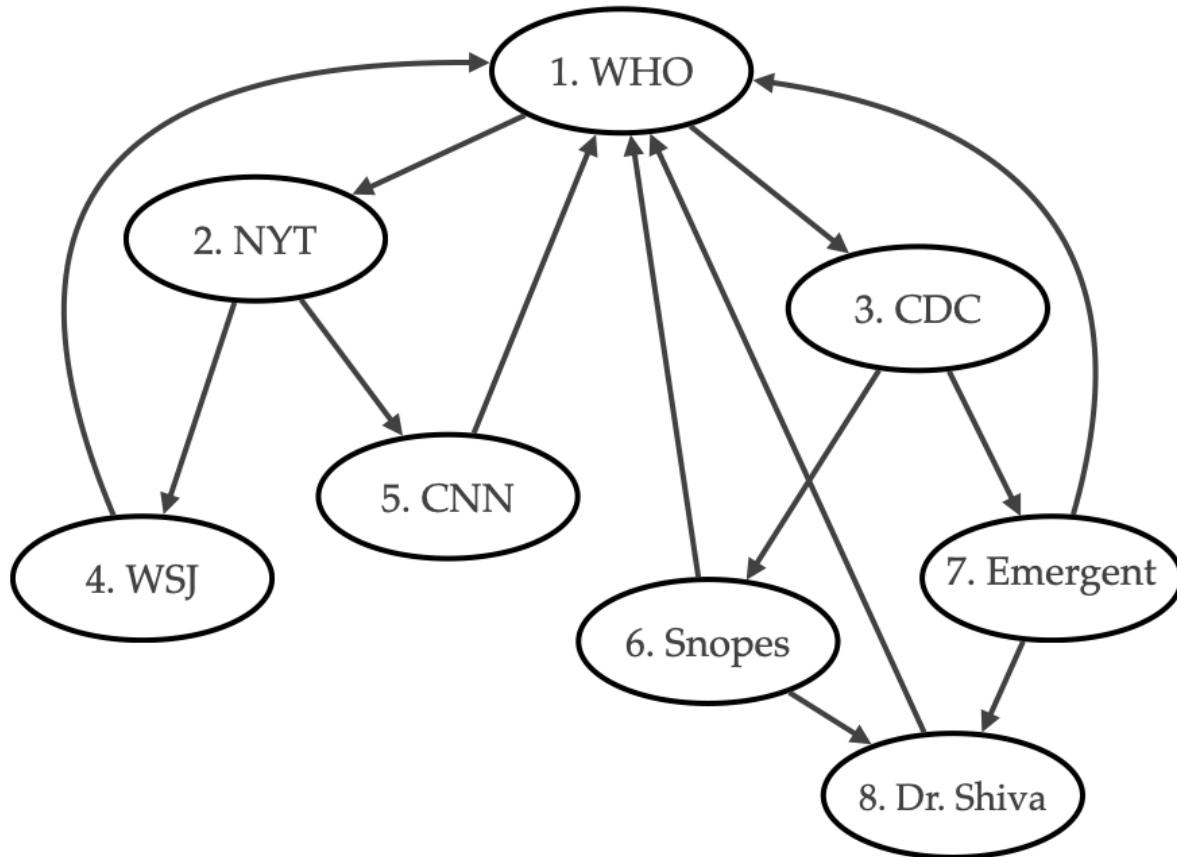
$$\begin{aligned} \frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} &= \frac{c_1^k q_1 \mathbf{z}_1 + c_2^k q_2 \mathbf{z}_2 + \dots + c_l^k q_l \mathbf{z}_l}{c_1^k} + \frac{c_{l+1} n^k q_{l+1} \mathbf{z}_{l+1} + \dots + c_n^k q_n \mathbf{z}_n}{c_1^k} \\ &= q_1 \mathbf{z}_1 + q_2 \mathbf{z}_2 + \dots + q_l \mathbf{z}_l + 0 \end{aligned}$$

with  $k \rightarrow \infty$  is still a convergence

- (iv) authority values: the argument is very similar to hub values (multiplication by  $\mathbf{M}^T \mathbf{M}$ )

# Page Rank

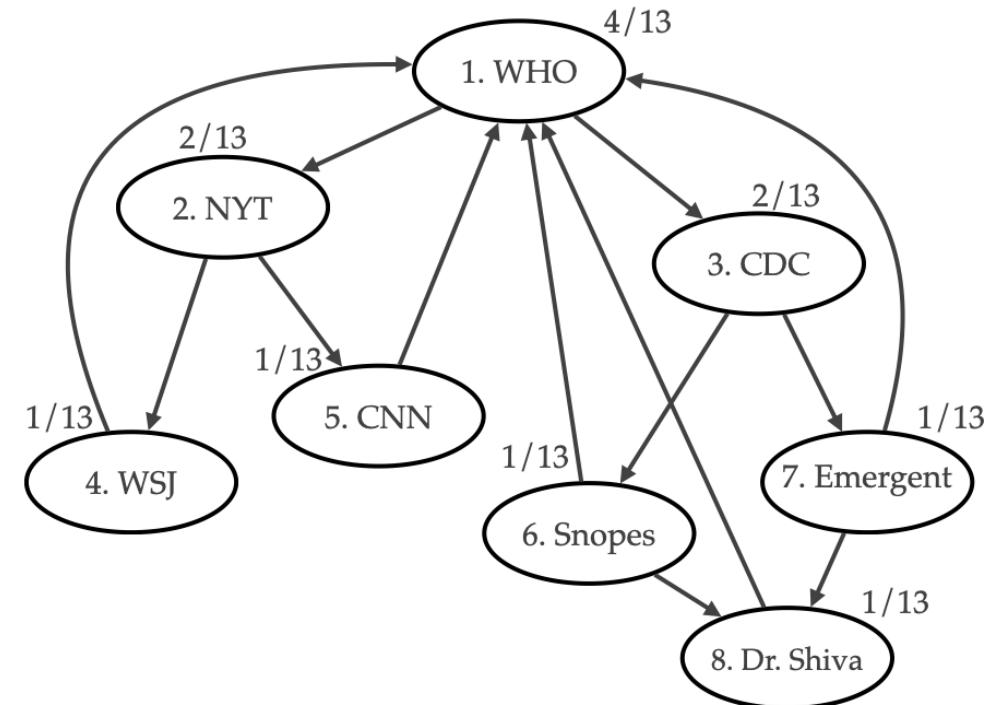
- **Endorsement** viewed as passing directly from one important node to another
  - endorsements received by in-links and passed across outgoing links
- Basic definition:
  - Step 0: Init all the pages  $p$  to a  $PR(p) = \frac{1}{n}$ , where  $n$  is the number of pages
  - Step k: Update all the  $PR(p)$  to the sum of all the receiving  $PR$  values, normalized by out-links



	1	2	3	4	5	6	7	8
0	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16
	...	...	...	...	...	...	...	...

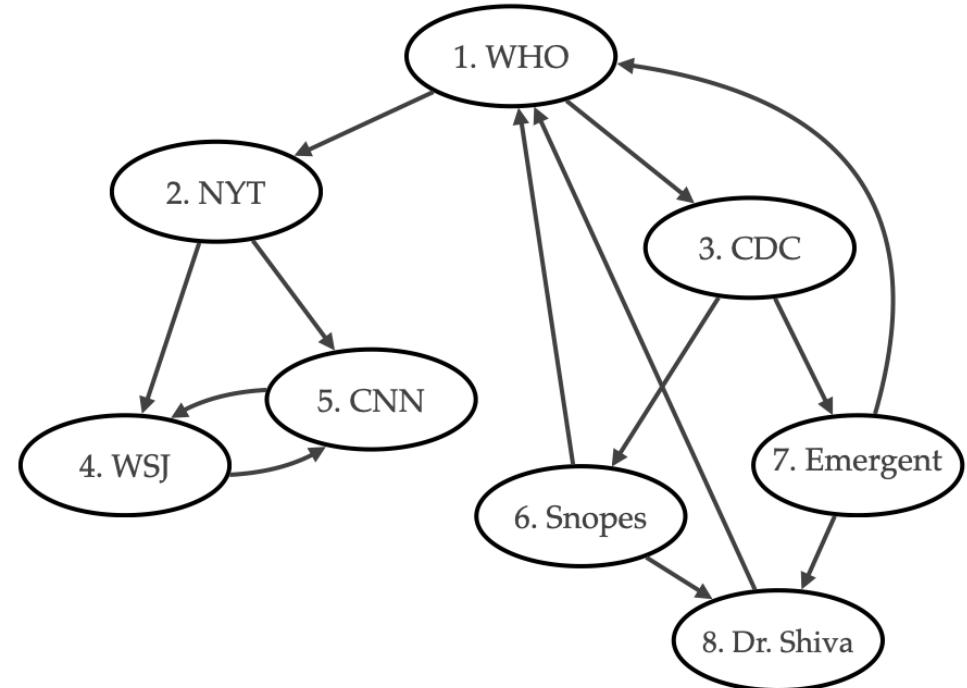
## PageRank and stabilization

- PR values of all the nodes converge when  $k \rightarrow \infty$  (but for some "degenerate cases")
- Equilibrium: if we apply our PR update rule, then our limiting values do not change



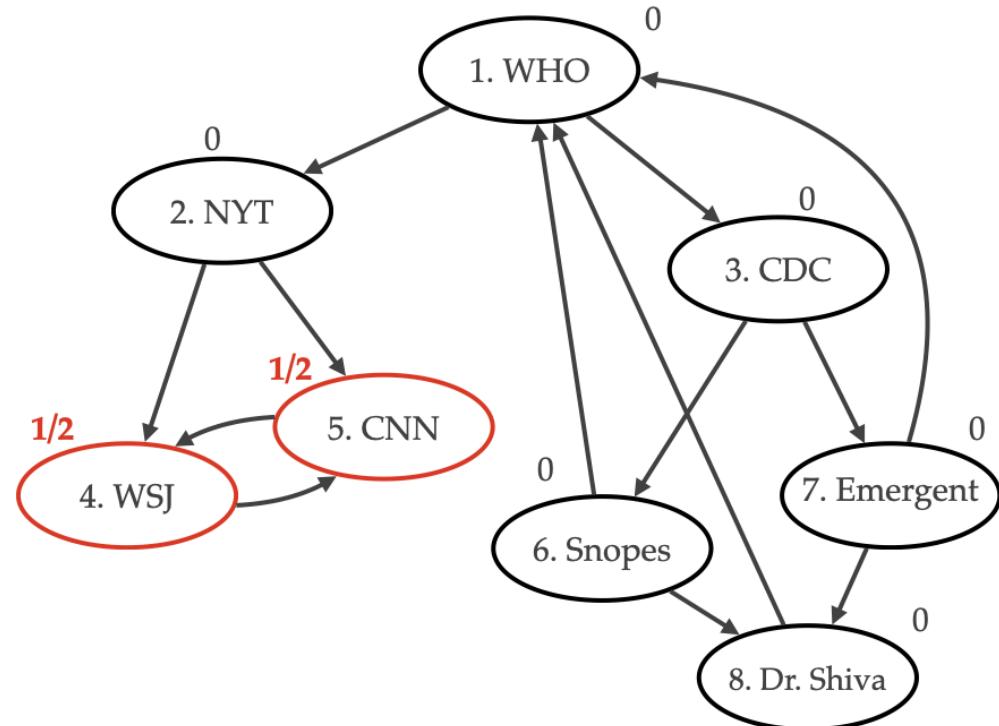
## Scaling the definition of PageRank

- "degenerate cases", the problem: in some networks some nodes receive all the PR values of the network



## Scaling the definition of PageRank

- "degenerate cases", the problem: in some networks some nodes receive all the PR values of the the network
- Applying PR update rule until we get equilibrium
- I can have degenerate cases in the Out Component of the Web



- Why? We do not have path back to some other nodes
- **Solution:** let's force this "fluid" to stream back to other nodes "sometimes":
- select a **scaling factor** (aka **damping factor**)  $s : s \in [0, 1]$
- get a portion  $s$  of PR values from in-links and then add  $(1 - s) \frac{1}{n}$
- Now we have convergence for  $k \rightarrow \infty$
- Observe: typically  $s \in [0.8, 0.9]$

# Spectral Analysis of Page Rank

- We need to analyze the methods to compute **page rank** values
- Pre-requisites:
  - linear algebra
  - vector and matrix multiplication
- Limiting values are coordinates in eigenvectors for given eigenvalues in matrices derived from our graphs
- Eigenvalues/eigenvectors calculation to study the structure of networks => spectral analysis

## Page Rank (revisited)

At step 0 (init):

$$\forall i : r_i = \frac{1}{n}; \ n: \# \text{ pages} \quad r_i = PR(i)$$

At step k:

$$\forall i : r_i = \sum_{j=1}^n M_{ji} \frac{r_j}{k_j^{\text{out}}} \quad (\text{basic PR update rule})$$

$$\forall i : r_i = s \cdot \sum_{j=1}^n M_{ji} \frac{r_j}{k_j^{\text{out}}} + (1 - s) \cdot \frac{1}{n} \quad (\text{scaled PR update rule})$$

## Using matrix notation

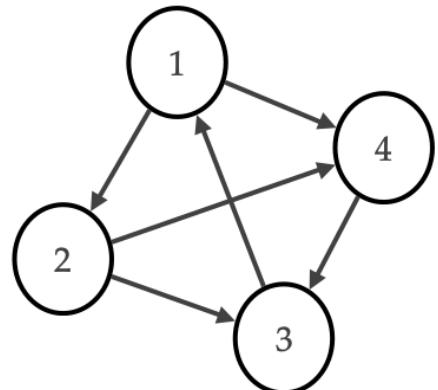
$\mathbf{N}$ : Matrix derived from  $M$

nodes:  $1, \dots, n$

$\mathbf{N} : n \times n$

$$N_{ij} = \begin{cases} \frac{1}{k_i^{\text{out}}}, & \text{if } (i, j) \\ 1, & \text{if } (i, j) \text{ is an edge, and } k_i^{\text{out}} = 0 \\ 0, & \text{otherwise} \end{cases}$$

$N_{ij}$ : the share of  $i$ 's  $PR$  that  $j$  should get in one update step



$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

# Update rule (basic and scaled)

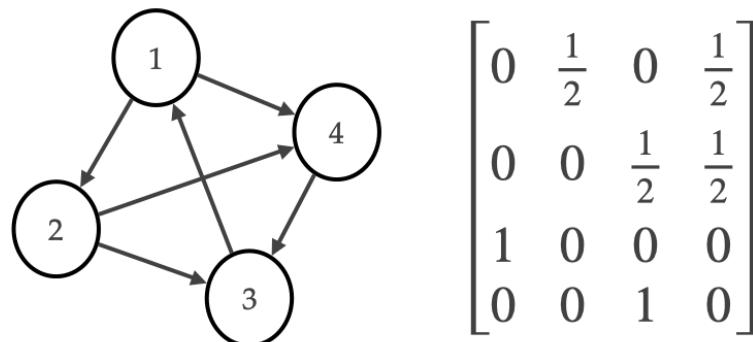
1. Basic update rule:

$$\forall i : r_i = \sum_{j=1}^n N_{ji} r_j \leftarrow N_{1i}r_1 + N_{2i}r_2 + \dots + N_{1n}r_n \mathbf{r} \leftarrow \mathbf{N}^T \cdot \mathbf{r}$$

2. Scaled update rule (factor  $s$ ):

$$\widetilde{N}_{ij} = s \cdot N_{ij} + (1 - s) \cdot \frac{1}{n}$$

3.



#### 4. Application of scaled update rule:

$$\forall i : r_i = \sum_{j=1}^n \widetilde{N}_{ji} r_j \leftarrow \widetilde{N}_{1i} r_1 + \widetilde{N}_{2i} r_2 + \dots + \widetilde{N}_{1n} r_n \mathbf{r} \leftarrow \widetilde{\mathbf{N}}^T \cdot \mathbf{r}$$

## Repeated improvement

$\mathbf{r}^{<0>} = \left( \frac{1}{n}, \dots, \frac{1}{n} \right)$ , initial PR vector

$$\mathbf{r}^{<k>} = (\widetilde{\mathbf{N}}^T)^k \cdot \mathbf{r}^{<0>}$$

Limiting vector  $\mathbf{r}^{<*>}$  satisfies  $\widetilde{\mathbf{N}}^T \cdot \mathbf{r}^{<*>} = 1 \cdot \mathbf{r}^{<*>}$

$\mathbf{r}^{<*>}$  should be an eigenvector of  $\widetilde{\mathbf{N}}^T$  with corresponding eigenvalue of 1 BUT  $\widetilde{\mathbf{N}}^T$  is not symmetric: this means that eigenvalues can be complex numbers and eigenvectors have no relationships to one another

## Convergence of the scaled PR update rule

$$\forall i, j : \tilde{N}_{ij} > 0$$

Perron's theorem

Matrix  $\mathbf{P}$  (with entries  $> 0$ )

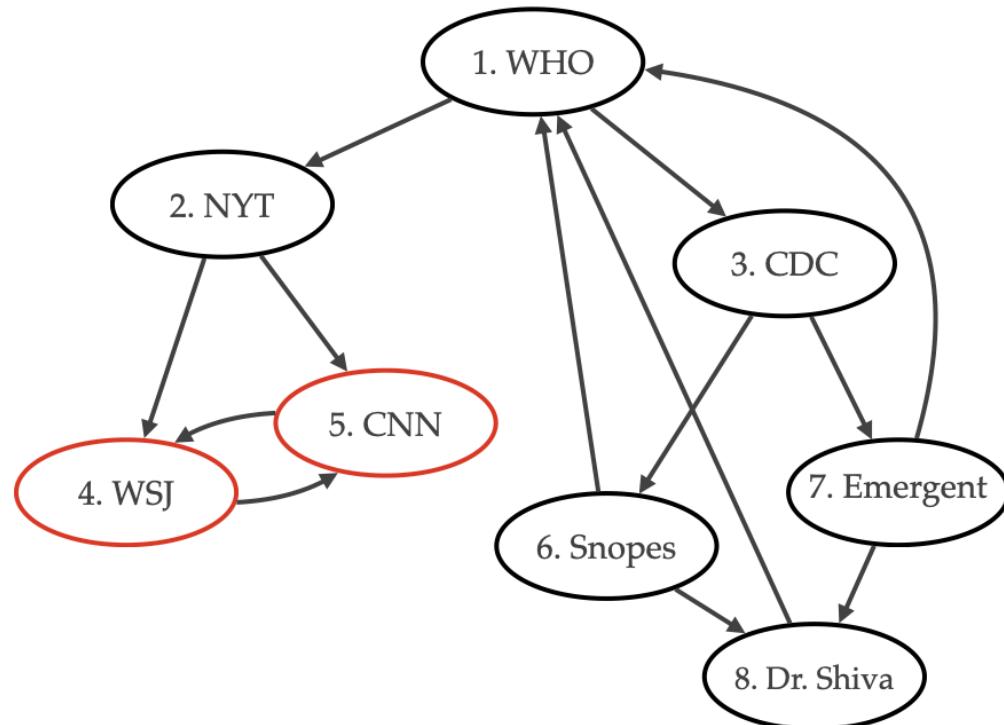
- i)  $\mathbf{P}$  has an eigenvalue  $c > 0$  s.t.  $c > c'$   $\forall c'$  (with  $c'$  another eigenvalue)
- ii) Exists an eigenvector  $\mathbf{y}$  with real positive values corresponding to  $c$ , and  $\mathbf{y}$  is unique (up to a multiplication constant)
- iii) if  $c = 1$ , then for any starting vector  $\mathbf{x} \neq \mathbf{0}$  with non negative coordinates, the sequence of vectors  $p^k \mathbf{x}$  converges to a vector in the direction of  $\mathbf{y}$  ( $k \rightarrow \infty$ )

## Random walks

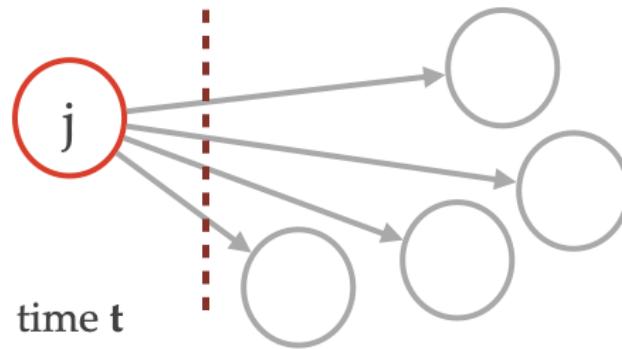
- randomly clicking from one page to another, picking each page with equal probability
- follow links for a sequence of length  $k$
- **claim:** the probability of being at page  $x$  after  $k$  steps is the application of the basic  $PR$  update rule
- **additional intuition:**  $PR(x)$  is the limiting probability that a random walk across hyperlinks will end up at  $x$  as we sum the walk for larger and larger number of steps

## Leakage

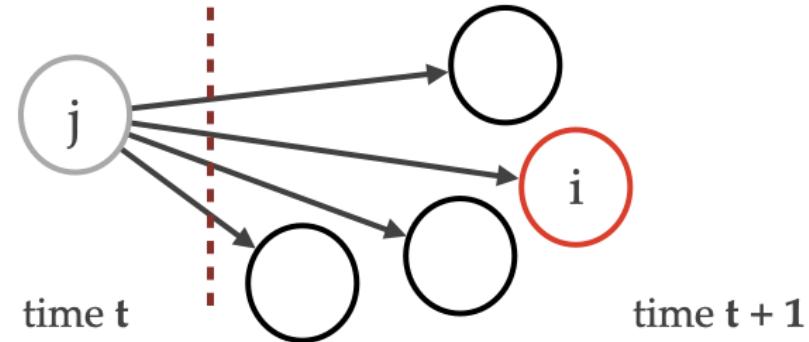
- The "leakage" of 4. and 5. has a natural interpretation: when the surfer reaches 4. or 5., then it is stuck forever
- Solution:
  - with probability  $s$ , the random walker clicks on an hyperlink in the page
  - with probability  $1 - s$ , it jumps to a randomly selected node



## Formulation of the PageRank using Random Walks



## Formulation of the PageRank using Random Walks



Which is the probability of being at node  $i$  at time  $t + 1$ ?

$b_1, b_2, \dots, b_n$ : the probabilities of being at node  $i$  in a given step.

$b_i \leftarrow \sum_{j=1}^n M_{ji} \frac{b_j}{k_j^{out}}$ : the probability of being at node  $j$  in the following step

Using matrix  $\mathbf{N}$  :  $b_i \leftarrow N_{1i}b_1 + N_{2i}b_2 + \dots + N_{1n}b_n \Rightarrow \mathbf{b} \leftarrow \mathbf{N}^T \cdot \mathbf{b}$

**claim:**  $PR$  of page  $i$  is exactly the probability of being at node  $i$  after  $k$  step.

## A scaled version of the random walk

For a given probability  $s$ : the walker follows a random outgoing edge

With prob  $(1 - s)$ : the walker is teleported uniformly at random to another node

$$b_i \leftarrow s \cdot \sum_{j=1}^n M_{ji} \frac{b_j}{k_j^{out}} + \frac{(1 - s)}{n}$$

Using matrix:

$$\widetilde{N} : b_i \leftarrow \widetilde{N}_{1i}b_1 + \widetilde{N}_{2i}b_2 + \dots + \widetilde{N}_{1n}b_n \Rightarrow \mathbf{b} \leftarrow \widetilde{\mathbf{N}}^T \cdot \mathbf{b}$$

**claim:**  $PR$  is equivalent to the scaled version of random walks.

## Practical implications (also beyond the Web)

### Modern Web search

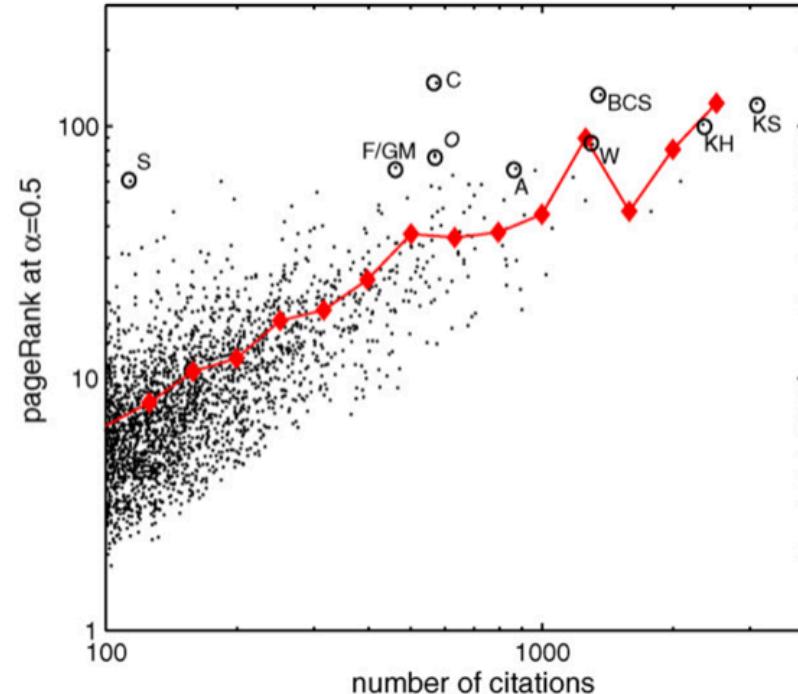
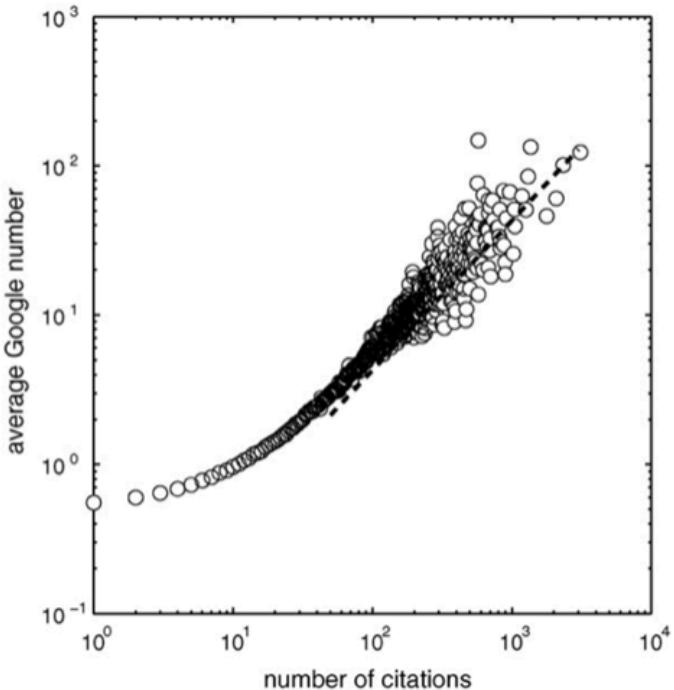
- Google today doesn't use *PR* anymore (original Page&Brin's paper: 2001)
- Hilltop (an extension of HITS) has been probably used for a while
- anchor texts
- clicking behavior
- and much more (and who knows what they are actually using!)

## SEO vs Google

- SEO: Search Engine Optimization
- SEO companies: reverse engineering of search engine's ranking functions
- SE companies: define new measures
- ... in loop!
- Feedback effects: perfect results are "moving targets"
- It is a game theoretic principle

# Page Rank and citation analysis

- paper: "Finding Scientific Gems with Google Page Ranks" (2007)
- dataset: collection of scientific papers with their references
- positive correlation between number of citations and average *PR* values
- BUT outliers are papers with "limited" number of citations



Google rank	Google # ( $\times 10^{-4}$ )	Cite rank	# cites	Publication			Title		Author(s)
1	4.65	54	574	PRL	10	531	1963	Unitary symmetry and leptonic...	N. Cabibbo
2	4.29	5	1364	PR	108	1175	1957	Theory of superconductivity	J. Bardeen, L. Cooper, and J. Schrieffer
3	3.81	1	3227	PR	140	A1133	1965	Self-consistent equations...	W. Kohn and L.J. Sham
4	3.17	2	2460	PR	136	B864	1964	Inhomogeneous electron gas	P. Hohenberg and W. Kohn
5	2.65	6	1306	PRL	19	1264	1967	A model of leptons	S. Weinberg
6	2.48	55	568	PR	65	117	1944	Crystal statistics I	L. Onsager
7	2.43	56	568	RMP	15	1	1943	Stochastic problems in...	S. Chandrasekhar
8	2.23	95	462	PR	109	193	1958	Theory of the Fermi interaction	R.P. Feynman and M. Gell-Mann
9	2.15	17	871	PR	109	1492	1958	Absence of diffusion in...	P.W. Anderson
10	2.13	1853	114	PR	34	1293	1929	The theory of complex spectra	J.C. Slater

**Pros:** *PR* helps to find "gems" in networks!

**Cons:** Indicators can change our behaviors



## Reading material

[ns2] Chapter 14 (14.1-14.6) Link Analysis and Web Search

[ns1] Chapter 4 (4.3) [Simplified description of PageRank]



# Q & A

