HG2052

Language, Technology and the Internet

Citation, Reputation and PageRank

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

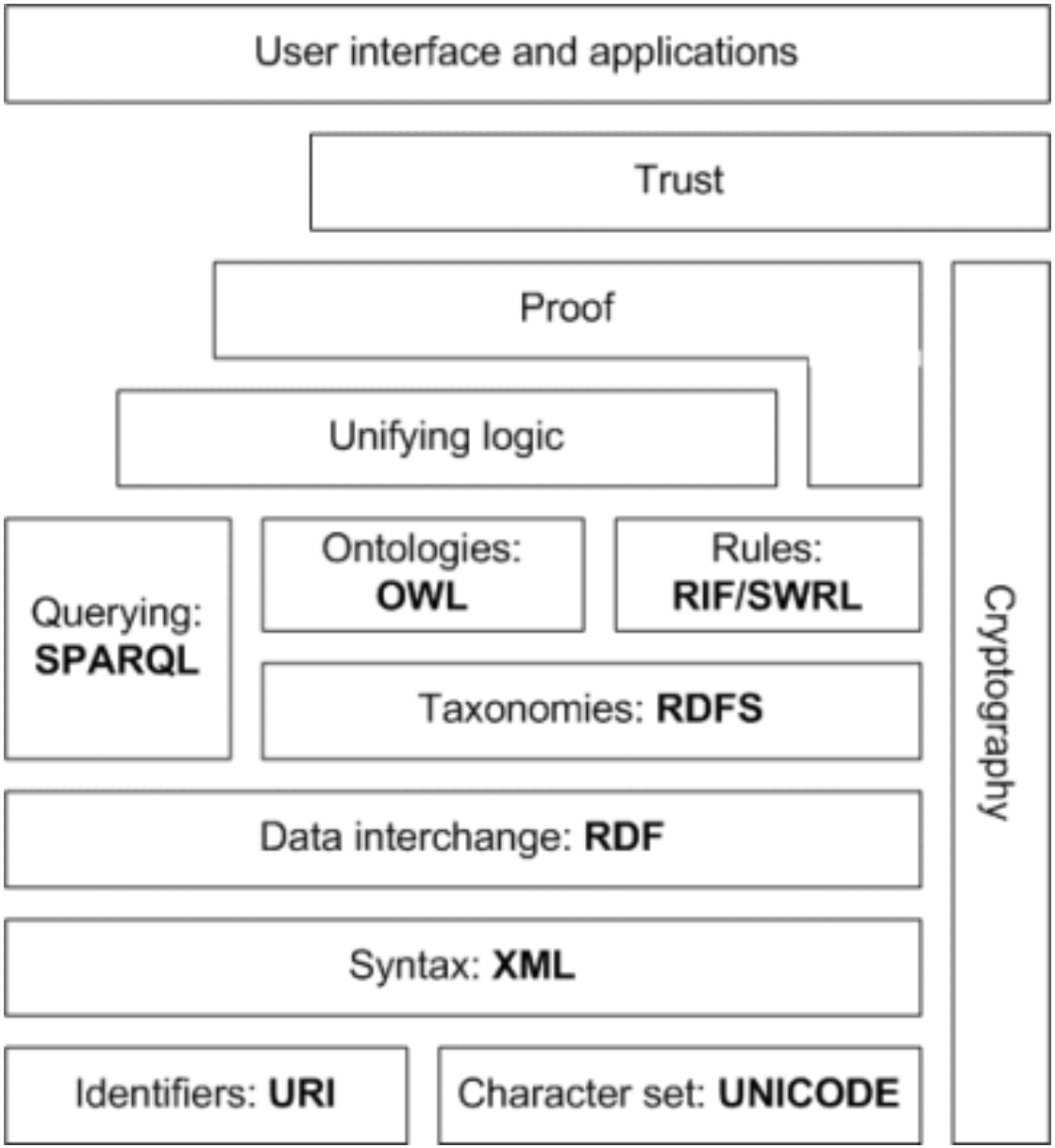
Lecture 11

Review of the Semantic Web

- Web of data
  - provides common data representation framework
  - makes possible integrating multiple sources
  - so you can draw new conclusions
- Increase the utility of information by connecting it to definitions and context
- More efficient information access and analysis

E.G. not just "color" but a concept denoted by a Web identifier:
<http://pantone.tm.example.com/2002/std6#color>

Semantic Web Architecture



Semantic Web Architecture (details)

- Identify things with Uniform Resource Identifiers
  - Universal Resource Name: `urn:isbn:1575864606`
  - Universal Resource Locator:: `http://www3.ntu.edu.sg/home/fcbond/`
- Identify relations with Resource Description Framework
  - Triples of <subject, predicate, object>
  - Each element is a URI
  - RDFs are written in well defined XML
  - You can say anything about anything
- You can build relations in ontologies (OWL)
  - Then reason over them, search them, ...

Overview of Citation, Reputation and PageRank

- Citation and Reputation
  - Identifying high quality research
- PageRank (Google's algorithm for ranking web pages)
  - Identifying interesting web pages

Citation Networks

- How can we tell what is a good scientific paper?
  - Content-based
    - Read it and see if it is interesting (hard for a computer)
    - Compare it to other things you have read and liked
  - Context based
    - See who else read and thought it interesting enough to cite

Citation Networks

- Citation networks are information networks formed by citations between articles. Citation networks arise from scientific publications, patent filings, or other more exotic information systems.

- Citation networks are directed graphs
  - Can be cyclic — papers can cite each other!

- Citation networks feature a distinct time-arrow, because researchers can only cite articles that have already been written. As a result, movement on a citation network is always backwards in time.

- Citations are rarely altered after an article is published. Thus, unlike the www, links in citation networks are never updated. One known consequence of this is that citation networks are prone to serious ageing effects.

Reputation and Citation Analysis

- One major use of citation networks is in measuring productivity and impact of the published work of a scientist, scholar or research group

- Some scores are
  - Total Number of Citations (Pretty Useful)
  - Total Number of Citations minus Self-citations
  - Total Number of (Citations / Number of Authors)
  - Average (Citation * IF / Number of Authors)

- Problems
  - Not all citations are equal: citations by 'good' papers are better
  - Newer publications suffer in relation to older ones
  - Newer researchers suffer in relation to older ones

Impact Factor (IF)

- (IF) is a measure reflecting the average number of citations to articles published in journals

  IF(Journal) = the average number of citations received per paper published in that journal during the n preceding years (default $n = 2$)

  Depends a lot on finding all the citations

- It is used to estimate the relative importance of a journal within its field

- Journals with higher impact factors are more important than those with lower ones.

Impact Factor (Validity)

- The impact factor is highly discipline-dependent
- The percentage of total citations occurring in the first two years after publication varies highly among disciplines
  - 1–3% in the mathematical and physical sciences
  - 5–8% in the biological sciences
- In the short term - especially in the case of low-impact-factor journals - many of the citations to a certain article are made in papers written by the author(s) of the original article.

The Hirsch Index (H-Index)

- Hirsch (2005) writes:

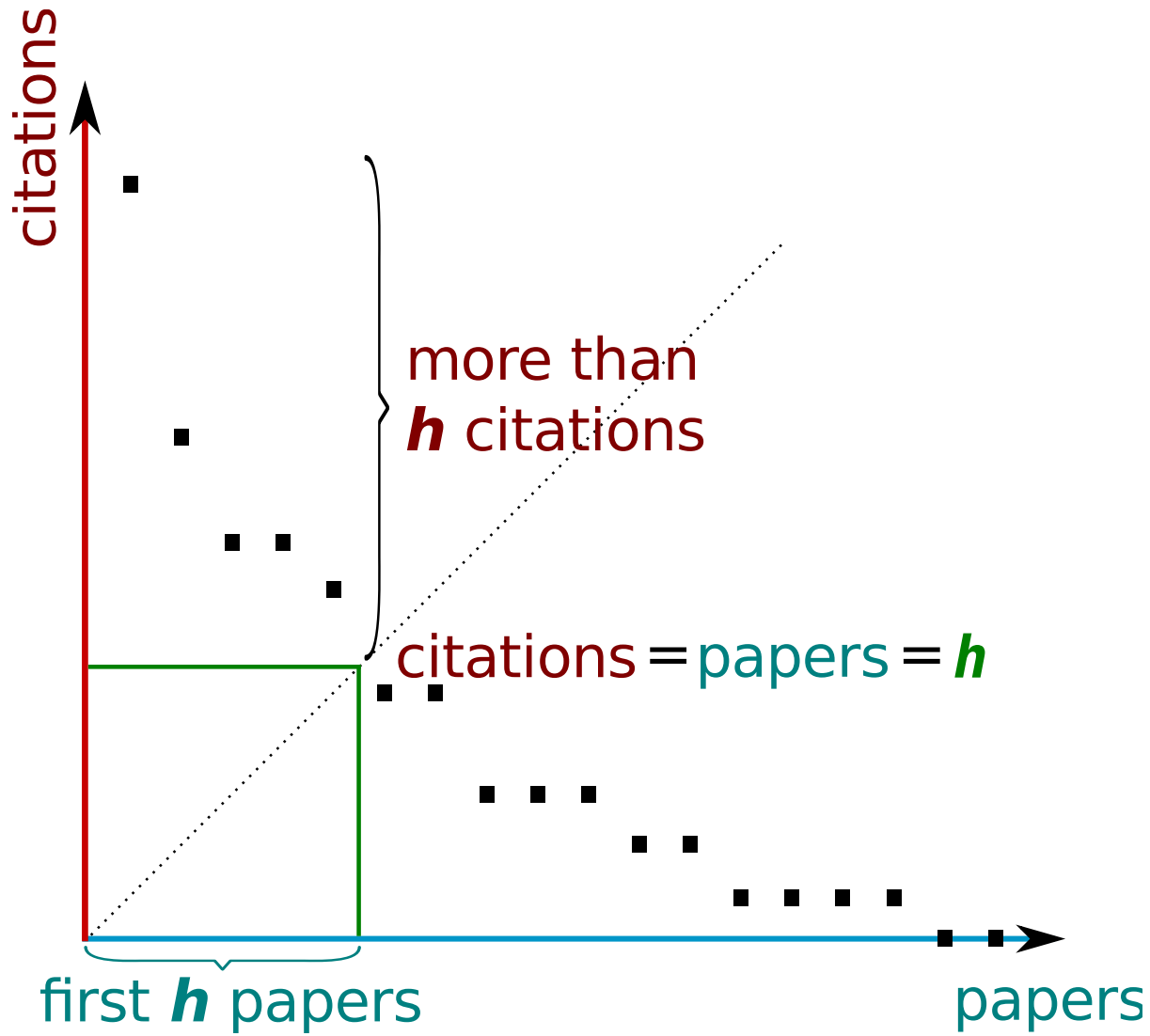
  A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.

- In other words, a scholar with an index of h has published h papers each of which has been cited in other papers at least h times.

- Average h differs for different disciplines (and what you count as a citation)

- In Physics, high h correlates well with whether a scientist has won honors like National Academy membership or the Nobel Prize.

- H-Index is not affected strongly by few papers with many citations and many papers with few citations.


citations
more than
h citations
citations = papers = h
first h papers
papers

Multiple Authors

Natural Sciences In the life sciences, first listing is usually given to the researcher who did most of the work, both physical and intellectual, and last billing goes to the mentor or person who guided the project and whose grant money paid for the project - the PI.

Chemistry The senior author is sometimes the first author on a paper, even if a postdoc completed the bulk of the work.

Archeology, Economics Strict alphabetical by surname

- Dr Aardvark gets promoted more than Dr Zwilensky

Liran Einav and Leeat Yariv (2006) "What's in a Surname? The Effects of Surname Initials on Academic Success", *The Journal of Economic Perspectives*, **20(1)**, pp. 175-188

Practices also vary by country, ...

Who to include

Honorary Authorship In many places, the project PI or research group leader is always included (even if they have not read the paper). The United States National Academy of Sciences, warns that such practices "dilute the credit due the people who actually did the work, inflate the credentials of those so 'honored,' and make the proper attribution of credit more difficult.".

Ghost Authorship Ghost authorship occurs when an individual makes a substantial contribution to the research but is not listed as an author.

- Technicians and data gatherers are typically not included.
- Some pharmaceutical companies ghost write papers
- Many big names have farms of research assistants, who are typically not given authorship.

Gaming Citations

- Least/Minimum Publishable Unit
  - Break research into small chunks to increase the number of citations
  - Sometimes there is very little new information
- Self citation, in-group citation
- Write only proceedings (some journals are not often read)
- Submitting only to High Impact factor journals

You improve what gets measured
not necessarily what you want to improve

Judging Citations

- Not all citations are positive:
  - Contrary to Bond *et al* (2005), we found ...
  - This invalidates Bond *et al* (2005).
- Not all citations are equal:
  - We follow closely Bond *et al* (2005) as it is efficient and accurate.
  - Other approaches include Band *et al* (2004); Bind 2006; Bend 2000.
- Recent research tries to classify the citation types using cue phrases:
  - Statement of weakness
  - Contrast or comparison
  - Agreement/Usage
  - Neutral

Other uses of Citation NetWorks

- Measure the similarity of two articles by the overlap of other articles citing them.
- This is called co-citation similarity.
  Two ways of measuring similarity based on hyperlinks:
  - Cocitation similarity: The two articles are cited by the same articles.
  - Bibliographic coupling similarity: The two articles cite the same articles.
- Co-citation similarity on the web: Google's "find pages like this" or "Similar" feature
- Citation analysis is a big deal: The budget and salary of this lecturer are / will be determined by the impact of his publications!

How do you decide what to cite?

Ranking Web Pages

- Static Ranking
- Page Rank
- Gaming Rankings
- Digital Object Identifier (DOI)

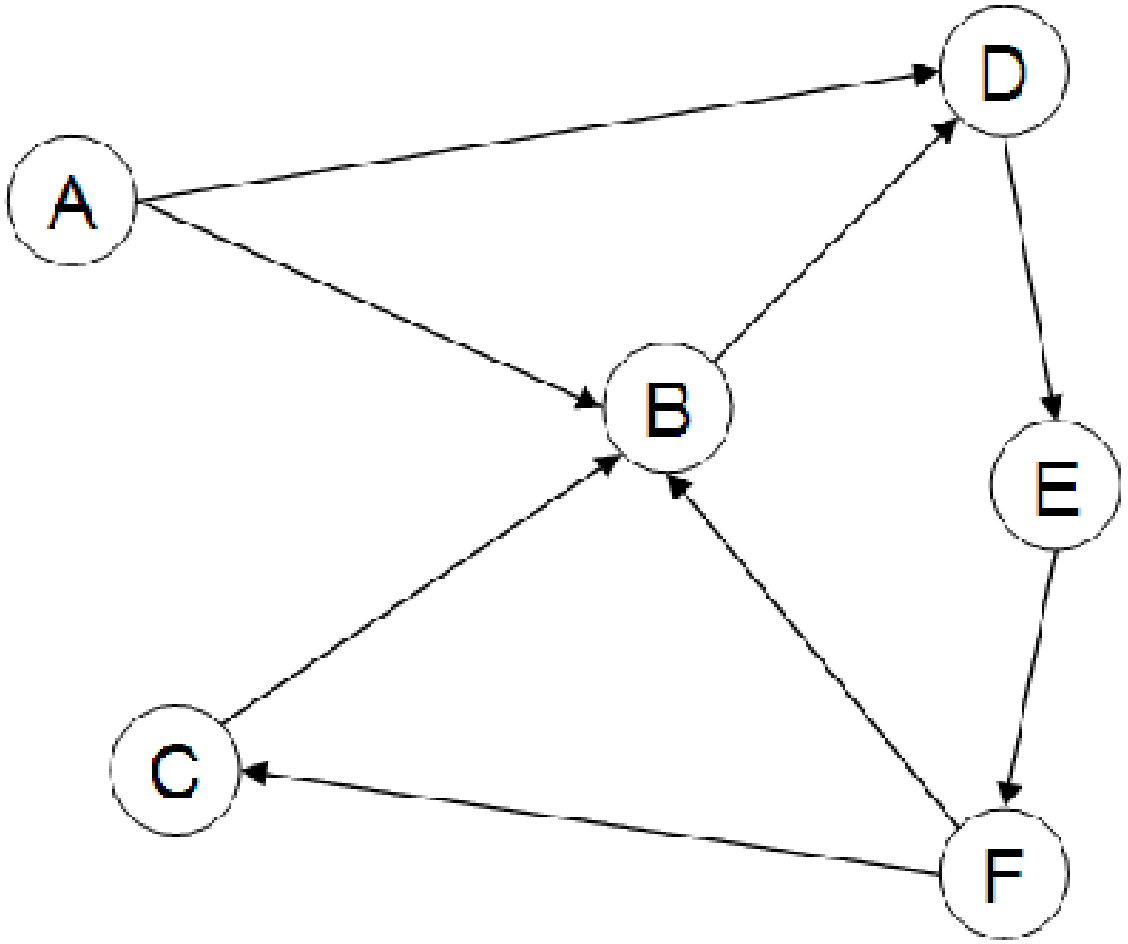
Static Ranking

- Known Reputable Sites
  - Wikipedia, `.edu` domains, ...
- Spam detection
- (Real) Popularity (number of click throughs in search)
- Page features
  - Length of page, length of URL
- Anchor Text features
  - How much text in links to page, how varied, ...

Ranking Web Pages

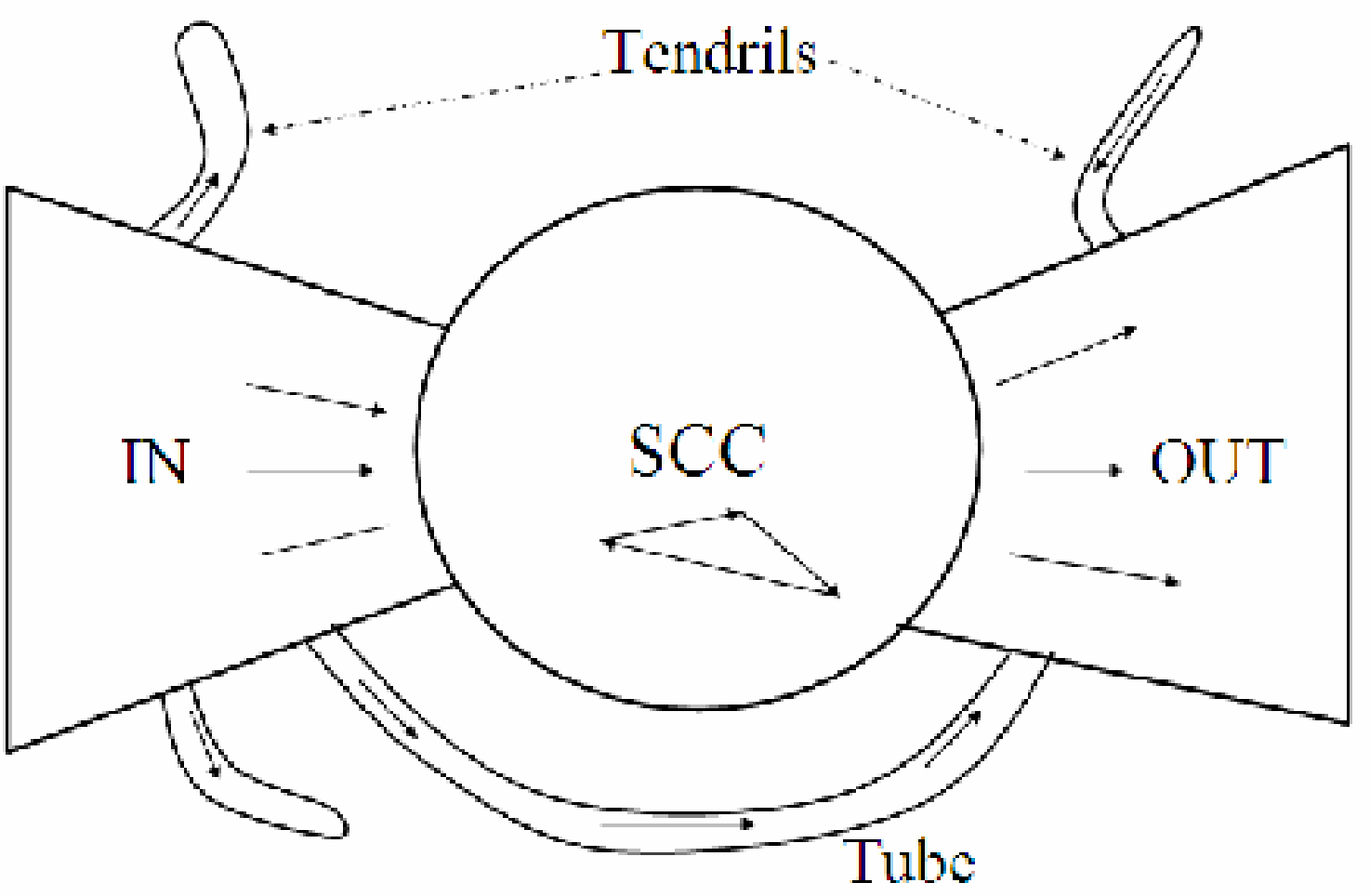
- Web Characteristics: What does the web look like
- Anchor text: What exactly are links on the web and why are they important for finding web pages?
- Citation analysis: the mathematical foundation of PageRank and link-based ranking
- PageRank: the original algorithm that was used by Google for link-based ranking on the web
- Gaming PageRank: Search Engine Optimization
- Other Applications

Characteristics of the Web: Graph



You can't go anywhere from anywhere!

Characteristics of the Web: Bow Tie



SCC: Strongly Connected Core — can travel from any page to any page

Anchor Text

- Recall how hyperlinks are written:

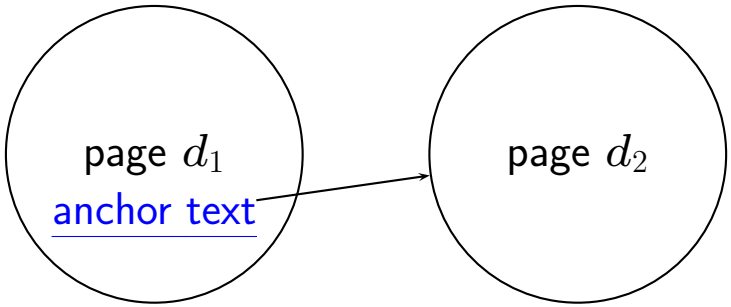
```
<a href="http://path.to.there/page/HG803/">HG803:
Language, Technology and the Internet.</a>

For more information about Language, Technology and the
Internet, see the <a href="http://..">HG803 Course Page.</a>
```

- Link analysis builds on two intuitions:

1. The hyperlink from A to B represents an endorsement of page B, by the creator of page A.
2. The (extended) anchor text pointing to page B is a good description of page B.

This is not always the case; for instance, most corporate websites have a pointer from every page to a page containing a copyright notice.

The web as a directed graph

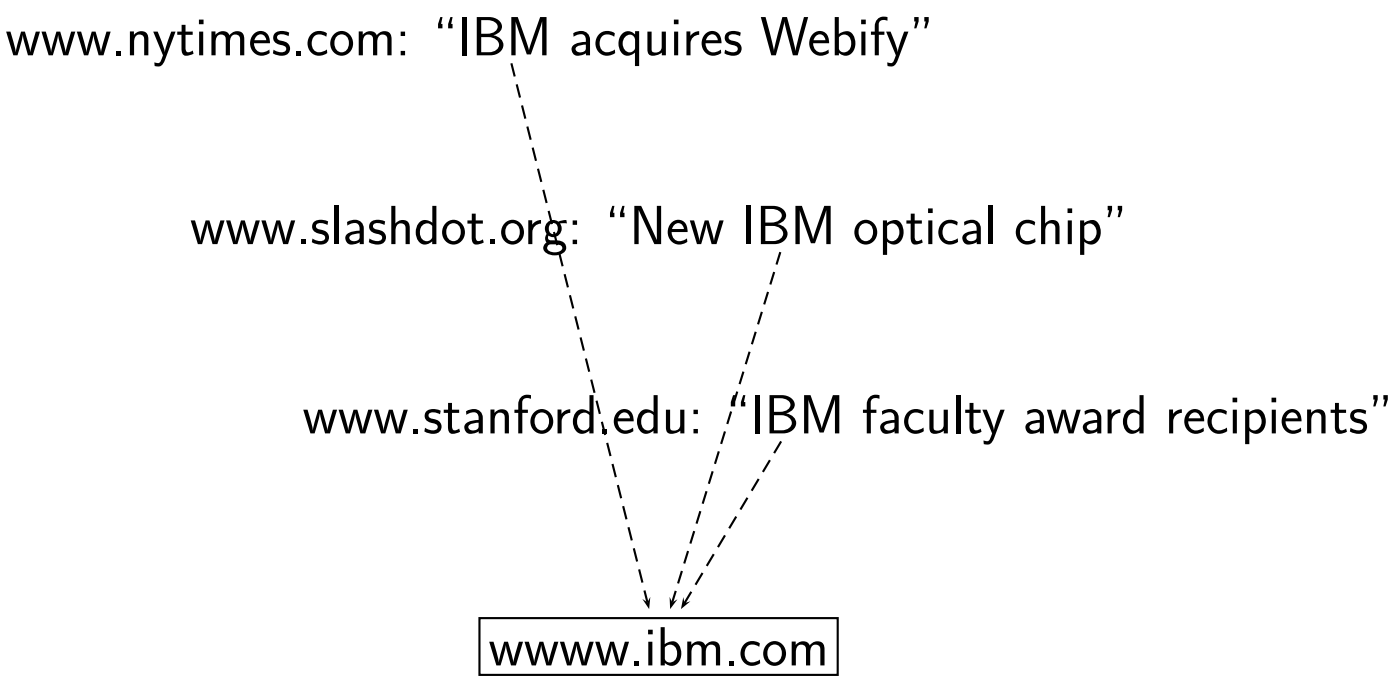


- Assumption 1: A hyperlink is a quality signal.
  - Hyperlink $d_1 \rightarrow d_2$ implies that d_1's author deems d_2 high-quality and relevant.
- Assumption 2: The anchor text describes the content of d_2.
  - We use anchor text loosely here for any text surrounding the hyperlink.
  - Example: "You can find cheap cars here."
  - Anchor text: "You can find cheap cars here"

[text of d_2] only vs. [text of d_2] + [anchor text $\rightarrow d_2$]

- Searching on [text of d_2] + [anchor text $\rightarrow d_2$] is often more effective than searching on [text of d_2] only.
- Example: Query *IBM*
  - Matches IBM's copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - May not match IBM home page!
  - ...if IBM home page is mostly graphics
- Searching on [anchor text $\rightarrow d_2$] is better for the query *IBM*.
  - In this representation, the page with the most occurrences of *IBM* is www.ibm.com.

Anchor text containing *IBM* pointing to www.ibm.com



Exercise: Assumptions underlying PageRank

- Assumption 1: A link on the web is a quality signal – the author of the link thinks that the linked-to page is high-quality.
- Assumption 2: The anchor text describes the content of the linked-to page.
- Is assumption 1 true in general?
- Is assumption 2 true in general?

Google bombs

- A Google bomb is a search with "bad" results due to maliciously manipulated anchor text.
- Google introduced a new weighting function in January 2007 that fixed many Google bombs.
- Still some remnants: [dangerous cult] on Google, Bing, Yahoo
  - Coordinated link creation by those who dislike the Church of Scientology
- Defused Google bombs: [dumb motherf....], [who is a failure?], [evil empire]

Origins of PageRank: Citation analysis (1)

- Citation analysis: analysis of citations in the scientific literature
- Example citation: "Miller (2001) has shown that physical activity alters the metabolism of estrogens."
- We can view "Miller (2001)" as a hyperlink linking two scientific articles.
- One application of these "hyperlinks" in the scientific literature:

Origins of PageRank: Citation analysis (2)

- Another application: Citation frequency can be used to measure the impact of an article.
  - Simplest measure: Each article gets one vote – not very accurate.
- On the web: citation frequency = inlink count
  - A high inlink count does not necessarily mean high quality ...
  - ...mainly because of link spam.
- Better measure: weighted citation frequency or citation rank
  - An article's vote is weighted according to its citation impact.
  - Circular? No: can be formalized in a well-defined way.

Origins of PageRank: Citation analysis (3)

- PageRank: weighted citation frequency or citation rank
- PageRank was invented in the context of citation analysis by Pinsker and Narin in the 1960s.

Origins of PageRank: Summary

- We can use the same formal representation for
  - citations in the scientific literature
  - hyperlinks on the web
- Appropriately weighted citation frequency is an excellent measure of quality ...
  - ...both for web pages and for scientific publications.
- Next: PageRank algorithm for computing weighted citation frequency on the web

Model behind PageRank: Random walk

- Imagine a web surfer doing a random walk on the web (virtual)
  - Start at a random page
  - At each step, go out of the current page along one of the links on that page, equiprobably

- In the steady state, each page has a long-term visit rate.
  what proportion of the time someone will be there

- This long-term visit rate is the page's PageRank.

- PageRank = long-term visit rate
  = steady state probability of being at a page

Long-term visit rate

- Recall: PageRank = long-term visit rate
- Long-term visit rate of page d is the probability that a web surfer is at page d at a given point in time.
- To calculate this the web graph must not contain dead ends.
  - But the web is full of dead ends
  - Random walk can get stuck in dead ends
  - If there are dead ends, long-term visit rates are not well-defined (or non-sensical)

Teleporting – to get us out of dead ends

- At a dead end, jump to a random web page with prob. $1/N$.
- At a non-dead end, with probability 10%, jump to a random web page (to each with a probability of $0.1/N$).
- With remaining probability (0.9), go out on a random hyperlink.
  - For example, if the page has 4 outgoing links: randomly choose one with probability (1-0.10)/4=0.225
- 10% is a parameter, the teleportation rate.
- Note: "jumping" from a dead end is independent of teleportation rate.

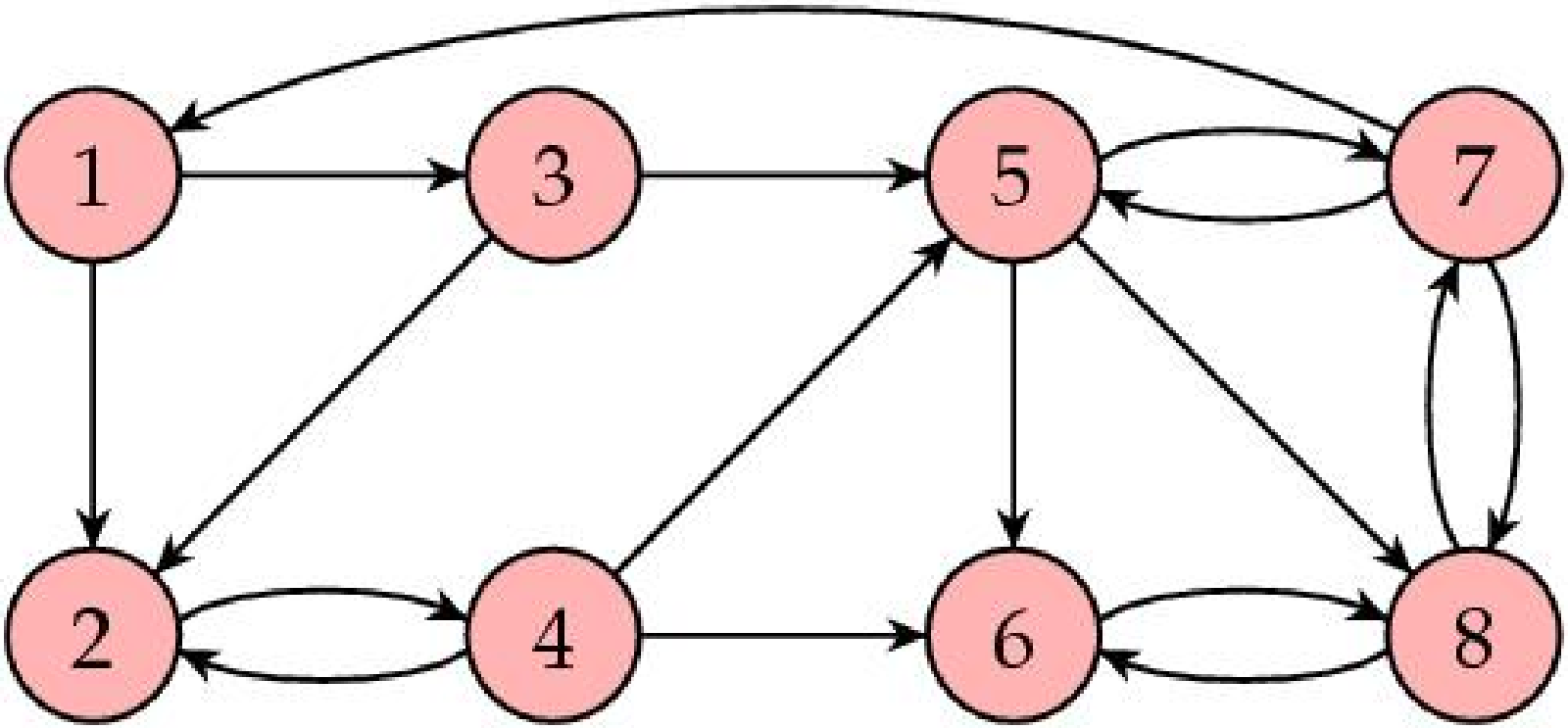
How do we compute the steady state vector?

- In other words: how do we compute PageRank?
- Recall: $\vec{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ is the PageRank vector, the vector of steady-state probabilities ...
- ...and if the distribution in this step is $\vec{x}$, then the distribution in the next step is $\vec{x}P$.
- But $\vec{\pi}$ is the steady state: $\vec{\pi} = \vec{\pi}P$
- Solving this matrix equation gives us $\vec{\pi}$.
- $\vec{\pi}$ is the principal left eigenvector for P
  (a well known mathematical concept)

One way of computing the PageRank $\vec{\pi}$

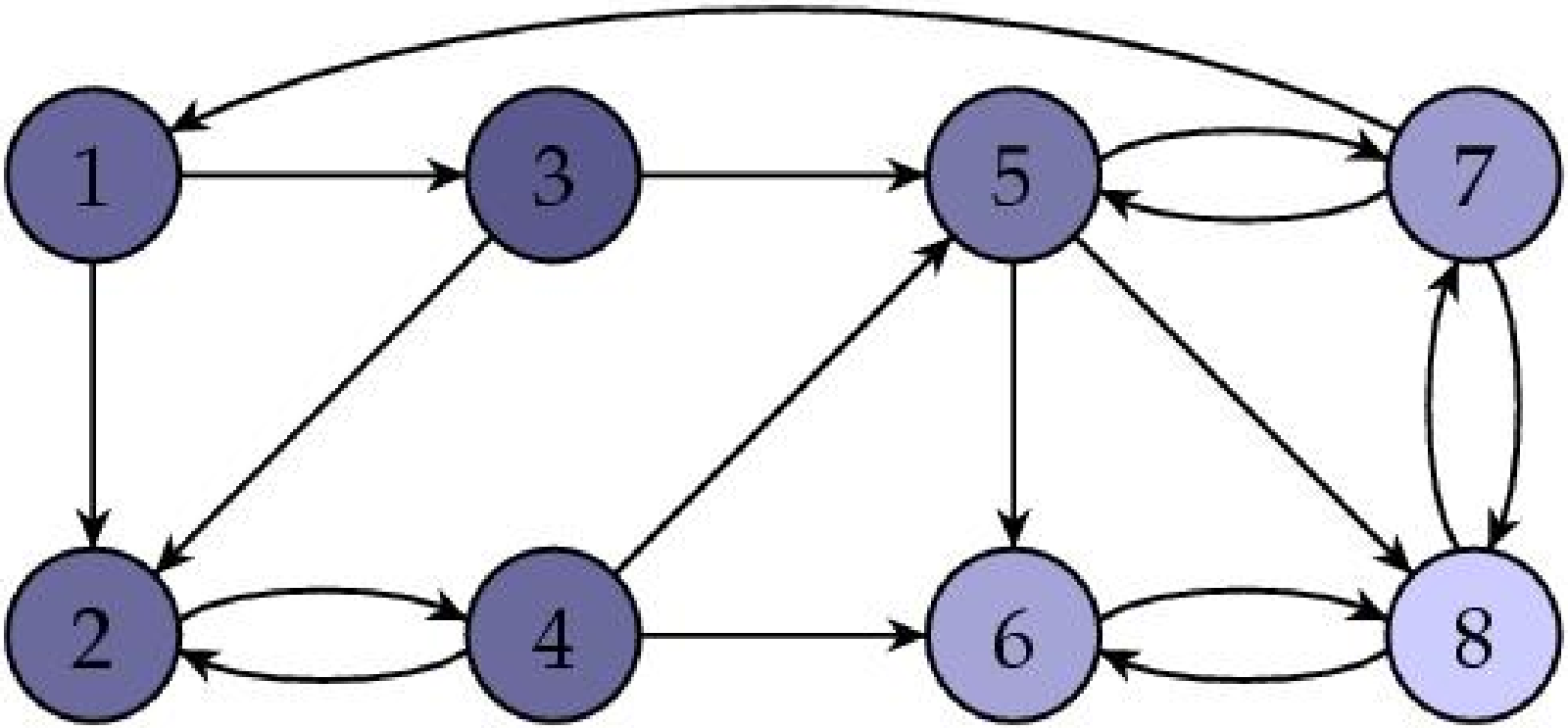
- Start with any distribution $\vec{x}$, e.g., uniform distribution
- After one step, we're at $\vec{x}P$.
- After two steps, we're at $\vec{x}P^2$.
- After k steps, we're at $\vec{x}P^k$.
- Algorithm: multiply $\vec{x}$ by increasing powers of P until convergence.
- This is called the power method.
- Regardless of where we start, we eventually reach the steady state $\vec{\pi}$.
- Thus: we will eventually (in asymptotia) reach the steady state.

Example Graph



Each inbound link is a positive vote.

Example Graph: Weighted



Pages with higher PageRanks are lighter.

PageRank summary

- Preprocessing
  - Given graph of links, build matrix P
  - Apply teleportation
  - From modified matrix, compute $\vec{\pi}$
  - $\vec{\pi}_i$ is the PageRank of page i.
- Query processing
  - Retrieve pages satisfying the query
  - Rank them by their PageRank
  - Return reranked list to the user

PageRank issues

- Real surfers are not random surfers.
  - Examples of nonrandom surfing:
    - back button
    - short vs. long paths
    - bookmarks
    - directories
    - search!
  - $\Rightarrow$ **Markov model** (future states depend only on the present state) is not a good model of surfing.
  - But it's good enough as a model for our purposes.

- Simple PageRank ranking produces bad results for many pages.
  - Consider the query **[video service]**
  - The Yahoo home page (i) has a very high PageRank and (ii) contains both *video* and *service*.
  - If we rank all Boolean hits according to PageRank, then the Yahoo home page would be top-ranked.
  - Clearly not desirable

- In practice: rank according to weighted combination of raw text match, anchor text match, PageRank & other factors

How important is PageRank?

- Frequent claim: PageRank is the most important component of web ranking.
- The reality:
  - There are several components that are at least as important:
    - anchor text
    - phrases
    - proximity
    - tiered indexes
  - Rumor has it that PageRank in its original form (as presented here) now has a negligible impact on ranking!
  - However, variants of a page's PageRank are still an essential part of ranking.
  - Adressing link spam is difficult and crucial.

Gaming PageRank

- Link Spam adding links between pages for reasons other than merit. Link spam takes advantage of link-based ranking algorithms, which gives websites higher rankings the more other highly ranked websites link to it. Examples include adding links within blogs.

- Link Farms creating tightly-knit communities of pages referencing each other, also known humorously as mutual admiration societies.

- Scraper Sites "scrape" search-engine results pages or other sources of content and create "content" for a website. The specific presentation of content on these sites is unique, but is merely an amalgamation of content taken from other sources, often without permission.

- Comment spam is a form of link spam in web pages that allow dynamic user editing such as wikis, blogs, and guestbooks. Agents can be written that automatically randomly select a user edited web page, such as a Wikipedia article, and add spamming links.

- The nofollow link: a value that can be assigned to the rel attribute of an HTML hyperlink to instruct some search engines that a hyperlink should not influence the link target's ranking in the search engine's index.

- Google does not index the target of a link marked **nofollow**.

- Yahoo! does not include the link in its ranking

- ...

Search Engine Optimization

- Search engine optimization (SEO) is the process of improving the visibility of a web site or a web page in search engines.
  - Getting Indexed
  - Cross Linking
  - URL normalization
  - Meta Tags
  - Paid Links (AdSense)
- The ultimate method:
  - Writing interesting pages with content people want to read

Current Status

- There is a continuous battle between
  - Search companies, who want to get the most useful page to the user
  - Page writers, who want to get their page read
- All metrics get gamed

Improving Citation Networks

- When most scholarship was in journals or books, then citation was easy or at least bounded
  - ⊗ Although in reality most people could access few resources
- Now many papers or sources of information only exist online or as files.
- Things on line move around, which makes citations unreliable
- We need **Persistent Identifiers**
  - **Uniform Resource Name (URN)**
    - ISBN
    - Archived Web Pages
  - **Digital object identifier (DOI)**

Digital object identifier

- DOI: a string used to uniquely identify an electronic document or object
- Metadata about the object is stored with the DOI name
- The metadata includes a location, such as a URL
- The DOI for a document is permanent, whereas the metadata may change
- Thus the DOI provides more stable linking than URLs
- The DOI system is implemented through a federation of registration agencies coordinated by the International DOI Foundation
- By late 2009 approximately 43 million DOI names had been assigned by some 4,000 organizations
  - 100 millionth DOI assigned in 2014
  - 10 millionth DOI assigned in August 2003
  - millionth DOI was assigned in April 2000

DOI example

- DOI: 10.1007/s10579-008-9062-z has metadata

```
url = http://www.springerlink.com/content/v7q114033401th5u/
type = journal
last1 = Bond | first1 = F.
last2 = Fujita | first2 = S.
last3 = Tanaka | first3 = T.
title = The Hinoki syntactic and semantic treebank of Japanese
journal = Language Resources and Evaluation
volume = 42
pages = 243
year = 2008
```

DOI characteristics

- An ID backed by
  - Persistence, if material is moved, rearranged, or bookmarked
  - Interoperability with other data from other sources
  - Extensibility by adding new features and services through management of groups of DOI names
  - Single management of data for multiple output formats
  - Class management of applications and services
  - Dynamic updating of metadata, applications and services

Makes it easier to correctly count citations – makes analysis more reliable

Other networks you can analyse

- Disease transfer — who infected whom
- Social networks — who friended whom
  Kevin Bacon game
- Semantic Networks

Summary of PageRank

- Given a graph, ranks nodes according to their relative structural importance
  - If an edge from n_i to n_j exists, a vote from n_i to n_j is produced
  - Strength depends on the rank of n_i
  - The more important n_i is, the more strength its votes will have.

- PageRank can also be viewed as the result of a random walk process
  - Rank of n_i represents the probability of a random walk over the graph ending on n_i, at a sufficiently large time.

Fake News

- Fake news is false news
  - Malicously False News
  - Satire
  - Disinformation
  - Misinformation
  - Rumour
- Fake news is intentionally and verifiably false news published by a news outlet

Acknowledgments and Further Reading

- Excellent introduction to Information Retrieval, including web searching: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/information-retrieval-book.html

- David Austin (2010) *How Google Finds Your Needle in the Web's Haystack* American Mathematical Society, Monthly Column http://www.ams.org/samplings/feature-column/fcarc-pagerank

- Hirsch, J. E. (2005). "An index to quantify an individual's scientific research output". *Proceedings of the National Academy of Science* **102(46)**: 16569–16572.