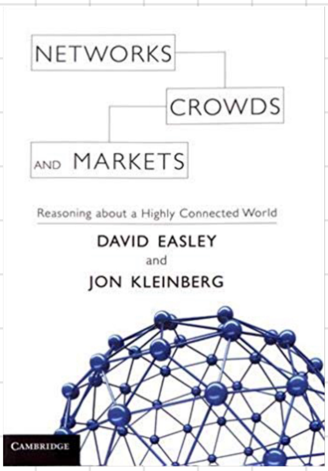
Lecture 12

Network Science

Link Analysis and Web Search

Today's topics

- Searching the Web
- Link Analysis
  - HITS: Hubs & Authorities
  - Page Rank
- Modern Web Search
- Link Analysis beyond the Web



Chapter 14
"Link Analysis and Web Search"

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,
...,
in response to keywords
based queries

- list of keywords "inexpressive"
  synonymy
  polysemy

- "diversity": given a topic
  we find pages written by
  many kind of "authors"

- pages are dynamic and always changing
- "news search" features
- "scarcity" vs "abundance"

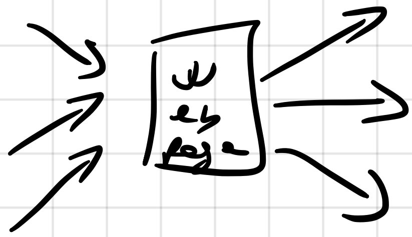
Filters

- what is "important"?

Can the structure of the web dominated by links, help us to find such "filters"

First attempt:
count "words" on documents

HITS: Link Analysis using Hubs and Authorities



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

"implicit endorsement" from the community

List Finding

query: "newspapers"

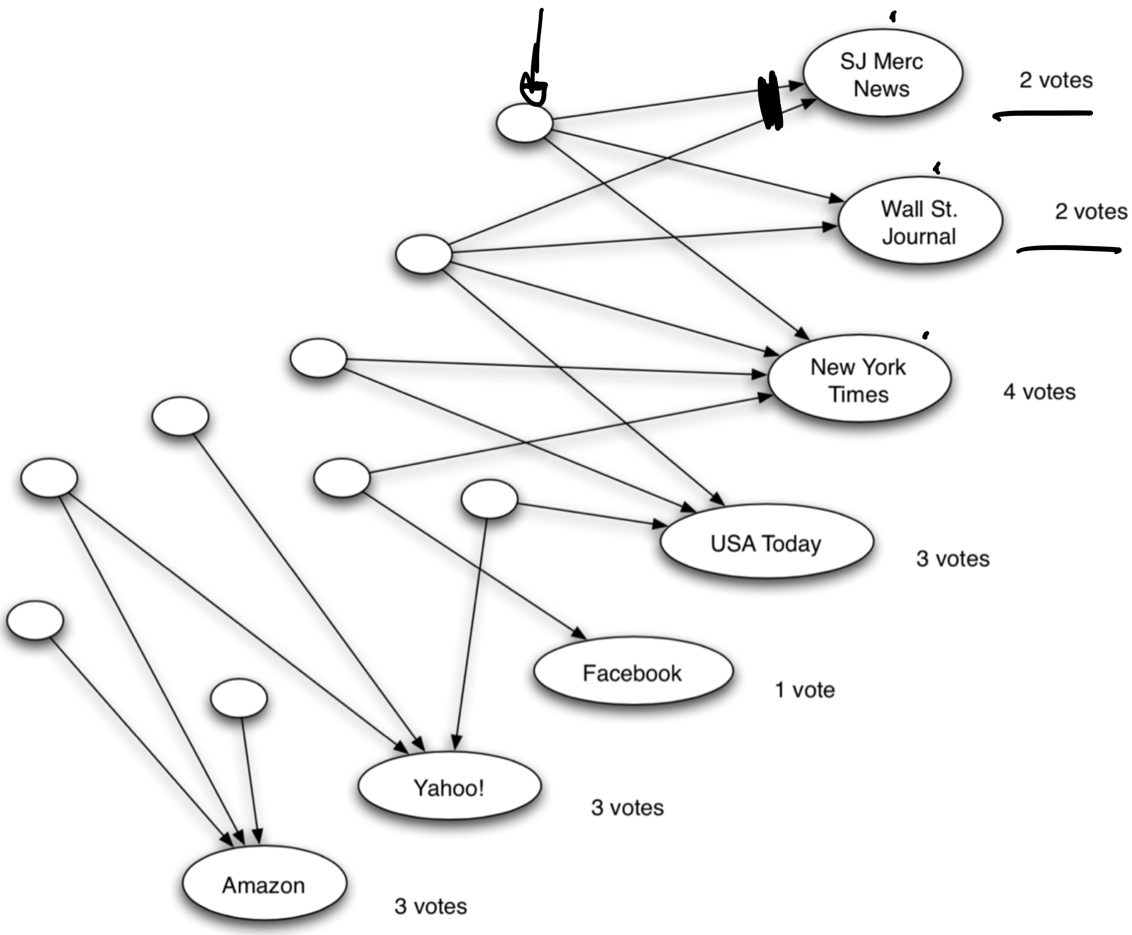


Figure 14.1: Counting in-links to pages for the query "newspapers."

We have "intuitively correct" results

- "Lists": pages that provide many different out-links to other pages.

a page has a "list value" <=> the sum of in-links received by all pages they link to

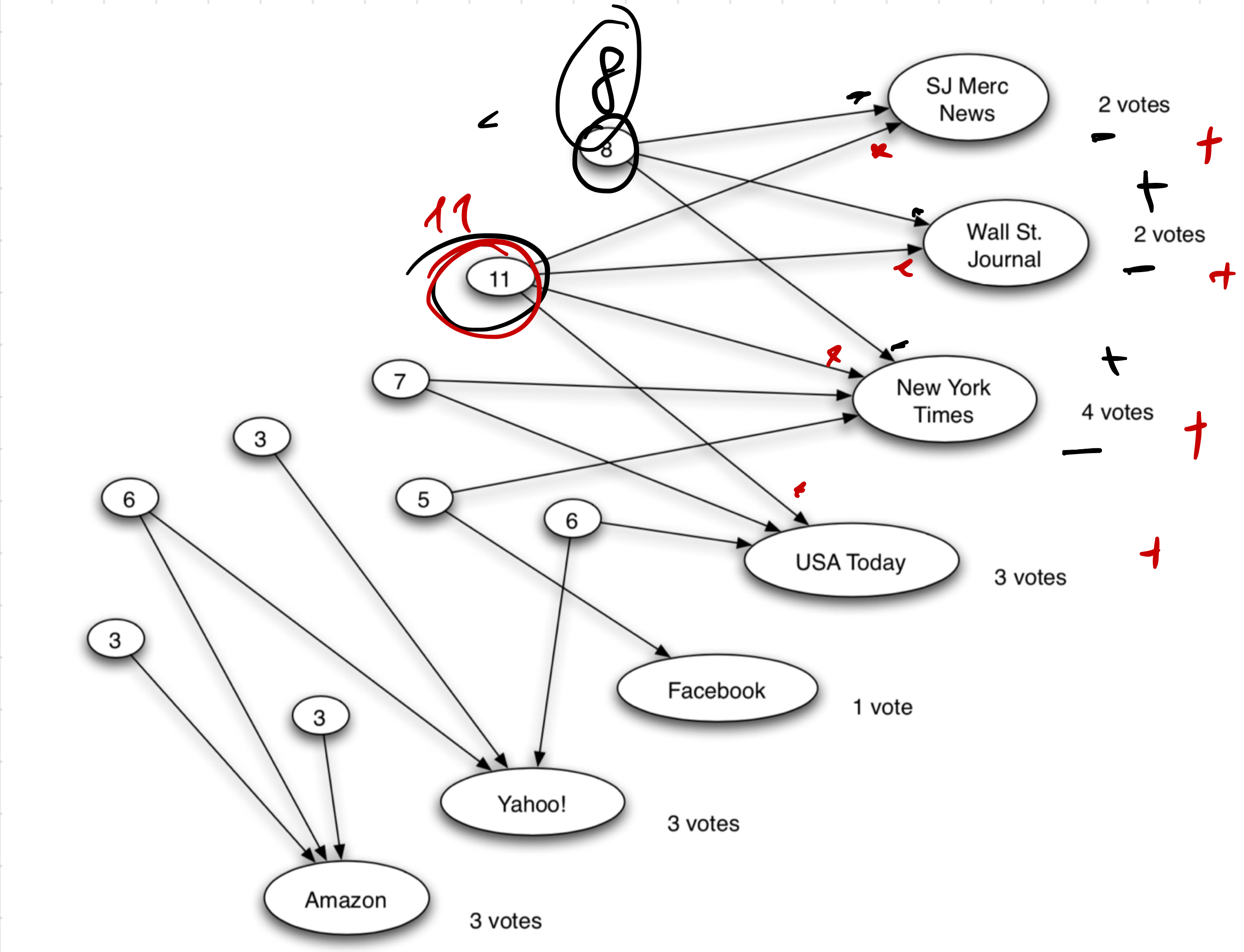


Figure 14.2: Finding good lists for the query "newspapers": each page's value as a list is written as a number inside it.

Assumption: pages behaving as lists have a better sense for where the good results are.
("Authorities" are competitors)

The Principle of Repeated Improvement

we want to weight lists' links more heavily.

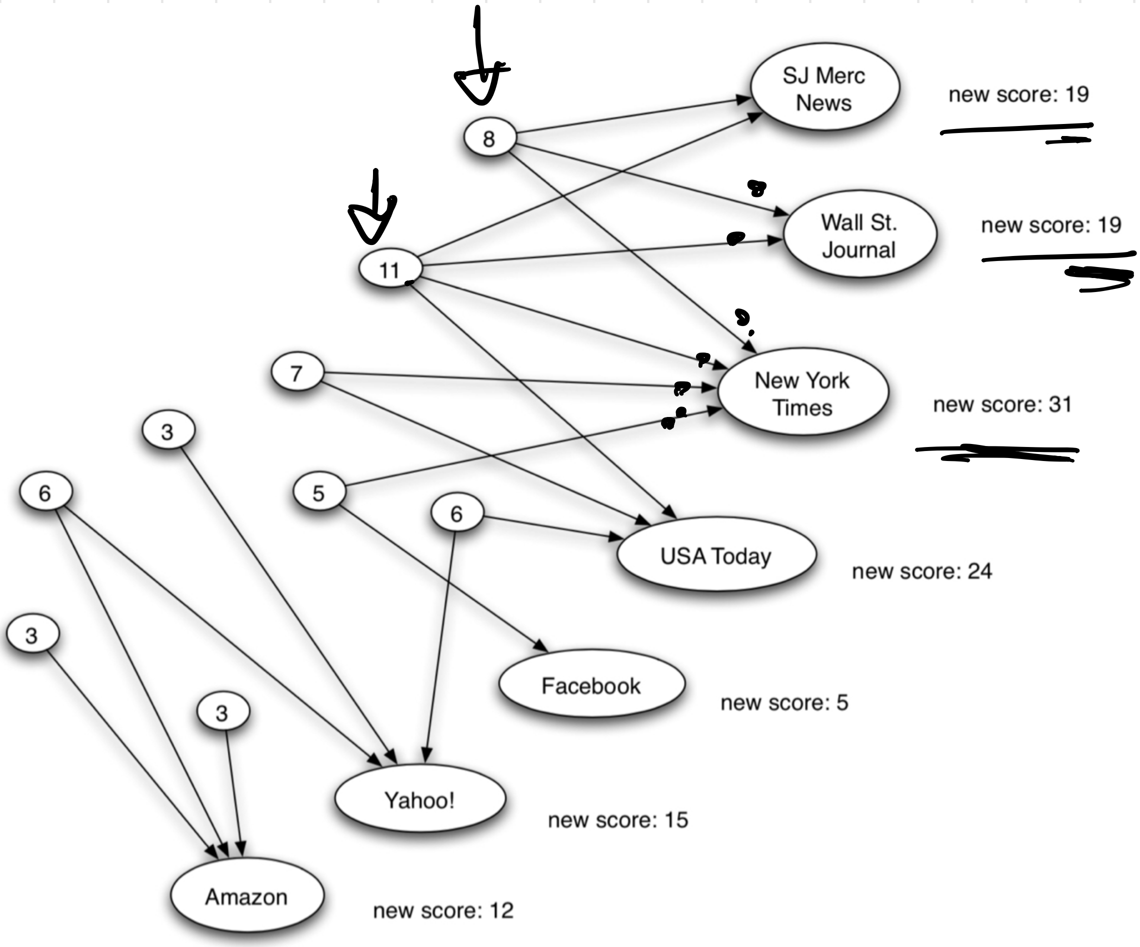


Figure 14.3: Re-weighting votes for the query "newspapers": each of the labeled page's new score is equal to the sum of the values of all lists that point to it.

why stop here?

we can refine values at both sides

Hubs and Authorities

pages that we are looking for: "authorities"

pages with high "list value" "hubs"

$\forall p$: hub(p), auth(p)

<u>initialize</u>. $\forall p$: hub(p) = auth(p) = 1

Authority Update Rule

$$\forall p: \quad auth(p) = \sum_{i=1}^{n} hub(i)$$

n: # pages connected <u>to</u> p

Hub Update Rule

$$\forall p: \quad hub(p) = \sum_{i=1}^{n} auth(i)$$

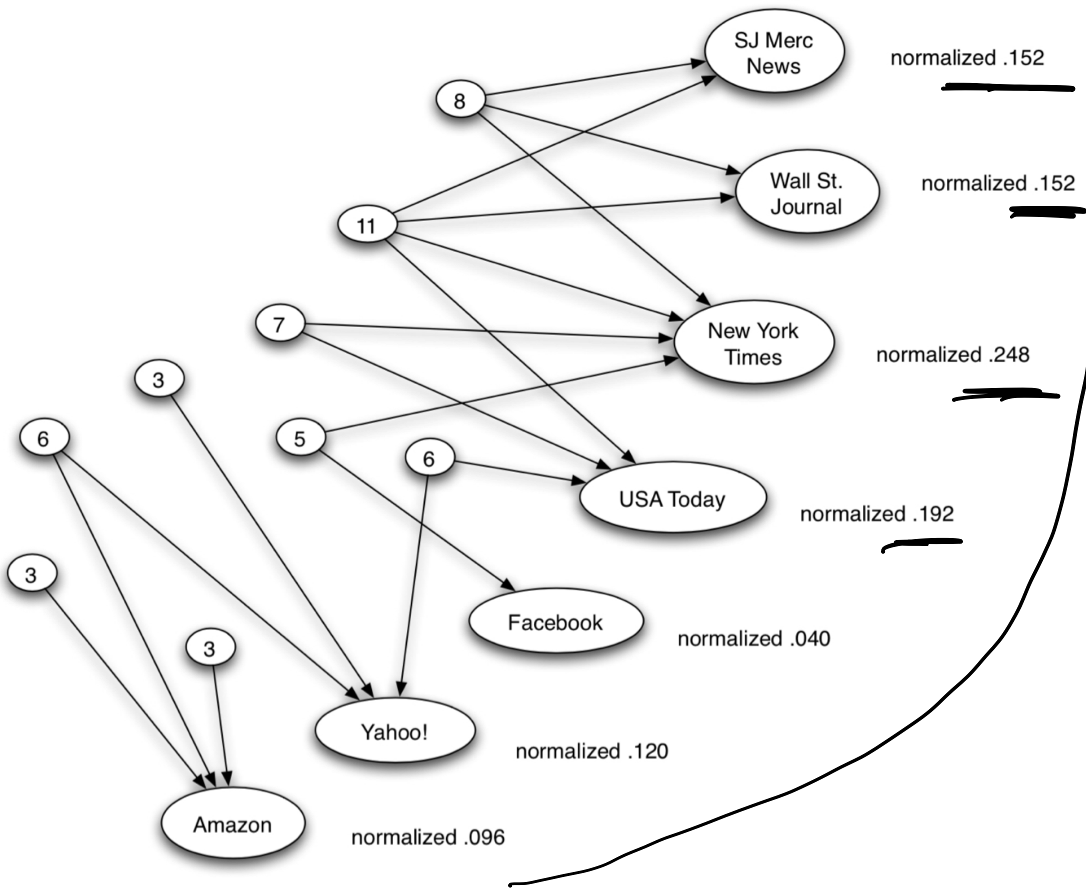
n: # pages p connects to

(p, i) are edges

lets decide K as the total number of steps

1. $\forall p$: $hub(p) = auth(p) = 1$
2. for K steps
   - 2a. apply auth update rule
   - 2b. apply hub update rule
3. normalize values

$$auth(p) = \frac{auth(p)}{\sum auth(i)}, \quad hub(p) \ldots$$



total = 1

Figure 14.4: Re-weighting votes after normalizing for the query "newspapers."

normalized values converge when

$K \to \infty$

STABILIZATION: init. values are not important stabilization : limiting values for hubs and authorities are properties of the links structure

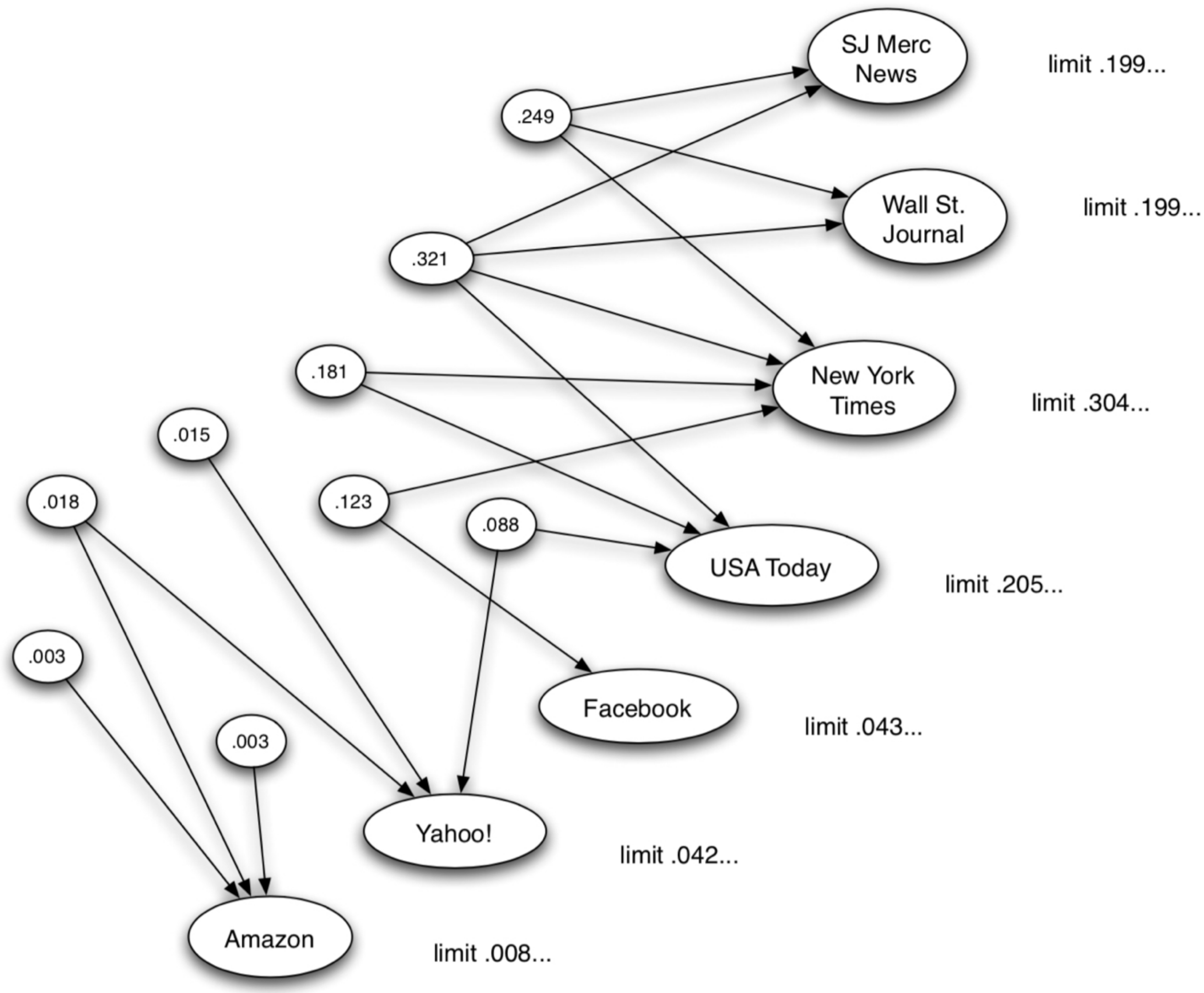


Figure 14.5: Limiting hub and authority values for the query "newspapers."

"Equilibrium"

Page Rank

"Endorsement" viewed as passing directly from one "important" node to another.

Endorsments are received by in-links and passed across out-going links

basic definition (number of steps)

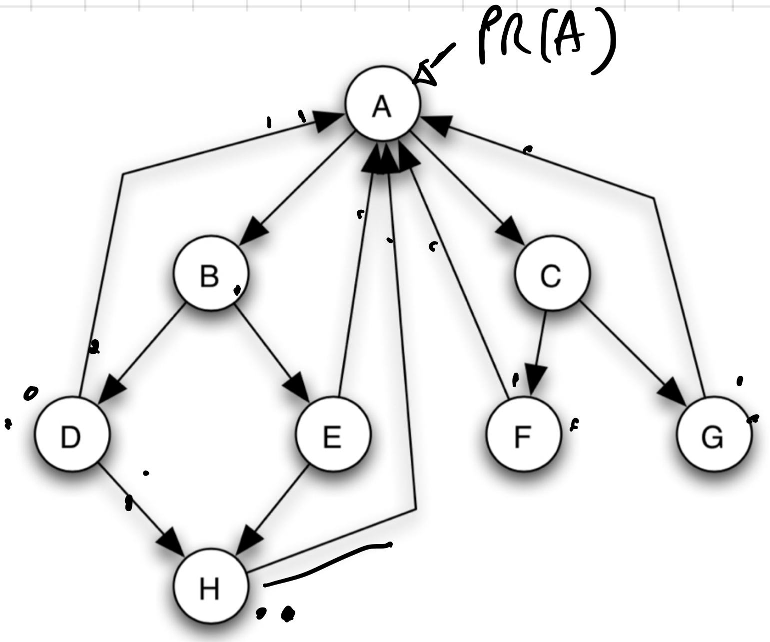
1. $\forall p: \quad PR(p) = \frac{1}{N}$; N: # pages
2. for K steps
   2a. apply "basic PR update rule"

$$PR(p) = \sum_{i=1}^{n} \frac{PR(i)}{out(i)}$$

n: # of pages connected to p

(i, p) are d. edges

out(i): # of outgoing links of page i.



8 pages

Figure 14.6: A collection of eight pages: A has the largest PageRank, followed by B and C (which collect endorsements from A).

$0 \quad \frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{8} \quad \text{—} \quad \text{-} \quad \text{—} \quad \text{—}$

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

$$PR(A) = \overset{D}{\frac{1}{2\cdot 8}} + \overset{E}{\frac{1}{2\cdot 8}} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} =$$

$$= \frac{1+1+2+2+2}{16} = \frac{8}{16} = \frac{1}{2}$$

Repeat the PR update rule fo K steps

PageRank at Equilibrium

PR values of all nodes converge when $k \to \infty$ (but for some "degenerate cases")

<u>Equilibrium:</u> if we apply our PR update rule, then our limiting values do not change

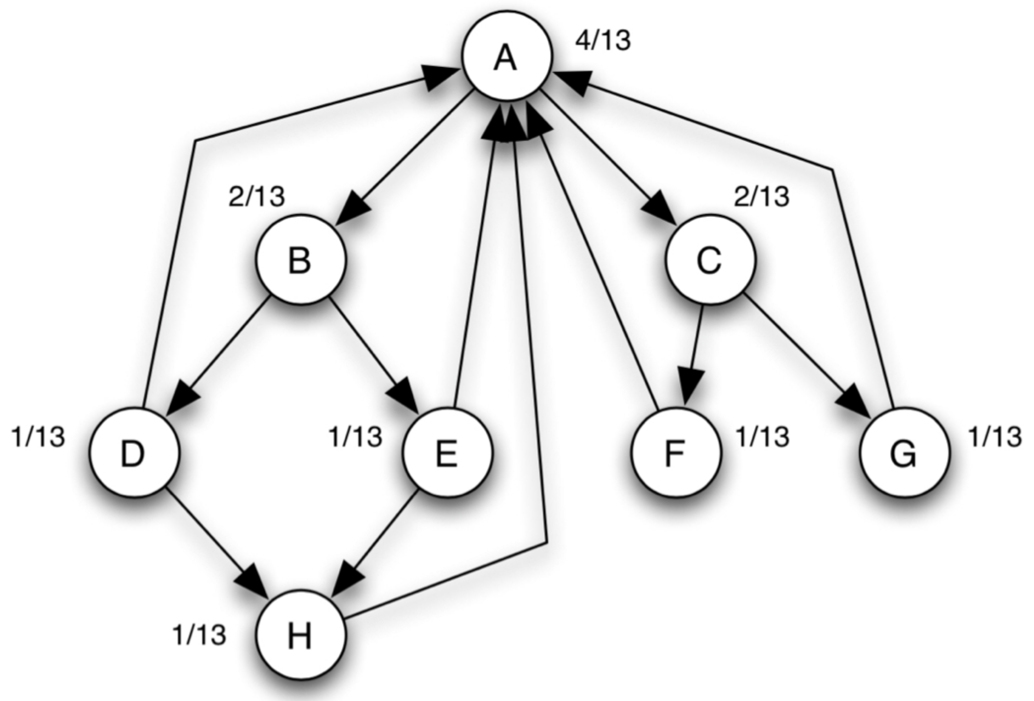


Figure 14.7: Equilibrium PageRank values for the network of eight Web pages from Figure 14.6.

Scaling the definition of PageRank

"degenerate cases"

the problem: in some networks some nodes receive all the PR values of the network

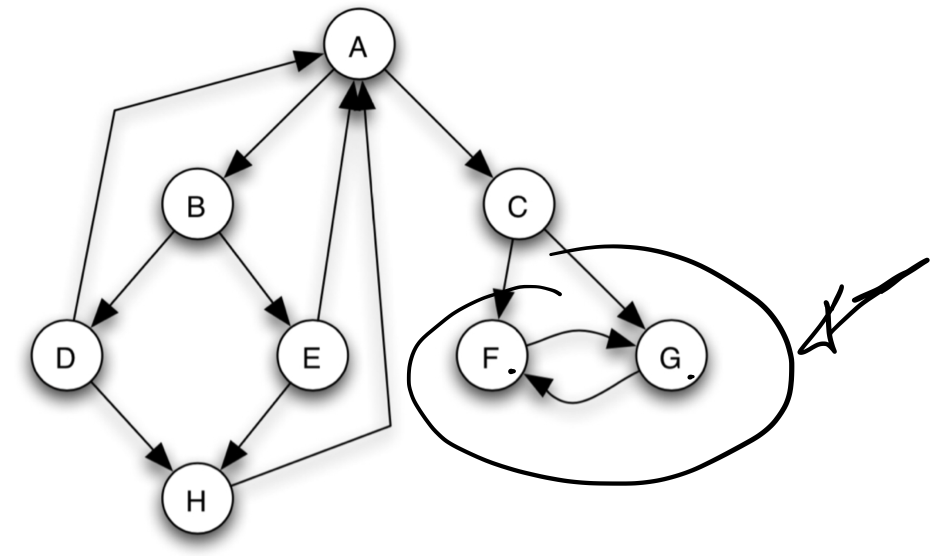
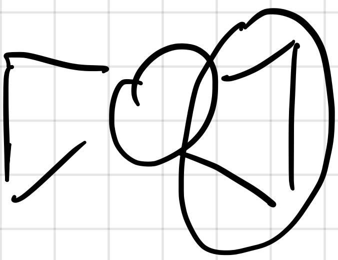


Figure 14.8: The same collection of eight pages, but F and G have changed their links to point to each other instead of to A. Without a smoothing effect, all the PageRank would go to F and G.

Repeating PR update rule

$$PR(F) = \frac{1}{2} \qquad PR(G) = \frac{1}{2}$$

$$\forall p \; (p \notin \{F, G\}) \quad PR(p) = 0$$



I can have degenerate cases in the OUT COMPONENT of the [illegible] the problem: we do not have path back to some other nodes

<u>solution</u>

lets force this "fluid" to stream back to other nodes "sometimes"

select a "scaling factor" (damping factor") "s"

$s \in [0, 1]$

<u>Scaled PR Update Rule (SPRU)</u>

$$PR(p) = s \sum_{i=1}^{n} \frac{PR(i)}{out(i)} + (1-s)\frac{1}{N}$$

the limits of SPRU rule:

$k \to \infty$: all the PR values are unique

values depend on "s" ($s \in [0.8, 0.9]$)

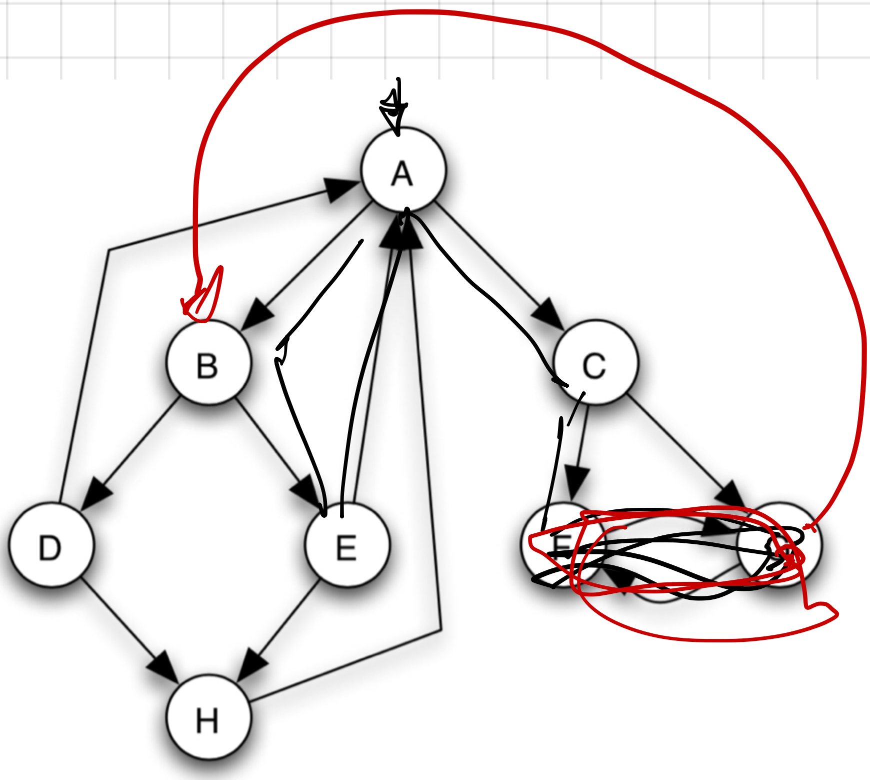
Random Walks

randomly clicking from one page to another picking each page with equal probability.

Follow links for a sequence of K links

<u>claim</u>: the probability of being at page X after K steps is the application of the basic PR update rule

<u>additional intuition</u>: 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.



the "leakage" of F and G has a natural interpretation: the probability of converging to 1, and once it reaches F or G, then it is stuck "forever"

<u>SOLUTION</u>: with prob. δ:

I click on an hyperlink

with prob. $1-\delta$:

I <u>JUMP</u> to a randomly selected node.

Applications of Link Analysis to Modern Web Search

Google to-day doesn't use PR anymore

(paper: 2001)

Hilltop (an extension of HITS)

anchor texts

clicking behavior

SEO (Search Engine Optimization)

Company

SEO

reverse engineering of SE's ranking functions

SEs define new measures

perfect result are "moving targets"

It is a game-theoretic principle

Link Analysis: Beyond the Web

citation analysis

"impact factor" of journal
average number of citations
received by papers
published in that journal

"influential weights"
↳ "page rank"

Apply PR to scientific paper

"Finding Scientific Gems with Google Page Rank" (2007)

dataset: collection of scient. papers with their references

correlation

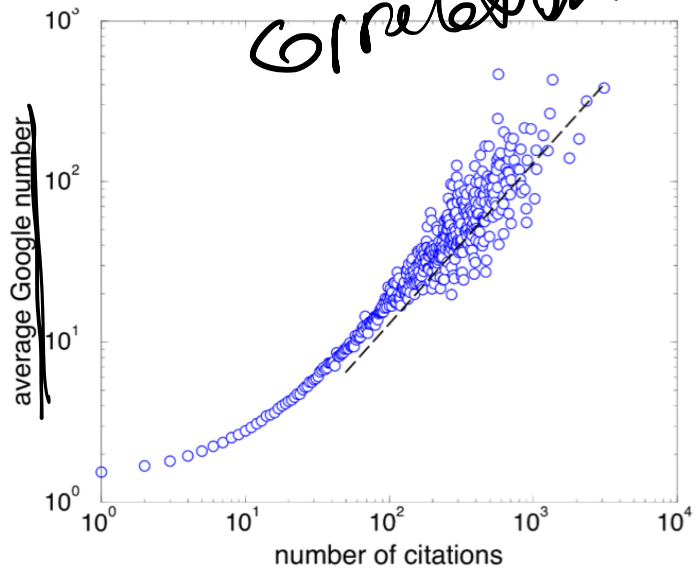


FIG. 2: Average Google number $\langle G(k)\rangle$ versus number of citations k. The dashed line of slope 1 is a guide for the eye.

correlation BUT

outliers are papers with "limited" number of citations but highly influential

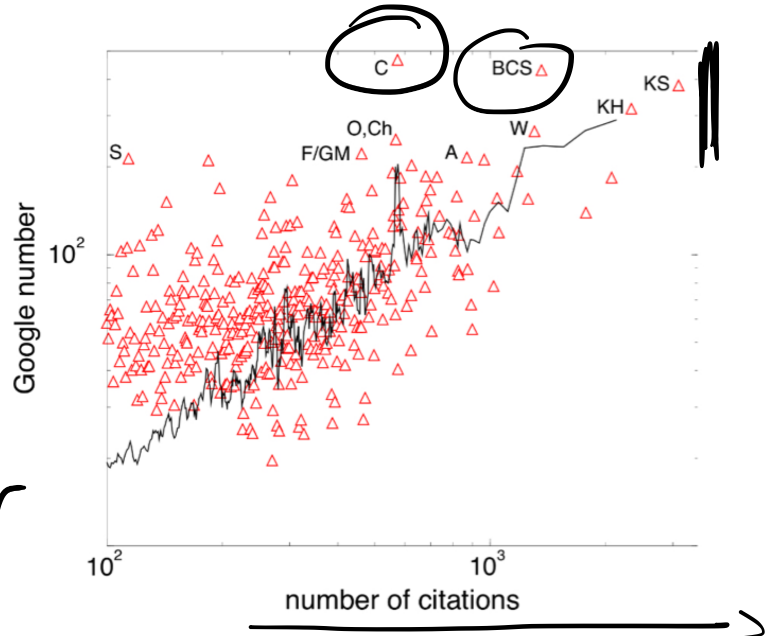


FIG. 3: Individual outlier publications. For each number of citations k, the publication with the highest Google number is plotted. The top-10 Google-ranked papers are identified by author(s) initials (see Table I). As a guide to the eye, the solid curve is a 5-point average of the data of $\langle G(k)\rangle$ versus k in Fig. 2.

TABLE I: The top 10 Google-ranked publications when $d = 0.5$

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, J. Schrieffer
3	3.81	1	3227	PR	140	A1133	1965	Self-Consistent Equations...	W. Kohn & L. J. Sham
4	3.17	2	2460	PR	136	B864	1964	Inhomogeneous Electron Gas	P. Hohenberg & 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	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 & 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

"Finding Sc. Gems"

Pros
↓
PR to help ~~to~~ discover "gems"

Cons
↓
indicators can change our behavior

Take Home Messages

- two methods to find important nodes:
  - HITS
  - Page Rank

  they are both iterative (k steps)
- (Modifications of) Both methods are widely used in modern web search engines and other domains
- Indicators change social behaviors: "perfect results" are moving targets
- Limiting PR and HITS values: for $k \to \infty$ same values are returned after each iteration
- Proof? next...
- Algorithmic complexity? HIGH, but numerical methods exist to solve the problem very efficiently (e.g. the power method)