TOR VERGATA
UNIVERSITY OF ROME

Searching the Web: The PageRank Algorithm

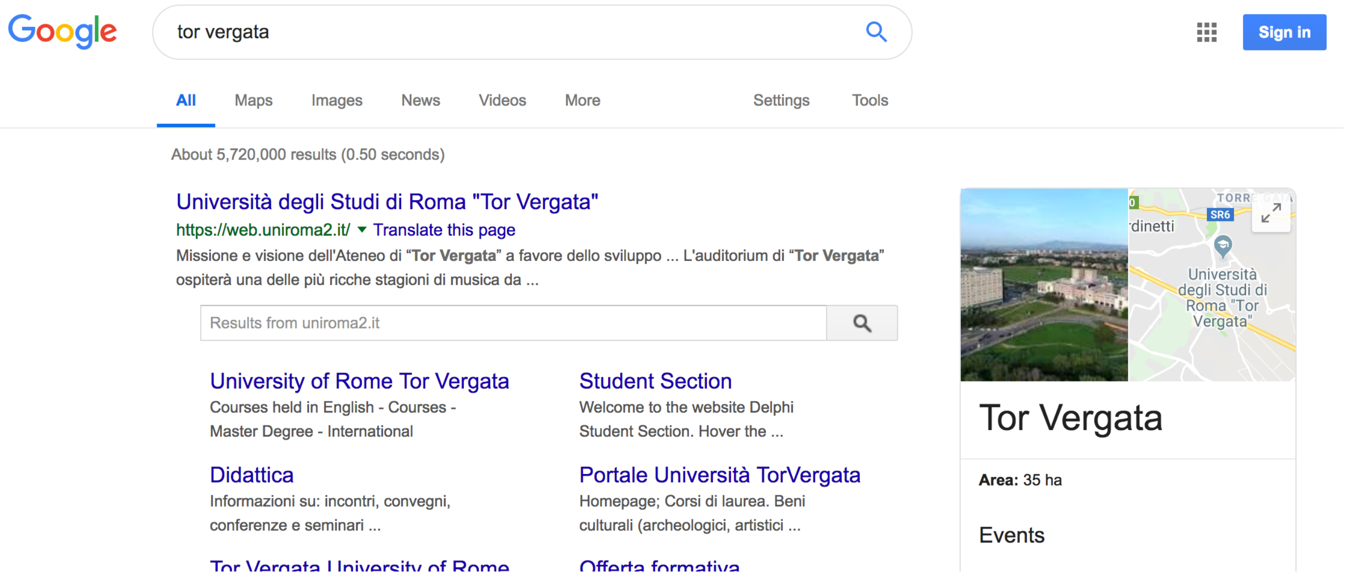
Algorithms, Data and Security
A.Y. 2023/24

Valeria Cardellini

Global Governance, 3rd year
Science and Technology Major

Searching the Web

- When we go to Google and type “tor vergata”, the first result we get is the home page of the university



About 5,720,000 results (0.50 seconds)

Università degli Studi di Roma "Tor Vergata"
https://web.uniroma2.it/ ▾ Translate this page
Missione e visione dell'Ateneo di “**Tor Vergata**” a favore dello sviluppo ... L'auditorium di “**Tor Vergata**” ospiterà una delle più ricche stagioni di musica da ...

University of Rome Tor Vergata
Courses held in English - Courses - Master Degree - International

Student Section
Welcome to the website Delphi Student Section. Hover the ...

Didattica
Informazioni su: incontri, convegni, conferenze e seminari ...

Portale Università TorVergata
Homepage; Corsi di laurea. Beni culturali (archeologici, artistici ...

Tor Vergata University of Rome

Offerta formativa

Tor Vergata

Area: 35 ha

Events

Searching the Web

- How did Google “know” that this was the best answer?
- Search engines determine how to rank pages using automated methods that look at the Web itself, without using some external source of knowledge
- There must be enough information intrinsic to the Web and its structure to figure this out

Search: a hard problem

- Search is a hard problem for computers to solve in any setting, not just on the Web
  - Information retrieval studied from the 1960s
- Keywords are a very limited way to express a complex information need
  - Synonymy: words that are similar or have a related meaning to another word
    - E.g., scallions and green onions
  - Polysemy: words may have more than one meaning
    - E.g., Jaguar

Search: a hard problem

- The Web introduces new kinds of problems in searching
- Dynamic and constantly-changing nature of Web content
- From a problem of *scarcity* to a problem of *abundance*
  - How to filter, from among a huge number of relevant Web pages, the few that are most important?

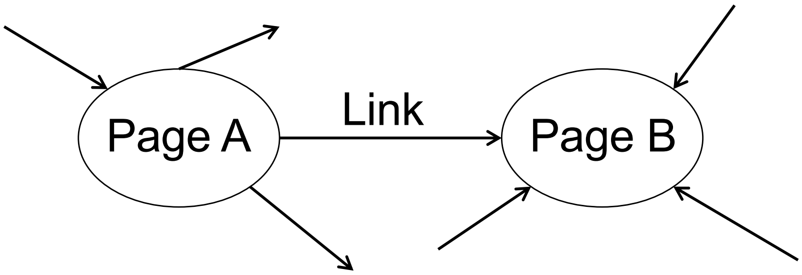
Search: tor vergata

- What are the clues that suggest the university home page as a good answer?
- The university home page does not use the word “tor vergata” more frequently than other pages (same as uniroma2)
- There is nothing on the page itself that makes it clearly stand out

Link analysis

- So how? **Link analysis**
- Links are essential to ranking
  - We can use them to assess the authority of a page on a topic
- Assumption: if a page has many links, then it receives a kind of “collective endorsement”
- Voting by in-links!

The Web as a link graph



- Nodes (Web pages) connected by directed edges (links)
- Assumption: a (hyper)link between Web pages denotes a conferral of authority (quality signal)

Why link analysis?

- The Web is not just a collection of documents: its links are important!
- Link from page A to page B may indicate:
  - A is related to B
  - A is recommending, citing, voting for, or endorsing B
- Links can be:
  - Referential: click here and get back home
  - Informational: click here to get more detail
- Links affect ranking of Web pages and thus have commercial value

Citation analysis

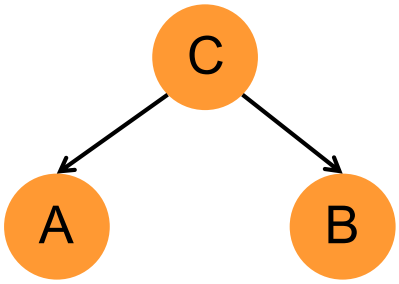
- The idea of using links is somehow “borrowed” by academic citation analysis
  - Since 1940s idea to evaluate the quality of an article based on citations
- Popular metric: impact factor (IF) of a journal

$$IF_y = \frac{Citations_{y-1} + Citations_{y-2}}{Publications_{y-1} + Publications_{y-2}}$$

  - The number of current year citations to articles appearing in the journal during previous two years divided by the number of articles published in the journal during previous two years
  - E.g., Nature’s IF in 2014

$$IF_{2014} = \frac{Citations_{2013} + Citations_{2012}}{Publications_{2013} + Publications_{2012}} = \frac{29753 + 41924}{860 + 869} = 41.456$$

Citation analysis: Co-citations



- Articles A and B are co-cited by article C, implying that A and B are related or associated
- The strength of co-citation between A and B is the number of times they are co-cited
- Co-citations are useful to infer topical relation between articles A and B without an explicit link

Citations vs. Web links

- Web links are a bit different from citations:
  - Many links are navigational
  - Many pages with high in-degree are Web portals
  - Not all links are endorsements
  - Company Web sites do not point to their competitors
- However, the general idea that "many citations = authority" has been borrowed in link analysis and in PageRank algorithm

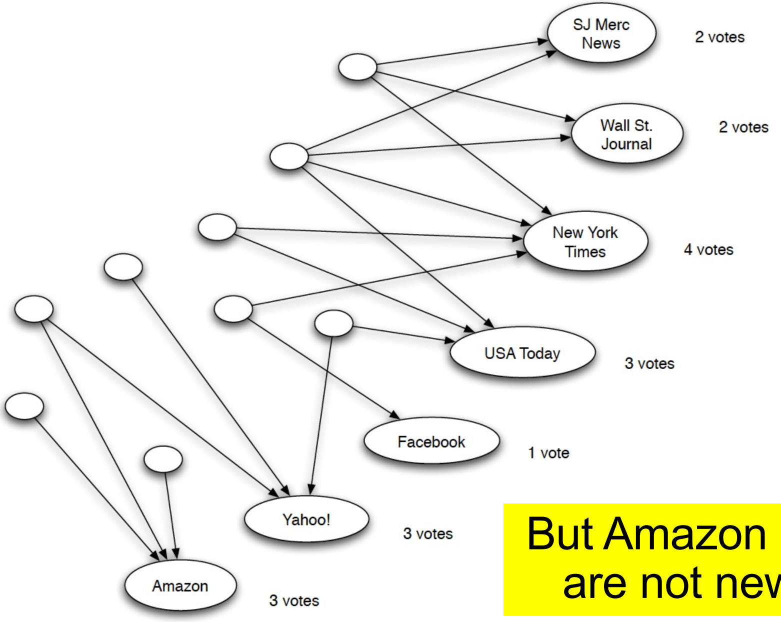
A simple approach to ranking: link-counting

- Collect a large sample of pages relevant to the query (text-only information retrieval)
- Each page in the sample "votes" through its links
- Count "votes" and sort results by number of in-links

- In the case of the query "tor vergata", which page on the Web receives greatest number of in-links from pages that are relevant to "tor vergata"?

A simple approach to ranking: example

- Counting in-links to pages for the query "newspapers"



But Amazon and Yahoo! are not newspapers!

A simple approach to ranking: pitfalls

1. Simple ranking algorithm based on link-counting does not work in all cases
   - Works well only when there is a single page that most people agree should be ranked first (e.g., home page)

- Consider the query “newspapers”
  - Not necessarily a single, intuitively “best” answer
  - Ideal answer would be a list of the most prominent newspapers on the Web

A simple approach to ranking: pitfalls

2. Google bombing
   - Easy to inflate the “votes” unfairly, causing a Web page to rank highly in web search engine results for irrelevant, unrelated or off-topic search terms by linking heavily
   - Examples: en.wikipedia.org/wiki/Google_bombing

- More generally, link building in the field of search engine optimization (SEO)
  - Those actions aimed at increasing the number and quality of in-links to a Web page with the goal of increasing the search engine rankings of that page

How to improve the ranking algorithm?

- Idea: make deeper use of the network structure than just counting in-links
- *"A document which points to many others might be a good **hub**, and a document that many documents point to might be a good **authority**. Recursively, a document that points to many good authorities might be an even better hub, and similarly a document pointed to by many good hubs might be an even better authority."*

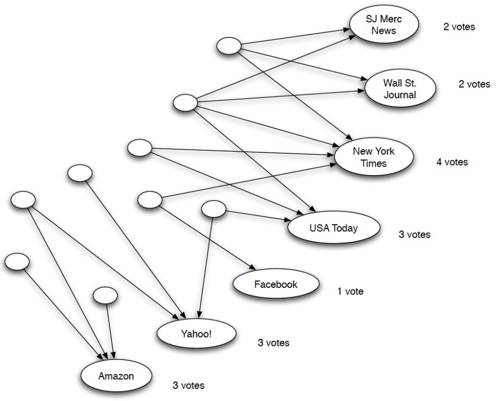
Monika Henzinger, former director of research at Google

List-finding technique

- We get high scores for well-known newspapers (results we wanted)
  - Along with pages that will receive lot of in-links no matter what the query is (like Yahoo!, Facebook, Amazon)
- In addition to newspapers themselves, we can find pages that compile list of links to online newspapers
- If we could find these list pages, we have another approach to find the newspapers themselves

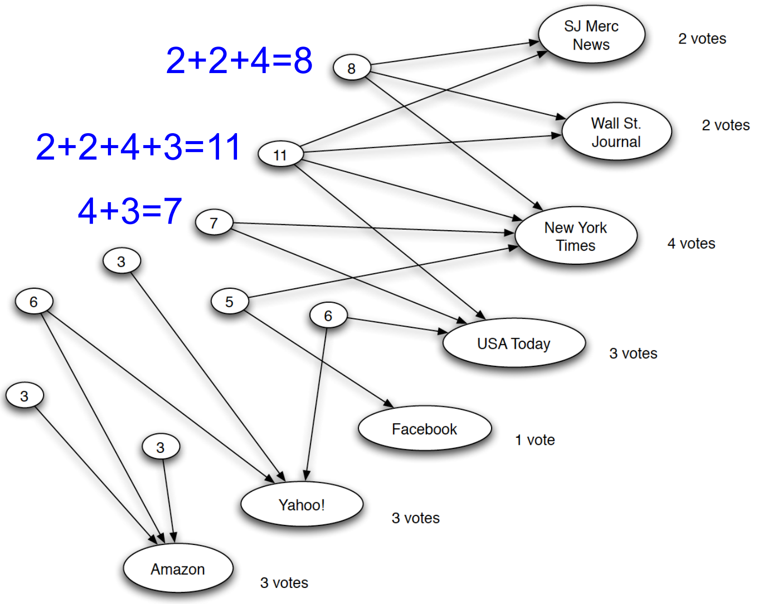
List-finding technique

- How to find good lists?
- Suspect that, among the pages casting votes, the ones that voted for popular pages (i.e., pages with many votes) should be lists
- Among the pages casting votes, a few of them voted for many of the pages that received a lot of votes



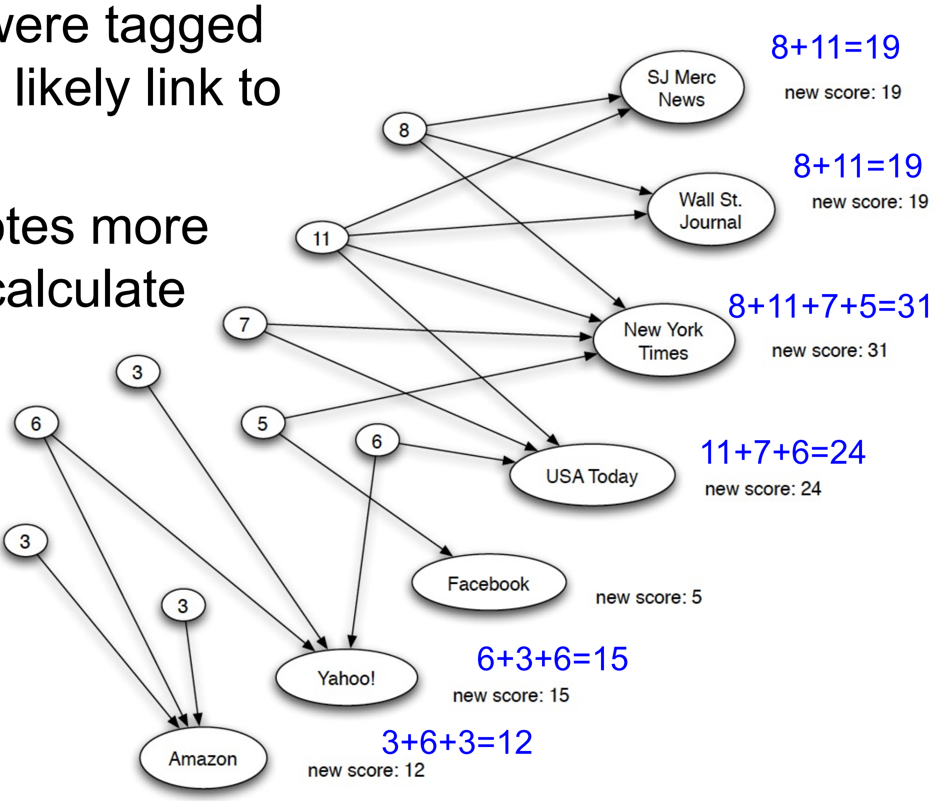
List-finding technique

- A page's value as a list is equal to the sum of the votes received by all pages that it voted for



Repeated improvement

- Pages that were tagged as lists most likely link to good results
- Give their votes more weight and calculate again
  - New score equal to the sum of the values of all lists that point to it

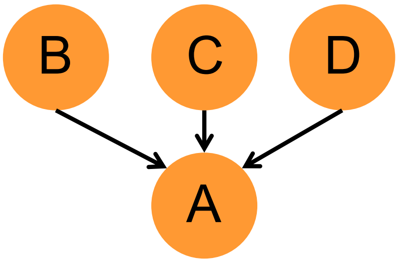


Repeated improvement

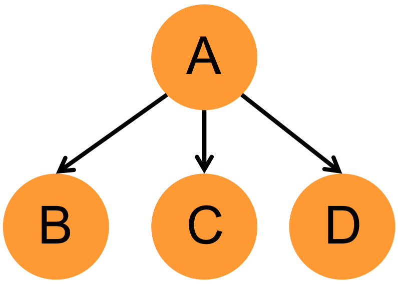
- Now bigger gap between good results and bad results (Yahoo! and Amazon), because the newspapers were endorsed by pages that were estimated to be good lists
- High score results are more likely to be good results now
- Why do we stop here?
- If we have better scores for pages on the right, we can use these to get still more refined values for the quality of the lists on the left

Authorities and hubs

- Authority: a page with many in-links
  - The idea is that the page may have good or authoritative content on some topic and thus many people trust it and link to it



- Hub: a page with many out-links
  - The page serves as an organizer of the information on a particular topic and points to many good authority pages on the topic



Authorities and hubs

- A good hub page for a topic points to many good authorities pages for that topic
- A good authority page for a topic is pointed to by many good hubs for that topic
- Kind of circular (mutual) reinforcement between authorities and hubs – which gives rise to *repeated improvement*

PageRank

- Google’s key algorithm to determine the order in which search results are displayed
- Based on the flow of authority via links
- Not query-dependent
- Applied to the entire web rather than to a local neighborhood of pages surrounding the results of a query (like in newspapers example)
- Used by Google in conjunction with other mechanisms (e.g., indexing) we do not study

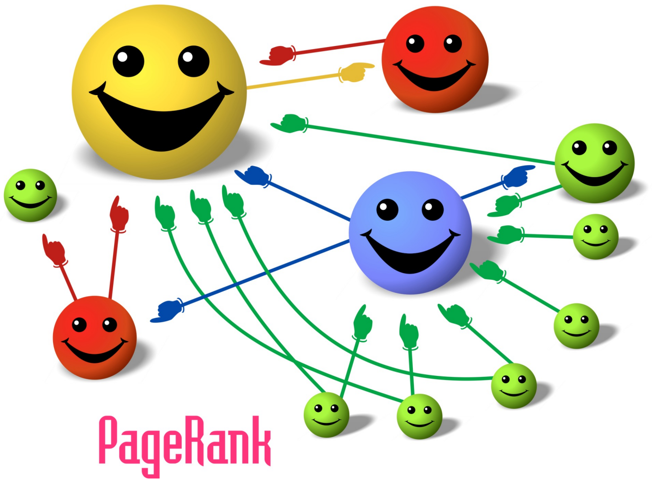
PageRank: some idea

- Observation: web pages vary greatly in terms of the number of in-links they have
  - E.g., Netscape home page had 62,804 in-links compared to most pages which had just a few in-links
- Highly linked pages are more “important” than pages with few links
- In-links coming from important pages convey more importance to a page
  - E.g., if a web page has a link off CNN’s home page, it may be just one link but it is a very important one

PageRank

- Count the number and quality of links to a web page to determine a rough estimate of how important that page is

- A page has high rank if the sum of the ranks of its in-links is high
- This covers both the case when a page has many in-links and when a page has a few highly ranked in-links

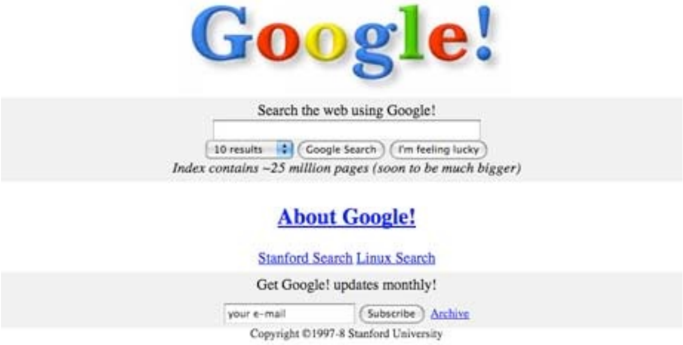


PageRank and Google

- Started with PhD thesis of Larry Pages at Stanford Univ. in 1996
- Sergey Brin, another PhD student, soon joined the project
- They realized that a search engine based on PageRank would produce better results than existing techniques
- Google was born!
  - First version of Google released in August 1996 on Stanford website
  - Company officially launched in 1998

PageRank and Google

- In 1997



Search the web using Google!

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

Get Google! updates monthly!

Copyright ©1997-8 Stanford University

- Google name: from Googol, the large number 10^{100}, and… its accidental misspelling!
- PageRank is not the only algorithm used by Google to order search results, but it is the first used algorithm and the best known

PageRank: basic definition

- Think of PageRank as a kind of “fluid” that circulates through the network, so that fluid passes from node to node across links
- Fluid pools at the nodes that are the most important

Computation of PageRank

- Iterative algorithm
- In a network with *N* nodes, assign all nodes the same initial PageRank, i.e., 1/*N*
- Choose a number of steps *k*
- Perform *k* updates to PageRank values, using PageRank update rule
- At each update, we refine the PageRank values and get closer to the final values

Basic PageRank update rule

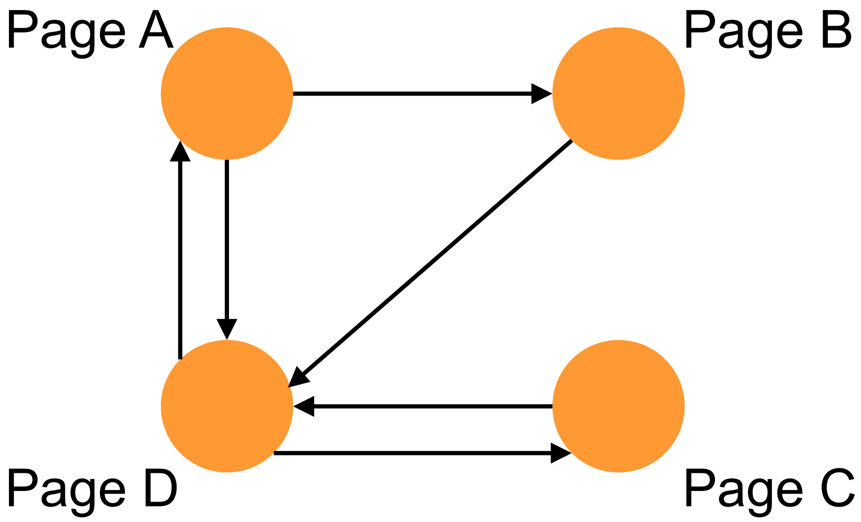
- Each page divides its current PageRank equally across its out-links, and passes these equal shares to the pages it points to
- PageRank is never created nor destroyed, just moved around from one node to another (no need to normalize)
- If a page has no outgoing links, it passes all its current PageRank to itself
- Each page updates its new PageRank to be the sum of the shares it receives

Basic PageRank update rule

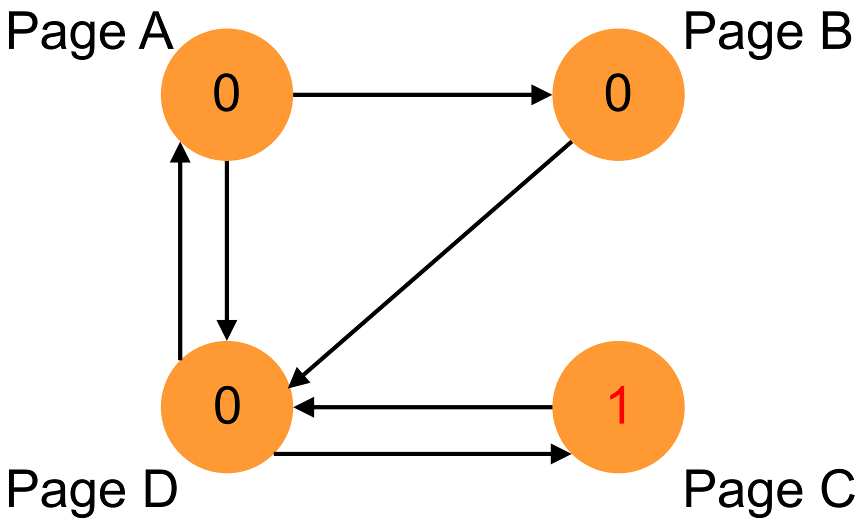
$$R_{i+1}(u) = \sum_{v \in B_u} \frac{R_i(v)}{N_v}$$

- where
  - *u*: a web page
  - B_u: set of pages that point to *u*
  - F_v: set of pages *v* points to
  - $N_v = |F_v|$: number of out-links from *v*
- Iterative equation
  - It may be computed by starting with any initial value of ranks (not only 1/*N*) and iterating the computation up to a given number of steps or until it converges

Example

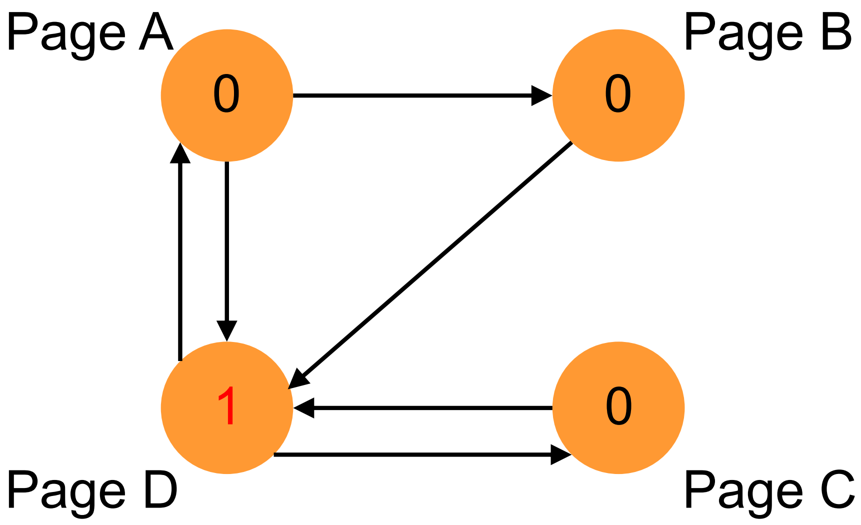


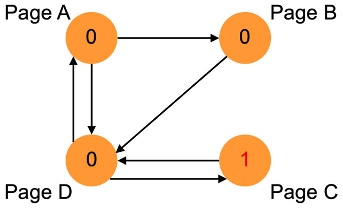
Example



Let’s start with all the PageRank assigned to page C

Example: step 1





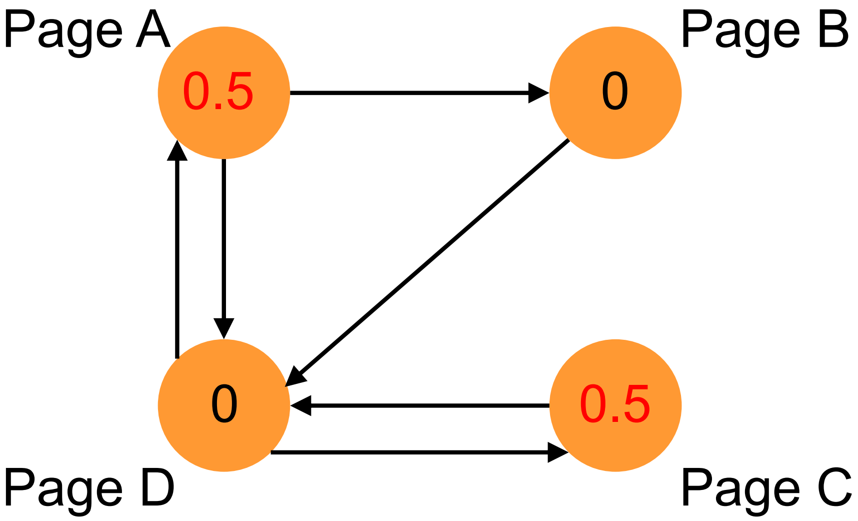
$R_1(A) = R_0(D)/2 = 0$

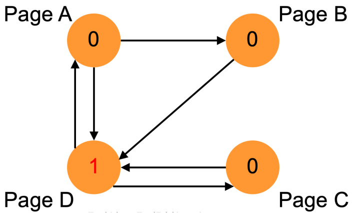
$R_1(B) = R_0(A)/2 = 0$

$R_1(C) = R_0(D)/2 = 0$

$R_1(D) = R_0(A)/2 + R_0(B) + R_0(C) = 1$

Example: step 2





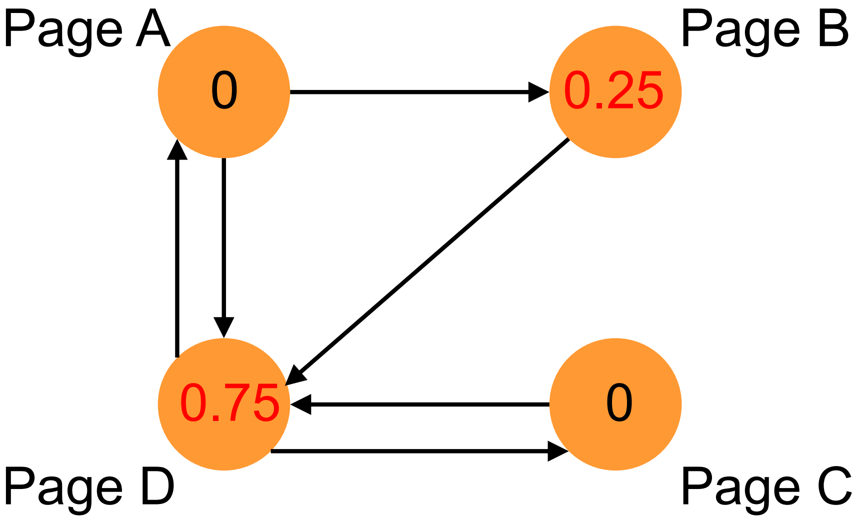
$R_2(A) = R_1(D)/2 = 1/2 = 0.5$

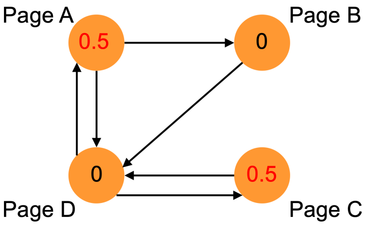
$R_2(B) = R_1(A)/2 = 0$

$R_2(C) = R_1(D)/2 = 1/2 = 0.5$

$R_2(D) = R_1(A)/2 + R_1(B) + R_1(C) = 0$

Example: step 3





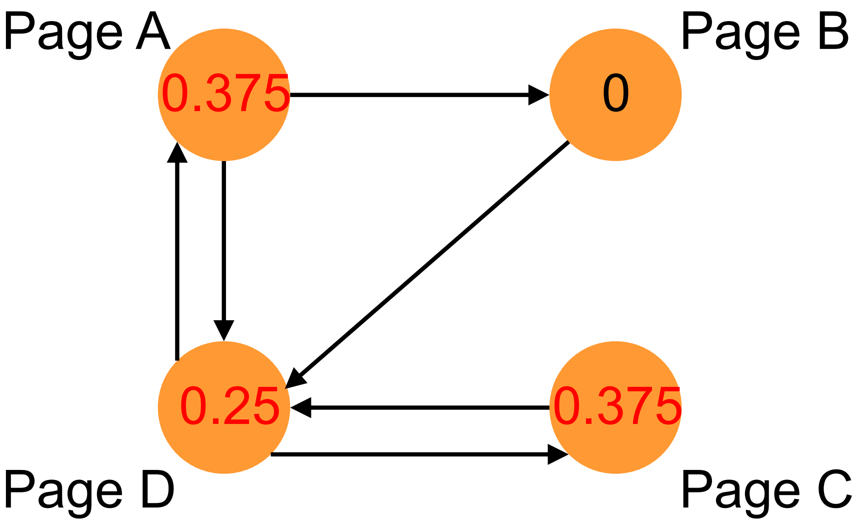
$R_3(A) = R_2(D)/2 = 0$

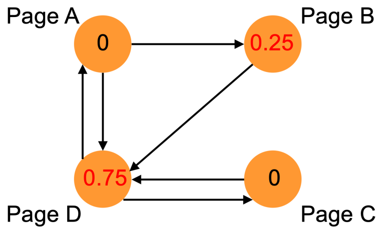
$R_3(B) = R_2(A)/2 = 0.5/2 = 0.25$

$R_3(C) = R_2(D)/2 = 0$

$R_3(D) = R_2(A)/2 + R_2(B) + R_2(C) = 0.5/2 + 0 + 0.5 = 0.75$

Example: step 4





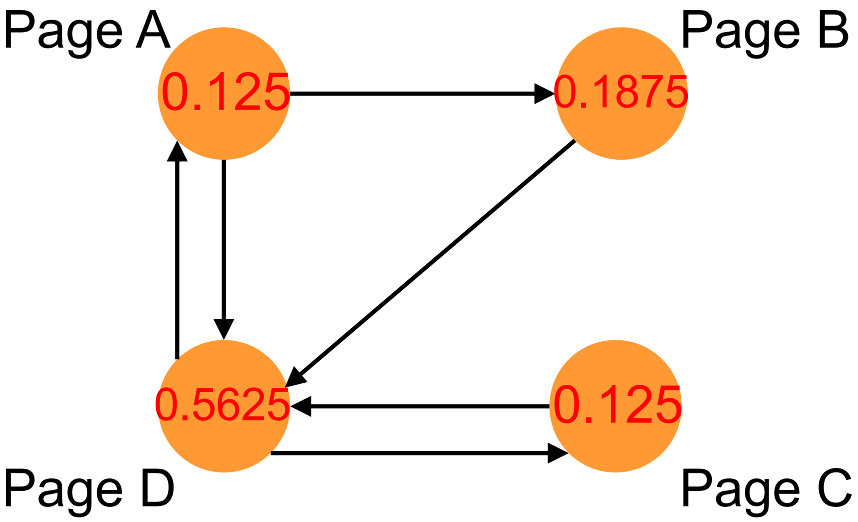
$R_4(A) = R_3(D)/2 = 0.75/2 = 0.375$

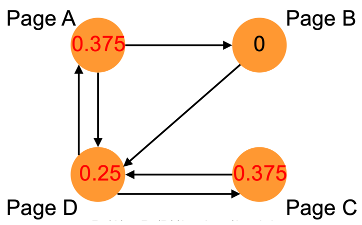
$R_4(B) = R_3(A)/2 = 0$

$R_4(C) = R_3(D)/2 = 0.75/2 = 0.375$

$R_4(D) = R_3(A)/2 + R_3(B) + R_3(C) = 0 + 0.25 + 0 = 0.25$

Example: step 5





$R_5(A) = R_4(D)/2 = 0.25/2 = 0.125$

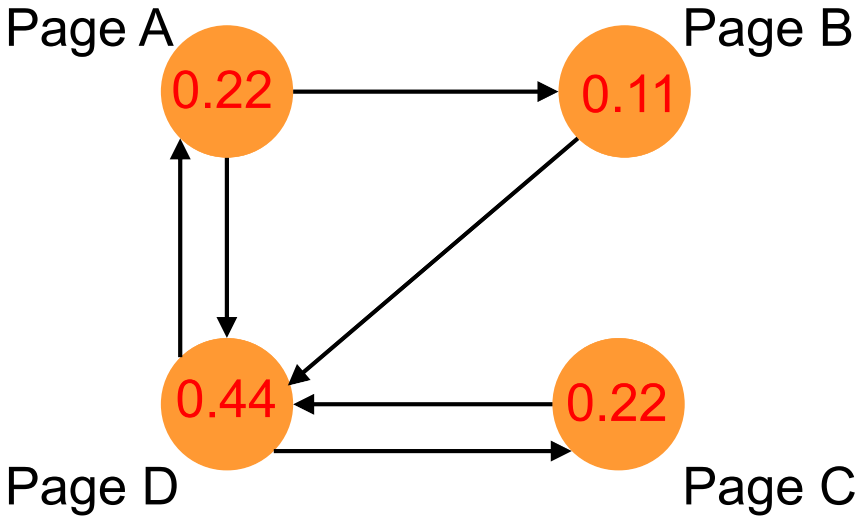
$R_5(B) = R_4(A)/2 = 0.375/2 = 0.1875$

$R_5(C) = R_4(D)/2 = 0.25/2 = 0.125$

$R_5(D) = R_4(A)/2 + R_4(B) + R_4(C) = 0.375/2 +0+0.375 = 0.5625$

Example: after k steps

- After few iterations we stop here: why?



$R_k(A) = R_{k-1}(D)/2 = 0.44/2 = 0.22$

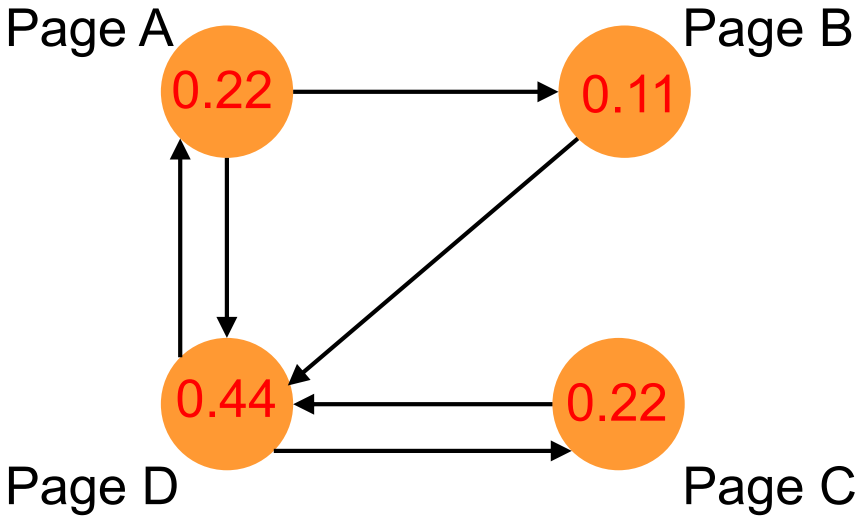
$R_k(B) = R_{k-1}(A)/2 = 0.22/2 = 0.11$

$R_k(C) = R_{k-1}(D)/2 = 0.44/2 = 0.22$

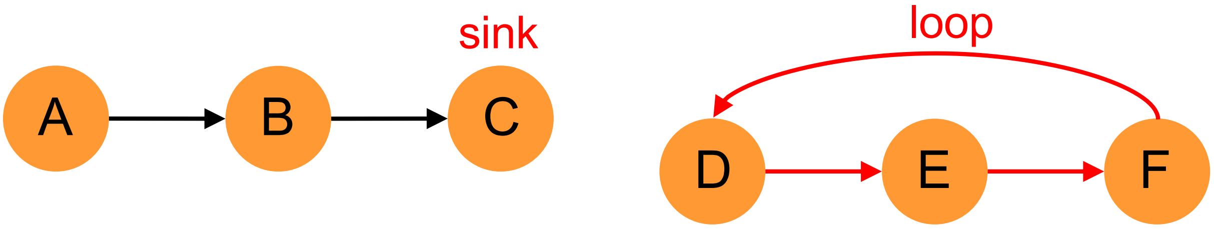
$R_k(D) = R_{k-1}(A)/2 + R_{k-1}(B) + R_{k-1}(C) = 0.22/2 + 0.11 + 0.22 = 0.44$

PageRank equilibrium

- The update process converges to limiting values
- At equilibrium, doing another PageRank update does not change anything



Problem: sink nodes and loops



- C is a “sink” (dead end, no out-links): it does not redistribute the PageRank
- Pages in a loop accumulate PageRank but do not redistribute it
- Fluid analogy: why water does not end up running downhill and residing exclusively at lowest points?
  - Counterbalancing factor: evaporation
- In PageRank algorithm: teleportation, i.e., with a certain probability jump to any other web page to get out of loop or sink

Scaled PageRank

$$R_{i+1}(u) = c \sum_{v \in B_u} \frac{R_i(v)}{N_v} + (1-c)\frac{1}{N}$$

- Scale down all PageRank values by a damping factor ***c*** (total PageRank in network shrunk from 1 to *c*<1)
- Divide the residual (1-*c*) units of PageRank equally over all, giving (1-*c*)/*N* to each
- Damping factor *c* is a user-designed parameter, usually chosen in [0.8, 0.9] and often equal to 0.85
- Equilibrium is sensitive to choice of *c*

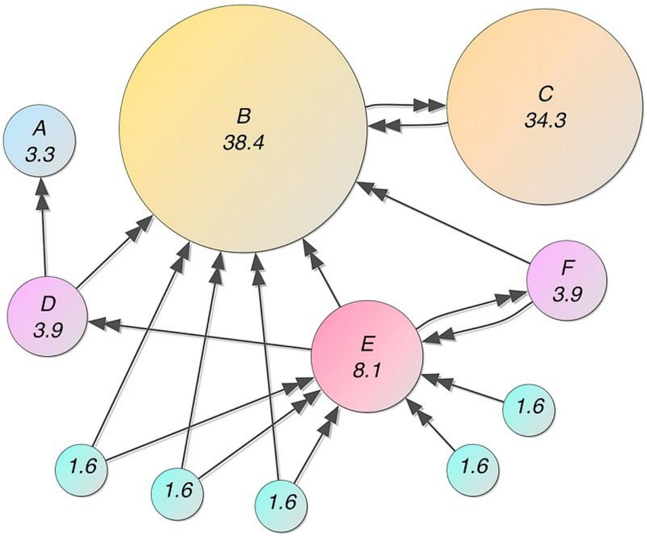
Probabilistic interpretation of PageRank

- *Basic PageRank* can be formulated as a random walk on the web graph
  - Start by choosing a page at random, picking each page with equal probability (i.e., 1/*N*)
  - Then follow links for a sequence of steps: in each step, pick a random outgoing link from current page, and follow it to where it leads (if current page has no outgoing links, then just stay where you are)
- PageRank of page X after *k* applications of *basic update rule* corresponds to probability of being at page X after *k* steps of random walk

Probabilistic interpretation of PageRank

- *Scaled PageRank* corresponds to “scaled” random walk
  - With probability *c*, walker follows a random outgoing link
  - With probability 1-*c*, walker jumps to some random node
- PageRank of page X after *k* applications of *scaled update rule* corresponds to probability of being at page X after *k* steps of scaled random walk

Example

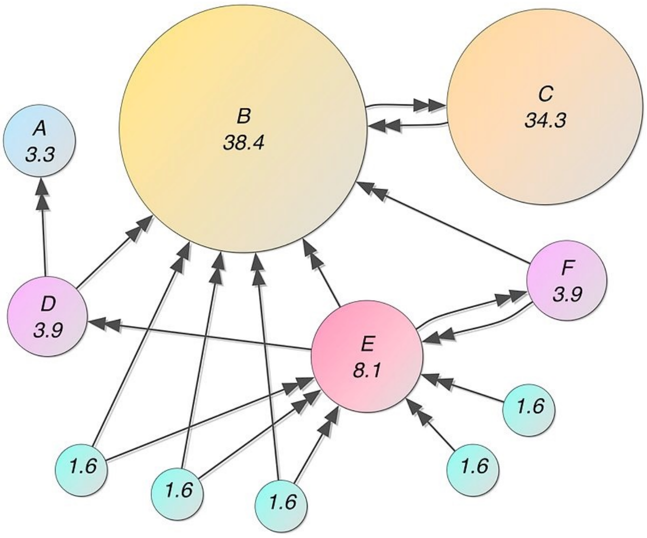


- In figure, sum of all PageRank scores is equal to 100 (instead of 1)

- Page C has a higher PageRank than Page E, even though there are fewer links to C
- The one link to C comes from an important page and hence is of high value

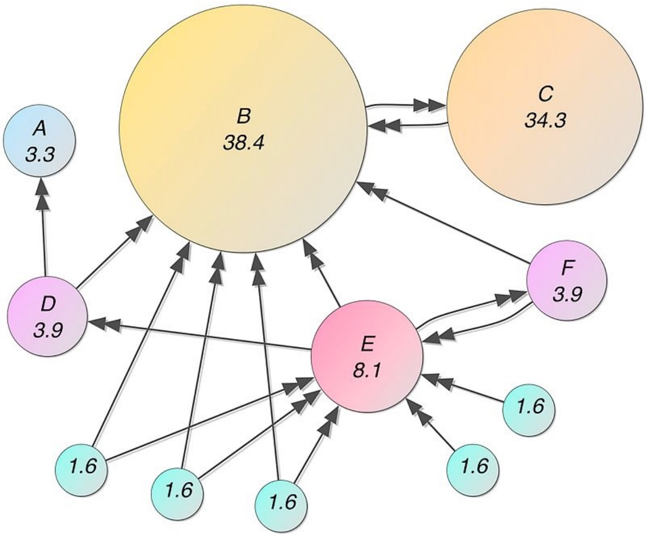
Example

- If a web surfer who starts on a random page has an 85% likelihood of choosing a random link from the page she is currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, she will reach Page E 8.1% of the time
  - The 15% likelihood of jumping to an arbitrary page corresponds to setting the damping factor *c*=0.85



Example

- Without damping (i.e., *c*=1), all web surfers would eventually end up on pages A, B, or C, and all other pages would have PageRank equal to 0
- In the presence of damping, page A effectively links to all pages in the web, even though it has no outgoing links of its own

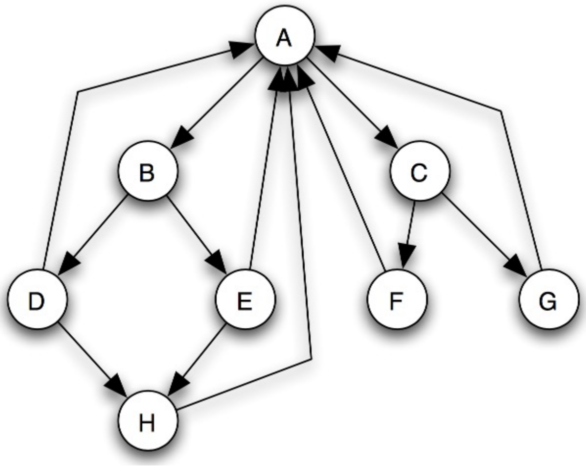


Exercise

a) Can you guess which page will achieve the highest PageRank?

b) Prove it using basic PageRank update rule (and stop after 4 iterations)

- Start with each page having PageRank equal to 1/8



Evolution of PageRank

- We have analyzed the original algorithm
- To improve the accuracy and relevance of search results, Google made updates to the PageRank algorithm, incorporating:
  - Keyword analysis
  - User behavior
  - Techniques to detect and penalize pages that use spammy or manipulative link-building practices to artificially improve their rankings (Penguin algorithm)
  - Techniques to Improve the ranking of high-quality pages (Panda algorithm)

Limitations of PageRank

- PageRank scores do not reflect current events
  - PageRank scores are not calculated at the time of search but rather determined at the time of indexing (Google scans by means of crawlers each page on the Web for topics and key phrases, and subsequently records the relevant pages to each topic or key phrase)
- Inability to handle queries containing natural language and information outside of keywords

PageRank beyond the Web

- PageRank is used in a variety of applications far beyond its origins in Google’s search engine
- In recommender systems to find the currently trending product that users might be willing to purchase
- In social networks as centrality measure
- In chemistry to study molecules
- In biology and bioinformatics to reveal localized information, e.g., to find correlated genes
- In neuroscience to study the human brain connectome
- In road and urban space to predict traffic flow and human movement
- In sport to rank football teams
- In literature, e.g., to find the most important books

References

- D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, chapter 14, Cambridge University Press, 2010.
- L. Page, S. Brin, The PageRank Citation Ranking: Bringing Order to the Web, 1999.

- Some video:
- A. Meyer, PageRank The Flow Formulation, Stanford Univ.
- V. Lavrenko, PageRank algorithm: how it works