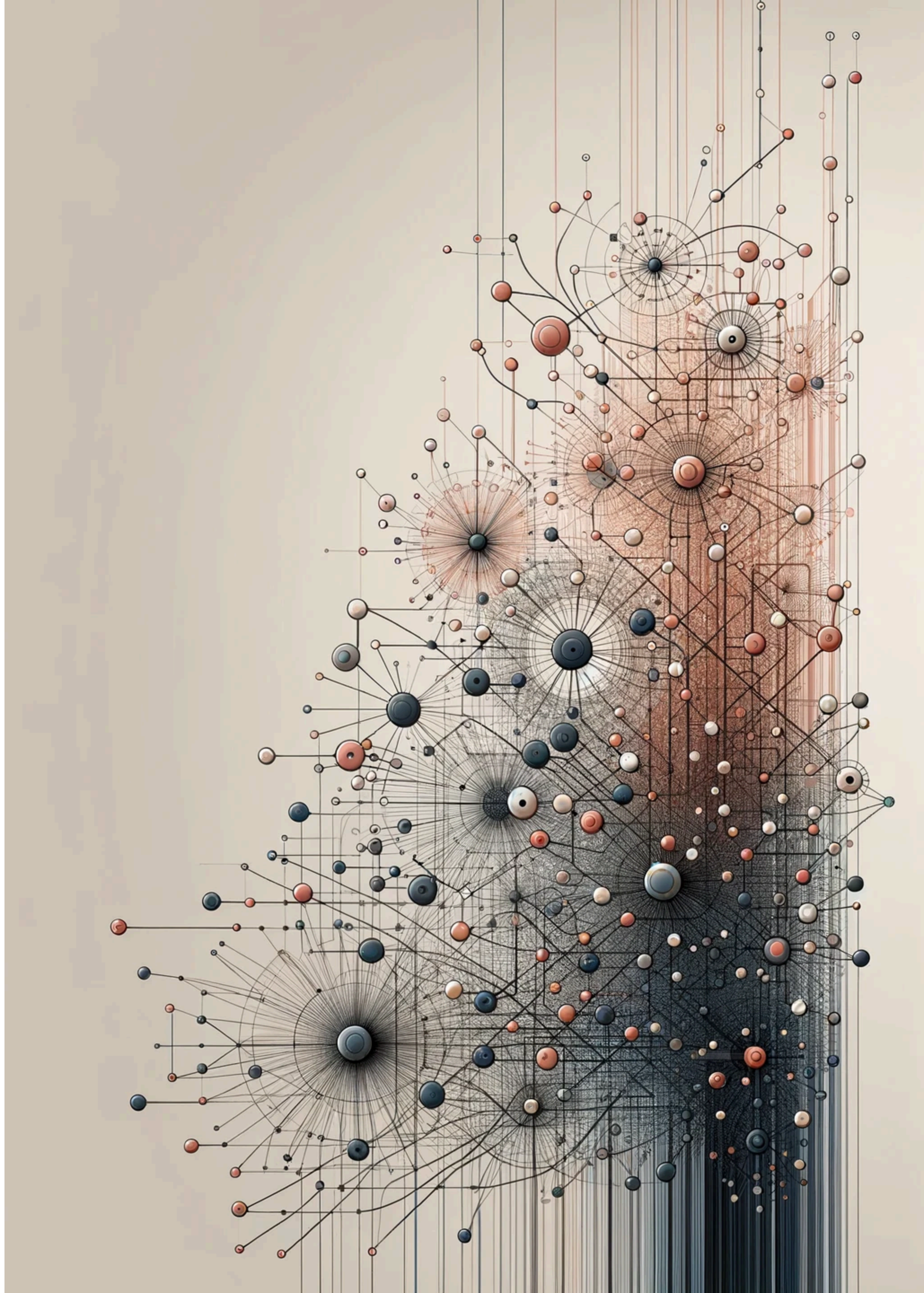
Analisi e Visualizzazione delle Reti Complesse

NS19 - 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

UNIVERSITÀ DI TORINO

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 a large variety of authors
  - e.g., experts, novices, children, conspiracy theorists
- Pages are **dynamic** and always changing
- **Filters**: abundance of information, 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?

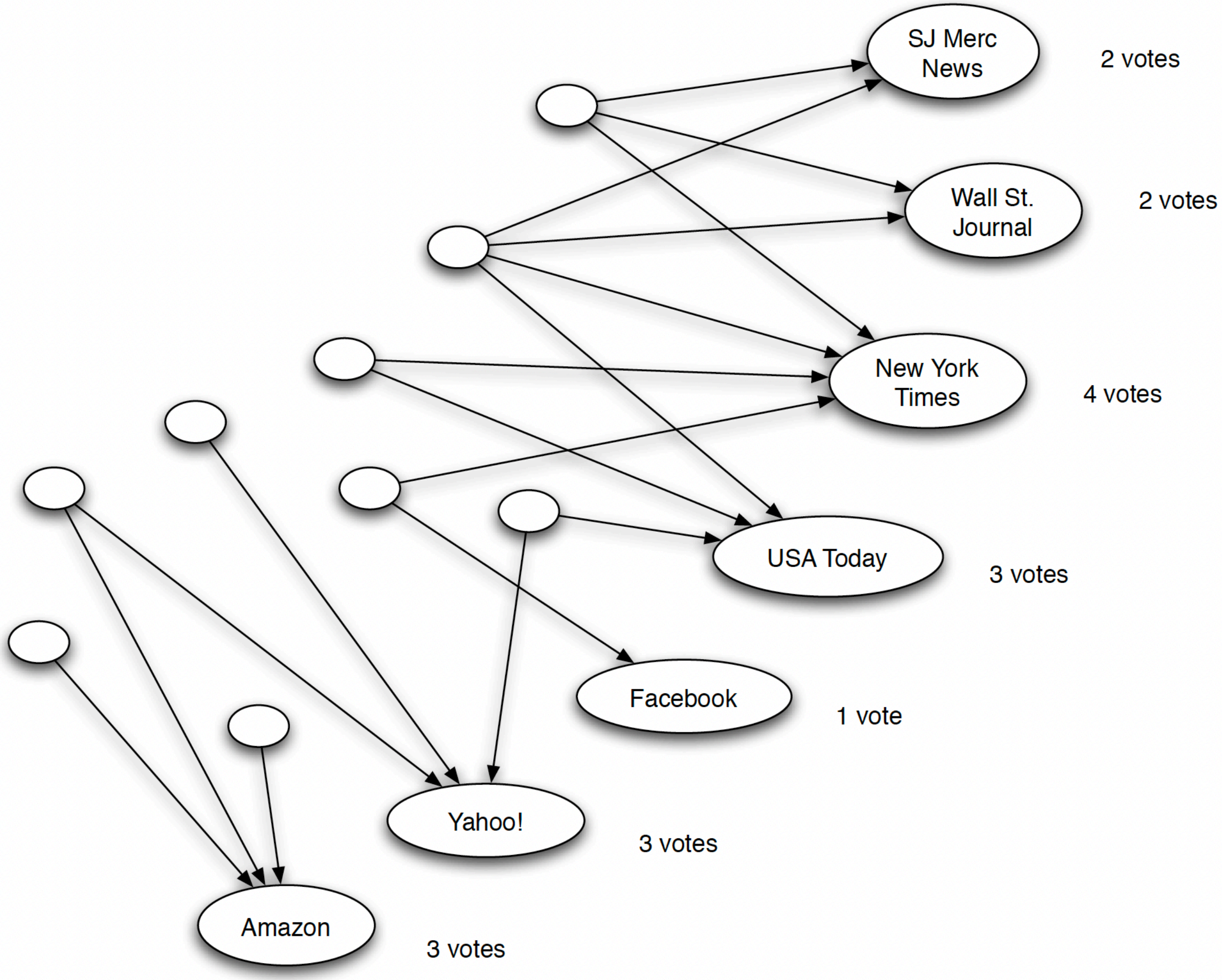
HITS: Link Analysis using Hubs and Authorities

- 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
  - It is hard for search engines to automatically assess the intent of each link. In aggregate, we assume they mean endorsement.

UNIVERSITÀ DI TORINO

Finding lists

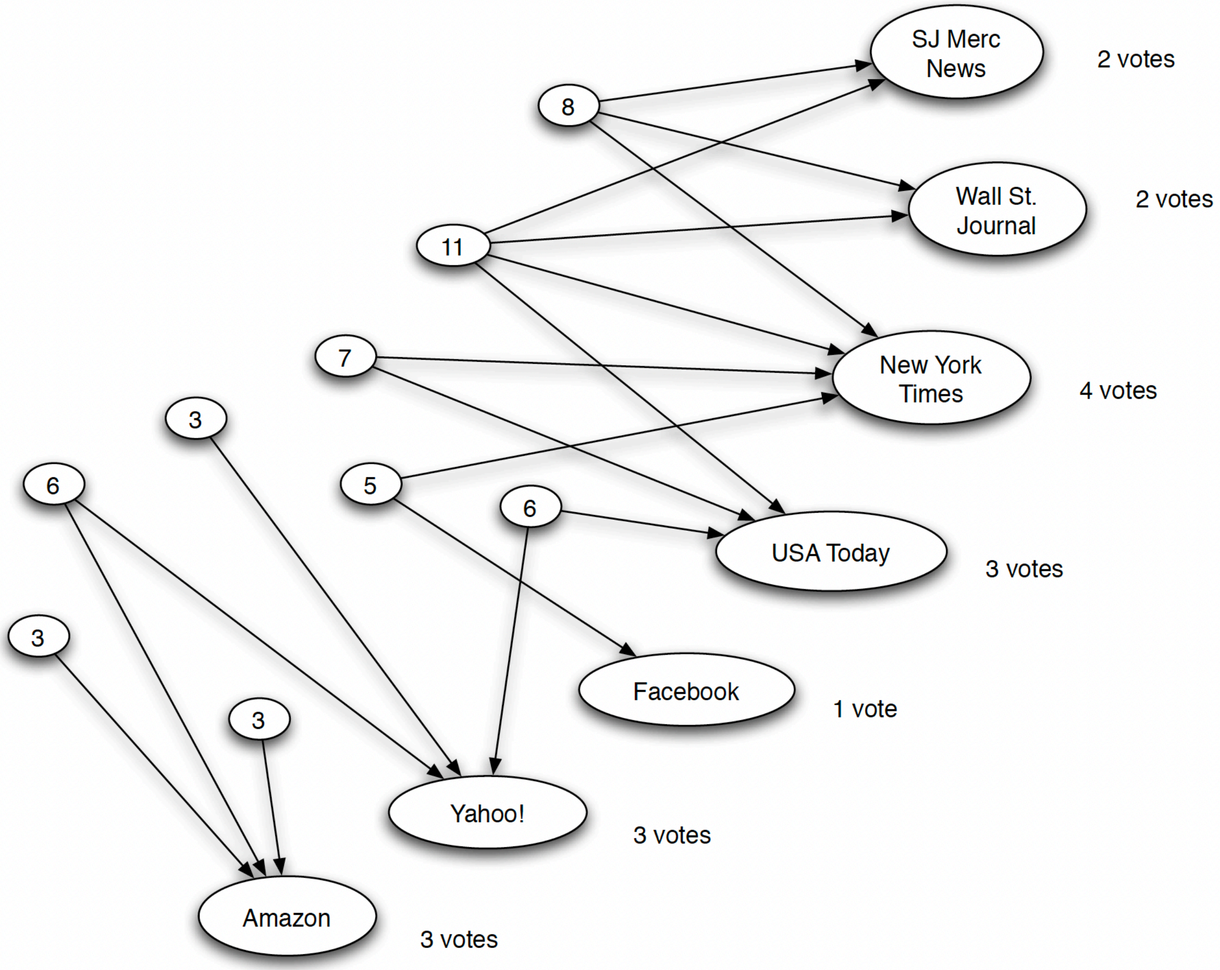
- query: "newspapers"
- we could have intuitively valid authority values with **in-degrees**
    - however, we get a mix of prominent newspapers along with pages that are going to receive a lot of in-links no matter what the query is
- **Lists**: pages that provide many different out-links to other pages
    - among the pages casting votes, a few of them voted for many of the pages that received a lot of votes



UNIVERSITÀ DI TORINO

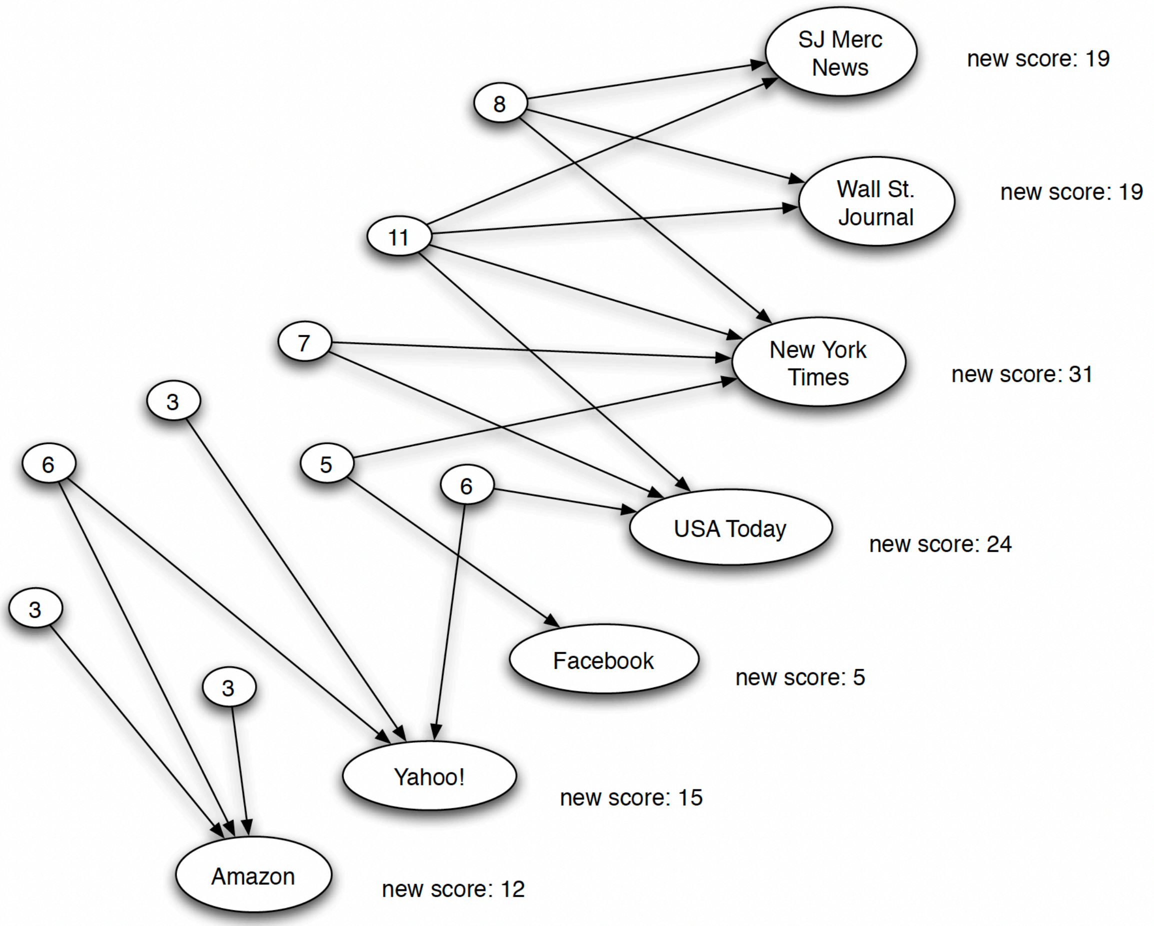
List values

- Hub value for a list page:
  the sum of in-links received by all pages it linked to
  - in other words: the sum of the votes received by all pages that it voted for
- Assumptions:
  - list pages have a better sense for where the good results are
  - authorities are often **competitors**!



The Principle of Repeated Improvement

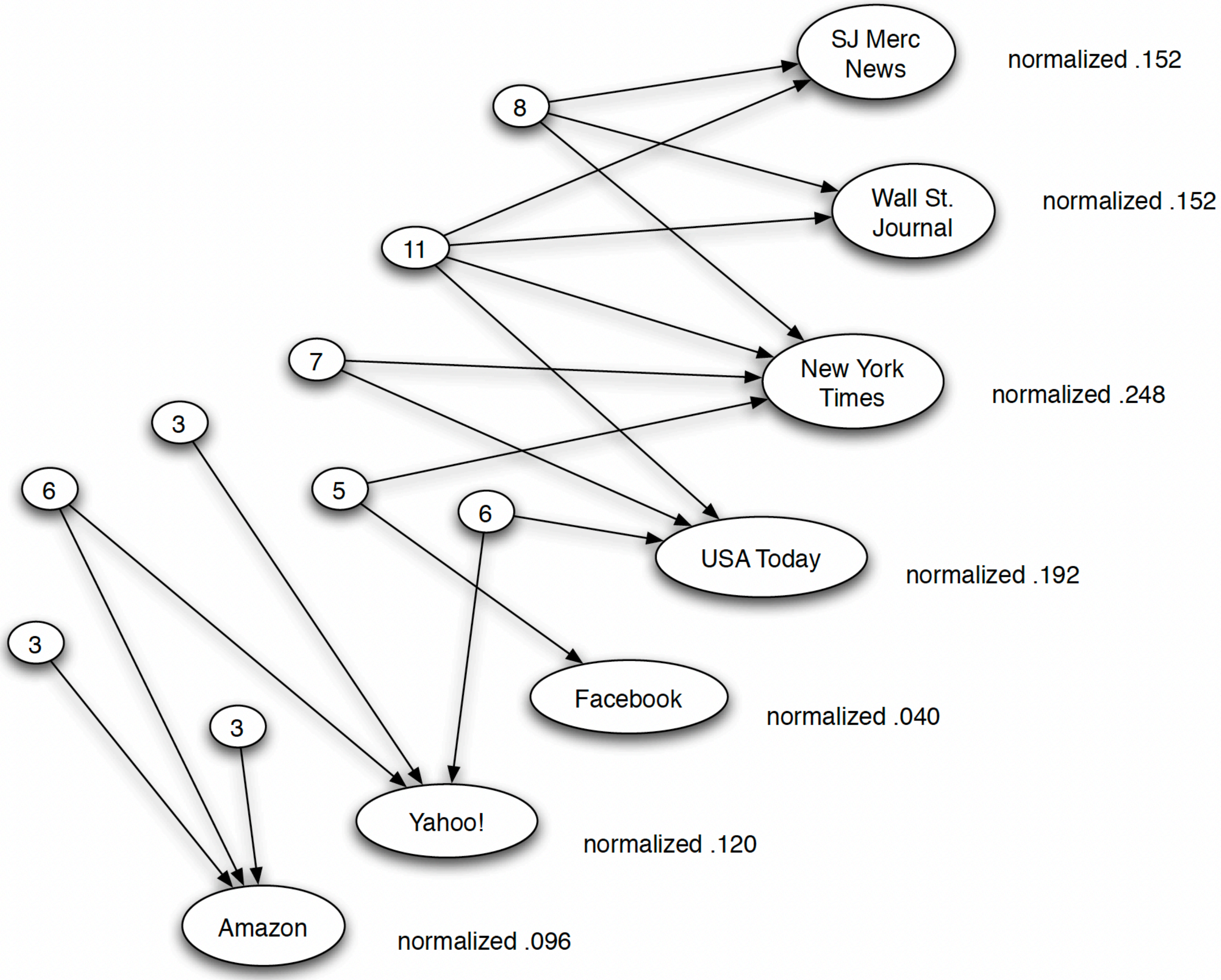
- If we think that high-scoring list pages (hubs) have a better sense of where good results are:
  - we can **weight links from hubs more heavily**
- Recalculate
  - the authority score is not just the number of in-links, it is the sum of the hub values of the in-link pages
- **Why stop here?** we can refine values at both sides



UNIVERSITÀ DI TORINO

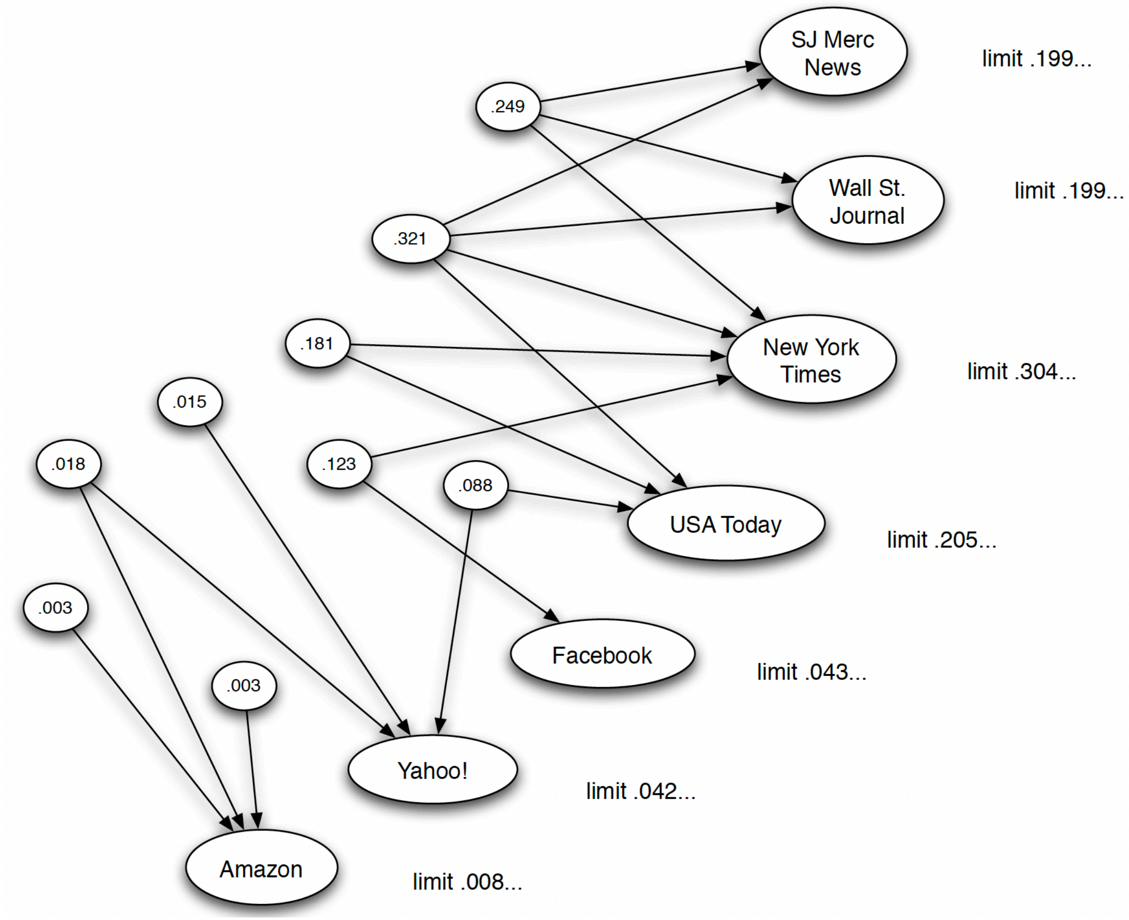
Hubs and Authorities

- Each page p has an **authority score** $auth(p)$ and **hub score** $hub(p)$ initialized to 1
- **Authority Update Rule**: for each page p, update $auth(p)$ to be the sum of the hub scores of all pages that point to it.
- **Hub Update Rule**: for each page p, update $hub(p)$ to be the sum of the authority scores of all pages that it points to.
- At step k, we performed k hub-authority updates. Each step involves:
  - First apply the **Authority Update Rule** to the current set of scores.
  - Then apply the **Hub Update Rule** to the resulting set of scores.
- Getting high values: **normalize hub and authority values** by divinding down the scores byt the sum of the scores of authoritys/hubs.


SJ Merc News
normalized .152
Wall St. Journal
normalized .152
New York Times
normalized .248
USA Today
normalized .192
Facebook
normalized .040
Yahoo!
normalized .120
Amazon
normalized .096
8
11
7
3
6
5
6
3
3

Stabilization

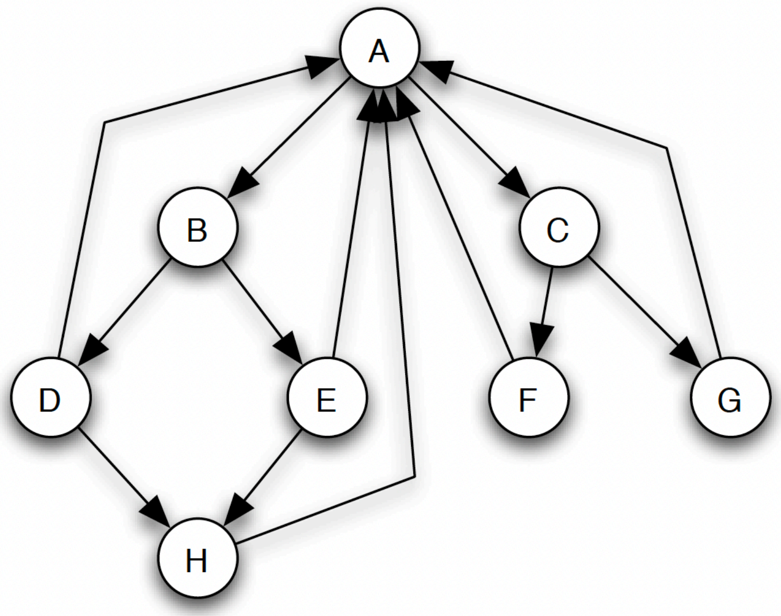
- It can be proved that normalized values **converge** when $k \to \infty$
    - **(optional proof)** see Section 14.6 Advanced Material: Spectral Analysis, Random Walks, and Web Search
- **Initial values are not important**
- Limiting values for hubs and authorities are **properties of the links structure**.
- Different form of game theoretical concept of equilibrium
    - **authority score** is proportional to the hub scores of the pages that point to you
    - **hub score** is proportional to the authority scores of the pages you point to.



Page Rank

- **Endorsement** viewed as passing directly from one important node to another
  - in other words, a page is important if it is cited by other important pages
  - endorsements received by in-links and passed across outgoing links
- Basic definition:
  - **Initilization**: Init all the pages p to a $PR(p) = \frac{1}{n}$, where n is the number of pages
  - $\boldsymbol{k^{th}}$ **step**: we perform a sequence of k updates to the PageRank values, using the following rule for each update:
  - **Basic PageRank Update Rule**:
    - each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to.
    - if a page has no out-going links, it passes all its current PageRank to itself.
    - each page updates its new PageRank to be the sum of the shares it receives.

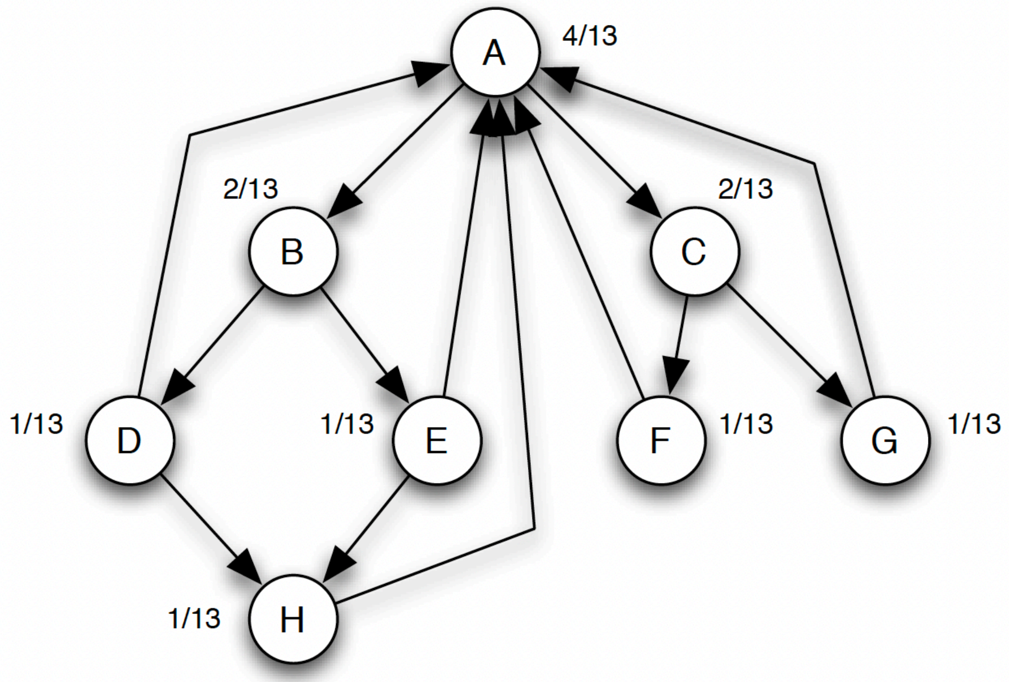
Example



Step	A	B	C	D	E	F	G	H
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

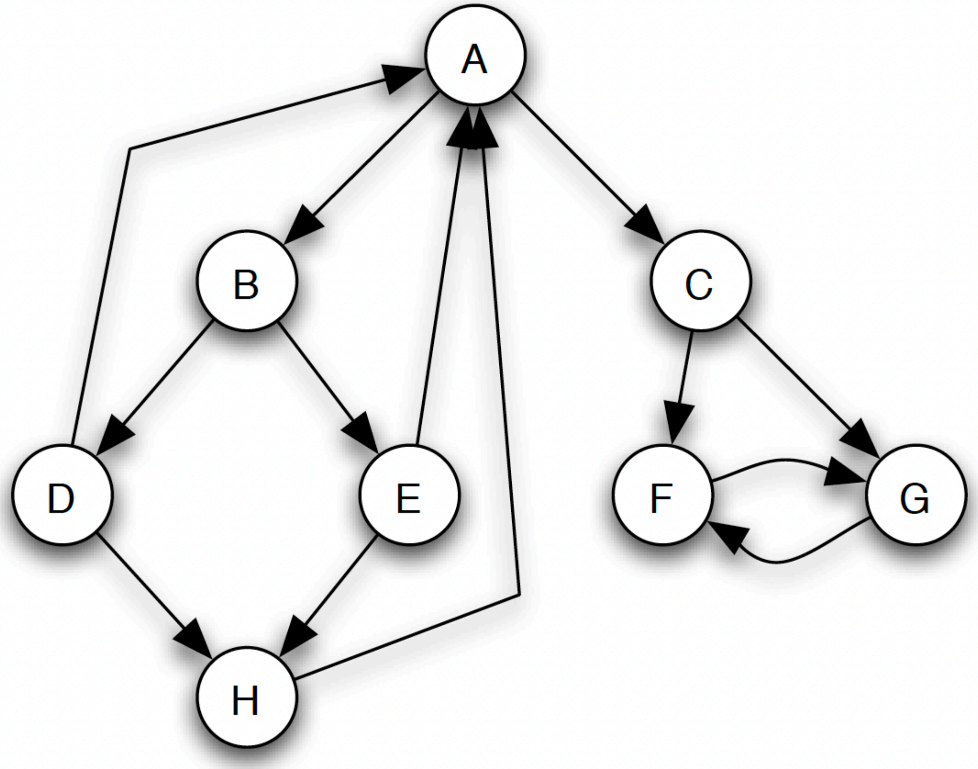
PageRank and stabilization

- PR values of all the nodes **converge** when $k \to \infty$
  - **but for some degenerate cases**.
- **Equilibrium**: if we apply our PR update rule, then our limiting **values do not change**.
- One can prove that if the network is **strongly connected**, then there is a **single set of equilibrium values** (when they exist)
  - **(optional proof)**: see Section 14.6 Advanced Material: Spectral Analysis, Random Walks, and Web Search



Scaling the definition of PageRank

- Which are these degenerate cases?
- **Problem:** in some networks some nodes receive all the PR values of the the network
  - **example on the right**: we converge to PR values of $\frac{1}{2}$ for each of F and G, and 0 for all other nodes
- Why?
  - We do not have path back to some other nodes



UNIVERSITÀ DI TORINO

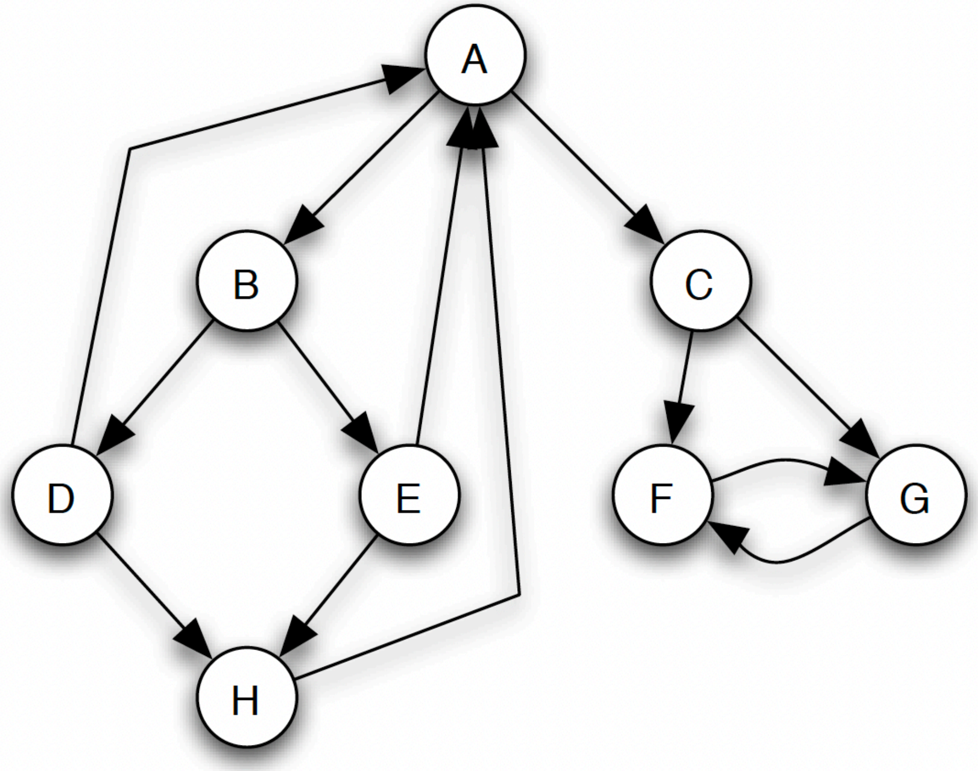
- **Solution:** let's force this "fluid" to stream back to other nodes "sometimes":
- Select a **scaling factor** (aka **damping factor**) $s : s \in [0, 1]$
- Get a portion s of PR values from in-links and then add $\frac{(1-s)}{n}$
    - This means that the total PR in the network has shrunk from 1 to s. We divide the residual $1 - s$ units of PR equally over all nodes, giving $\frac{(1-s)}{n}$ to each.
    - Called *Scaled PageRank Update Rule*
- Now we have convergence for $k \to \infty$
- **Note**: typically $s \in [0.8, 0.9]$

Random walks: An equivalent definition of PageRank

- Randomly browsing a network of Web pages, navigating each page with equal probability following links.
    - start by choosing a page at random
    - pick a random out-going link from their current page, and follow it to where it leads
- Follow links for a sequence of length k
- **Claim:** the probability of being at a page x after k steps of this random walk is precisely the PR of x after k applications of the Basic PageRank 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 F and G has a natural interpretation: when the surfer reaches F or G, 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.



Practical implications (also beyond the Web)

Modern Web search

- Google today use PR as one of the many features of their ranking framework
  - original Page&Brin's paper The Anatomy of a Large-Scale Hypertextual Web Search Engine.
  - e.g., 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!)

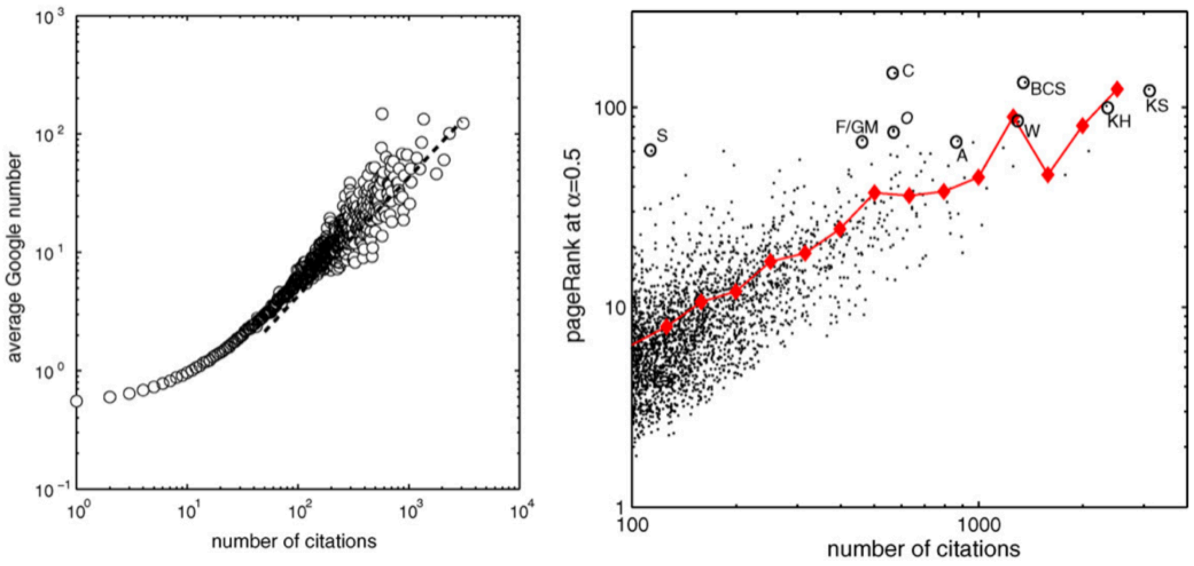
SEO vs Google

- 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.
  - you should always expect the world to react to what you do.

Applications beyond the Web

- Citation Analysis
- Link Analysis of U.S. Supreme Court Citations

Page Rank and citation analysis



- **Reference**: 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 a limited number of citations but highly influential anyhow

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

Introduction to spectral analysis

- **Pre-requisites**:
  - linear algebra
  - vector and matrix multiplication
- Eigenvalues/eigenvectors calculation to study the structure of networks
  - **spectral analysis**
- Limiting values in the **Hubs/Authorities** and **PageRank** algorithms are coordinates in the eigenvectors for given eigenvalues in matrices derived from our graphs.

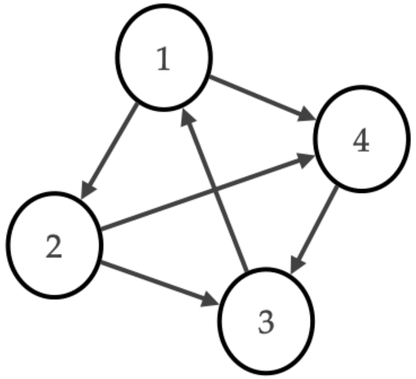
Definitions

Adjacency Matrix M

$1 \dots n$ nodes

$M : n \times n$

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



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

The adjacency matrix is not necessarily efficient for computational representations

- but conceptually very useful
- for practical use, consider adjacency lists or edge lists instead

Update rules of Hubs and Authorities

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

Hub Update Rule

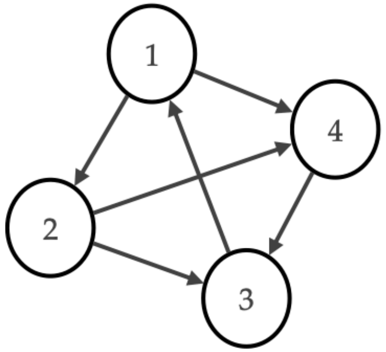
$$h_i \leftarrow \sum_{j=1}^{n} M_{ij}a_j = M_{i1}a_1 + M_{i1}a_2 + \ldots + M_{in}a_n$$

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

Authority Update Rule

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

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



$$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\mathbf{M}} \cdot \underbrace{\begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}}_{\mathbf{a}} = \underbrace{\begin{bmatrix} 3 \\ 4 \\ 1 \\ 2 \end{bmatrix}}_{\mathbf{h}}$$

UNIVERSITÀ DI TORINO

Understanding the k-step hub-authority computation

initilization: $\mathbf{h}^{<\mathbf{0}>} = \underbrace{(1, 1, \ldots, 1)}_{n}$

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

Understanding the k-step hub-authority computation

First step

$$\mathbf{a}^{<1>} = \mathbf{M^T h}^{<0>}$$
$$\mathbf{h}^{<1>} = \mathbf{Ma}^{<1>} = \mathbf{MM^T h}^{<0>}$$

Second step

$$\mathbf{a}^{<2>} = \mathbf{M^T h}^{<1>} = (\mathbf{M^T M})\mathbf{M^T h}^{<0>}$$
$$\mathbf{h}^{<2>} = \mathbf{Ma}^{<2>} = (\mathbf{MM^T})(\mathbf{MM^T})\mathbf{h}^{<0>} = (\mathbf{MM^T})^2\mathbf{h}^{<0>}$$

$\boldsymbol{k^{th}}$ **step**

$$\mathbf{a}^{<k>} = (\mathbf{M^T M})^{\mathbf{k-1}}\mathbf{M^T h}^{<0>}$$
$$\mathbf{h}^{<k>} = (\mathbf{MM^T})^{\mathbf{k}}\mathbf{h}^{<0>}$$

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

Multiplications and Eigenvectors (optional)

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

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

Focus on the hub vectors:

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

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

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

Eigenvectors and square matrices (optional)

- Observe that $\mathbf{MM^T}$ is a symmetric matrix
- Fact 1: "Any symmetric matrix $n \times n$ has a set of n eigenvectors $\mathbf{z_1}, \cdots, \mathbf{z_n}$ 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 $\mathbf{MM^T}$ we can find:
  - n mutual orthogonal eigenvectors: $\mathbf{z_1}, \cdots, \mathbf{z_n}$ (the spectrum of $\mathbf{MM^T}$)
  - n corresponding eigenvalues: $c_1, \cdots, c_n$
- Let's sort eigenvectors s.t. corresponding eigenvalues: $c_1 \geq c_2 \geq \cdots \geq c_n$
- Assume (for now): $c_1 > c_2$

UNIVERSITÀ DI TORINO

Eigenvectors and square matrices (optional)

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

$$\begin{aligned}(\mathbf{MM^T})\mathbf{x} &= (\mathbf{MM^T})(p_1\mathbf{z_1} + p_2\mathbf{z_2} + \ldots + p_n\mathbf{z_n}) \\ &= p_1(\mathbf{MM^T})\mathbf{z_1} + p_2(\mathbf{MM^T})\mathbf{z_2} + \ldots + p_n(\mathbf{MM^T})\mathbf{z_n} \\ &= p_1c_1\mathbf{z_1} + p_2c_2\mathbf{z_2} + \ldots + p_nc_n\mathbf{z_n}\end{aligned}$$

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

$$(\mathbf{MM^T})^{\mathbf{k}}\mathbf{x} = p_1c_1^k\mathbf{z_1} + p_2c_2^k\mathbf{z_2} + \ldots + p_nc_n^k\mathbf{z_n}$$

Convergence of the Hub-Authority computation (optional)

vector of hub scores at step k:

$$\mathbf{h}^{<\mathbf{k}>} = (\mathbf{M}\mathbf{M}^{\mathbf{T}})^k \cdot \mathbf{h}^{<\mathbf{0}>}$$
$$\mathbf{h}^{<\mathbf{0}>} = q_1\mathbf{z_1} + q_2\mathbf{z_2} + \ldots + q_n\mathbf{z_n}$$
$$\mathbf{h}^{<\mathbf{k}>} = c_1^k q_1\mathbf{z_1} + c_2^k q_2\mathbf{z_2} + \ldots + c_n^k q_n\mathbf{z_n}$$

Let's divide both sides by c_1^k:

$$\frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = \frac{c_1^k q_1\mathbf{z_1}}{c_1^k} + \frac{c_2^k q_2\mathbf{z_2}}{c_1^k} + \ldots + \frac{c_n^k q_n\mathbf{z_n}}{c_1^k}$$

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

Then:

$$lim_{k\to\infty} \frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = q_1\mathbf{z_1}$$

Wrapping up (optional)

- (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:

$$\mathbf{x} = p_1\mathbf{z_1} + p_2\mathbf{z_2} + \ldots + p_n\mathbf{z_n} \Rightarrow (\mathbf{MM^T})^\mathbf{k}\mathbf{x} = c_1^k p_1\mathbf{z_1} + c_2^k p_2\mathbf{z_2} + \ldots + c_n^k p_n\mathbf{z_n}$$

$$lim_{k\to\infty}\frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = p_1\mathbf{z_1}$$

- (ii) coefficient $p_!$ (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 $\mathbf{z_1}$

Wrapping up (optional)

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

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

$$\frac{\mathbf{h}^{<\mathbf{k}>}}{c_1^k} = \frac{c_1^k q_1 \mathbf{z_1} + c_2^k q_2 \mathbf{z_2} + \ldots + c_l^k q_l \mathbf{z_l}}{c_1^k} + \frac{c_{l+1} n^k q_{l+1} \mathbf{z_{l+1}} + \ldots\ldots + c_n^k q_n \mathbf{z_n}}{c_1^k}$$

$$= q_1 \mathbf{z_1} + q_2 \mathbf{z_2} + \ldots + q_l \mathbf{z_l} + 0$$

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

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

Spectral Analysis of Page Rank

At step 0 (init):

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

At step k:

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

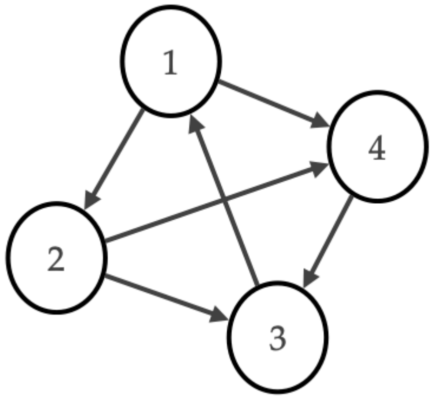
$\forall i : r_i = s \cdot \sum_{j=1}^{n} M_{ji} \frac{r_j}{k_j^{\text{out}}} + (1 - s) \cdot \frac{1}{n}$ (scaled PR update rule)

Using matrix notation

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

nodes: $1, \ldots, n$

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



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

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

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

If i has no outgoing links, then we define $N_{ii} = 1$

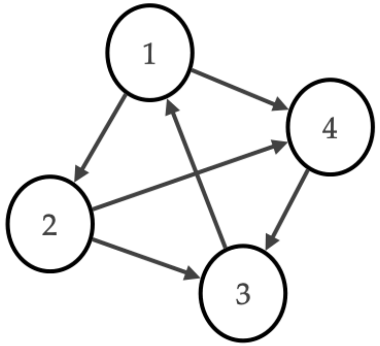
Update rule (basic and scaled)

1.Basic update rule:

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

2.Scaled update rule (factor s):

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



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

3.Application of scaled update rule:

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

Repeated improvement (optional)

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

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

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

$\mathbf{r}^{<*>}$ should be an eigenvector of $\widetilde{\mathbf{N}}^{\mathbf{T}}$ with corresponding eigenvalue of 1

Problem: $\widetilde{\mathbf{N}}^{\mathbf{T}}$ 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 (optional)

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

Perron's theorem

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

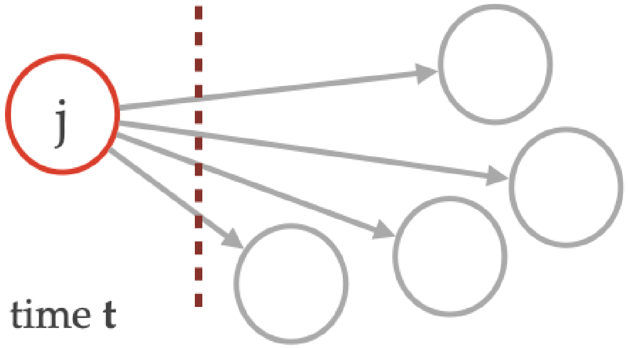
i) $\mathbf{P}$ has an eigenvalue $c > 0$ s.t. $c > c' \; \forall c'$ (with c' another eigenvalue)

ii) Exists an eigenvector $\mathbf{y}$ with real positive values corresponding to c, and $\mathbf{y}$ is unique (up to a multiplication constant)

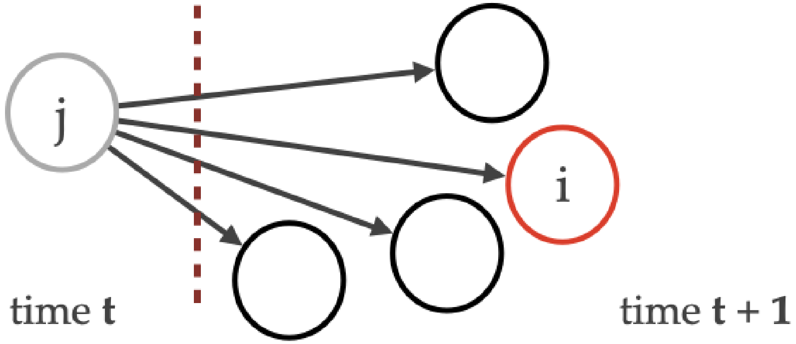
iii) if $c = 1$, then for any starting vector $\mathbf{x} \neq \mathbf{0}$ with non negative coordinates, the sequence of vectors $p^k\mathbf{x}$ converges to a vector in the direction of $\mathbf{y}$ $(k \to \infty)$

UNIVERSITÀ DI TORINO

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 $t+1$?

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

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:

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

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

Reading material

[ns2] **Chapter 14 (14.1-14.5)** Link Analysis and Web Search

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

Q&A